Pierre Cahuc, Stéphane Carcillo, and André Zylberberg

LABOR ECONOMICS

second edition

Labor Economics

Labor Economics

PIERRE CAHUC,
STÉPHANE CARCILLO,
AND
ANDRÉ ZYLBERBERG

The MIT Press
Cambridge, Massachusetts · London, England

The opinions expressed and arguments advanced are the exclusive responsibility of the authors, not the institutions with which they are affiliated.

For information about special quantity discounts, please email special_sales@mitpress.mit.edu.

Printed and bound in the United States of America.

Compositor: diacriTech

Library of Congress Cataloging-in-Publication Data

Cahuc, Pierre.
[Marché du travail. English]
Labor economics / Pierre Cahuc, Stéphane Carcillo, and André Zylberberg.
—Second Edition.
pages cm
Includes bibliographical references and index.
ISBN 978-0-262-02770-0 (hardcover : alk. paper)
1. Labor economics. I. Carcillo, Stéphane. II. Zylberberg, André. III.

Title.
HD4901.C2413 2014
331—dc23 2013048607

10 9 8 7 6 5 4 3 2 1

Contents

Preface

Ten years have now gone by since the first edition of *Labor Economics*. At the outset our purpose was to offer a survey of the theoretical foundations of labor economics and the empirical evaluations this discipline can furnish, expounding the models with enough detail to allow readers to see how they function. The perspective has changed, and this new edition aims to do rather more than bring the earlier one up to date. Methods of evaluation have advanced over the course of the last decade and hold a prominent place now in academic publications, a development due especially to the multiplication of individual databases and the organization of experiments. This new edition incorporates these advances: we set ourselves the goal of explaining current methods and instructing readers in how to use them by replicating research publications that have proved to be milestones in labor economics. We also devote more space to the analysis of public policy and the levers available to policy makers, with new chapters on income redistribution and the provision of protection against the risks inherent in the functioning of the labor market. There are now dedicated chapters on wage inequality and uneven access to employment, whether these phenomena arise out of technological progress, globalization, or discriminatory practices in the workplace.

In presenting current empirical methodology, we now draw heavily on research articles that have become references for the profession and so for this book, explaining their lines of reasoning and their techniques in detail. The data, as well as the corresponding Stata codes, are available at the website linked to this book, www.labor-economics.org. For each chapter, this site also puts at the disposition of readers current data related to the figures and tables in the book, up-to-date indications of important publications, and slides illustrating the main points of the chapter, which may be used as course aids. The site also includes a discussion forum.

The task we set ourselves was ambitious, too much so for just the pair of us. The gearbox needed an extra gear, so Stéphane Carcillo joined the team. Thanks to him, we have been able to bring this project to completion.

Pierre Cahuc and André Zylberberg

Introduction

On 4 April 1980, following a conflict with the Peruvian government, Cuban President Fidel Castro ordered the guards posted in front of the Peruvian embassy in Havana withdrawn. Seizing the chance offered by this absence, almost 11,000 Cubans stormed into the embassy and demanded political asylum. Images of a multitude of hungry and thirsty men, women, and children who were seen perched in trees and on the roof of the embassy, were immediately broadcast worldwide. After difficult negotiations and under strong pressure from the international community, the Cuban government agreed to allow all the asylum seekers in the embassy to leave the country. They were taken in by Costa Rica, Spain, Peru, and the United States. To counter the negative image created by this event, Fidel Castro announced, in his famous speech of 20 April 1980, that he was throwing open the port of Mariel, a municipality 25 miles west of Havana, so that anyone who wanted to leave the island of Cuba could do so. A veritable human tidal wave followed this speech: starting in May, almost 90,000 Cubans left their country for the United States. It is estimated that by the time the port of Mariel was closed again in September 1980, more than 125,000 Cubans had emigrated. Half of them settled in Miami, causing the labor force there to swell by 7%.

Between April and July 1980, in the space of three months, the unemployment rate in Miami soared, going from 5% to 7.1%. This flare-up of unemployment aroused the kind of reaction one can imagine. In some quarters the Cuban refugees were recklessly accused of stealing work from the least qualified Americans. Perhaps those who think that the number of jobs is bounded by (mysterious) limits independent of the size of the labor force were about to be proved right. If they were, this expansion of the labor force would do nothing other than expand the volume of unemployment. One way to test this thesis—the only way, actually—is to address this question: what would have happened in Miami if the Mariel boatlift had never taken place?

A Canadian economist, David Card, took on this problem. But he is not a seer, just a professor at the University of California at Berkeley, incapable of knowing any better than anyone else what would have happened in Miami if the Mariel boatlift had not occurred, since in reality it did occur. Economists very often encounter problems of this type. They try to solve them by comparing the real circumstance to a "control" situation, one that mirrors the real circumstance as closely as possible without the "disturbance" the effects of which they want to assess. In the case at hand, the Mariel boatlift constitutes the disturbance. David Card had the idea of taking as his controls U.S. cities with economic and demographic profiles similar to those of Miami but which had not been affected by the great wave of Cuban immigration in 1980. He picked Atlanta, Los Angeles, Houston, and Tampa–St. Petersburg. Like Miami, these four cities included large black and Hispanic communities and had undergone similar changes in employment and unemployment in the years before the Mariel boatlift. Card compared the average movements of wages and unemployment before and after the Mariel boatlift in the black,

Hispanic, and white communities in these five cities, taking into account differences of education, experience, marital status, industry sector, and amount of part-time work. This approach, called difference-in-differences, has become one of the core empirical strategies in labor economics.

In 1979, a year before the Mariel boatlift, the unemployment rate in the white population of Miami reached 5.1%; in 1981, a year after the Mariel boatlift, it fell to 3.9%. In other words, the unemployment rate in the white population of Miami went down by 1.2 percentage points between 1979 and 1981. Over the same period, the unemployment rate in the white population of the control cities also went down, but only by 0.1 percentage point. The comparison of these two figures, 1.2% and 0.1%, warrants the conclusion that the influx of Cubans into Miami did not have a negative effect on employment within the white population. This can be seen on the upper panel of figure I.1.

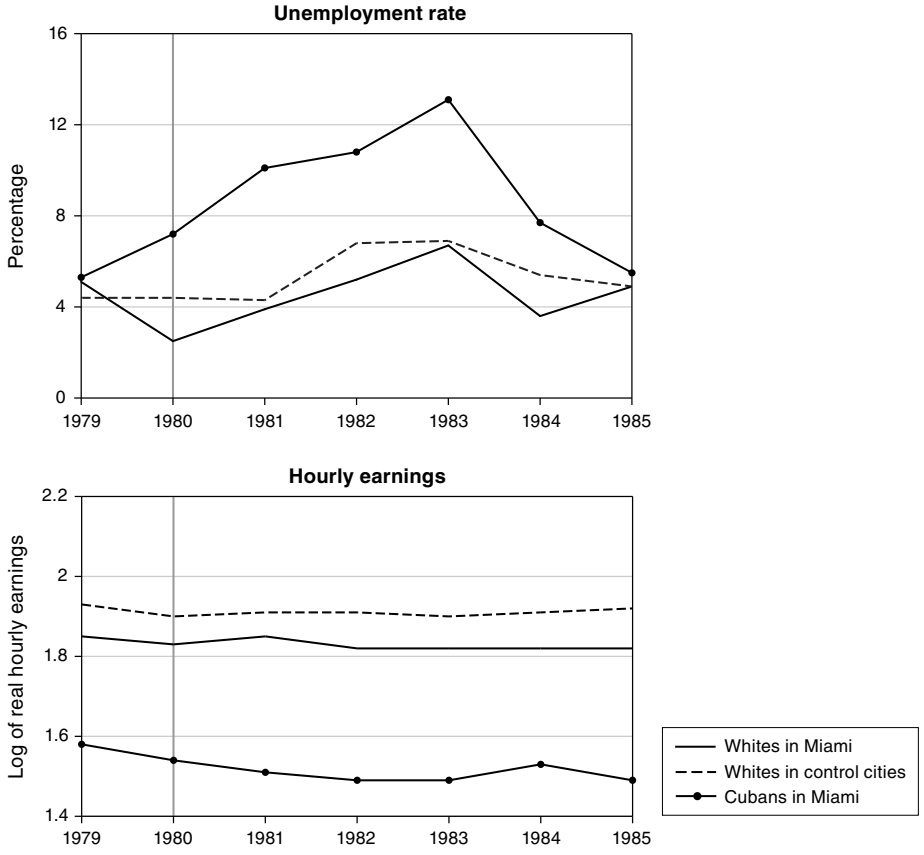


Figure I.1
Unemployment rate and earnings among whites and Cubans in Miami and control cities around 1980 (active population 16–61 years old).

Source: Card (1990, tables 3 and 4, pp. 250–251).

Neither did the Cuban influx have a negative effect on the black population, though that was the one most exposed to competition from the new refugees. It is true that the unemployment rate in the black population rose by 1.3 percentage points between 1979 and 1981, but in the same period it rose by 2.3 percentage points in the control cities. More generally, Card's study shows that patterns of development in the labor market in Miami and the control cities were very similar long after the Mariel boatlift. The earnings differentials across cities were also constant between 1979 and 1984, as shown on the lower panel of figure I.1 for the white population. So the Cuban immigration had no significant effect on the wages and employment of persons living in Miami, although it had a strong albeit temporary impact on the unemployment rate among Cubans because many had to look for a job upon their arrival. After a very strong upsurge of unemployment immediately following the arrival of the new immigrants (remember that the unemployment rate in Miami went from 5% to 7.1% between April and July 1980), all indications are that this city absorbed an exceptional influx of newcomers in the space of a year.

One objection immediately comes to mind. Might these results not be the result of the flight of a large proportion of the resident population, which left to seek jobs elsewhere because of the incoming tide of Cubans? Card was able to establish that in the years following the massive upswing of Miami's population in 1980, employment opportunities among the nonmigrants in this city did not degrade appreciably in comparison to the control cities. Thus, if there was a flight of the resident population, it was a small one. The majority of studies focusing on the United States, and on many other countries, come to analogous conclusions: inflows of migrants have a very weak impact on wages, employment, and the mobility of residents.

These results may occasion surprise, but they cannot be ignored. The rapid absorption of the Mariel immigrants was made possible by the presence in the Miami region of relatively low- to semi-skilled industries, such as apparel, textiles, agriculture, furniture, private household services, hotels and motels, and restaurants. These industries typically expanded at that time in cities with strong immigration. New cohorts of migrants also replaced earlier cohorts as the latter moved to more desirable jobs. These realities are incompatible with a static vision of the market, in which the number of jobs would be thought of as a set quantity and in which it would be taken for granted that immigration or any shock that increased the labor supply was certain to have strong, persistent, and detrimental effects on wages and unemployment.

We mention David Card's article right at the beginning of this book for a good reason: it can stand as a virtual emblem of the whole of labor economics in at least two respects. First, it brings to light relations of cause and effect. The Mariel boatlift constituted a sudden and unanticipated shift in a single variable, the quantity of immigrant labor present on the Miami market, just as in a laboratory experiment where scientists modify one factor in an environment while holding the rest constant. Thanks to this sudden and isolated change, Card was able to pinpoint, and throw into relief, the impact of the Cuban migration on a labor market in the United States. Hence the

Mariel boatlift qualifies as a "natural experiment." Over the last few decades the elaboration of experimental designs capable of registering genuine relations of cause and effect has profoundly reshaped labor economics. These techniques and their results are given plenty of space in our book. For that matter, David Card's experimental design yielded a "negative" result and so "disproved" a widely held misconception: massive as it was, the Cuban migration had no detectable impact on the path over time of the unemployment rate in Miami in comparison to other American cities that underwent no such influx. This counterintuitive result is a telling reminder of how quickly the labor market can react to change.

A second distinctive aspect of contemporary labor economics is this: the progress made since the 1980s in the acquisition of data has revealed that such phenomena as job creation and job destruction, and flows of manpower in general, have orders of magnitude hitherto unsuspected. These are fresh facts, and they too have profoundly reshaped our conception of the labor market. Today the labor economist conceives the labor market in a dynamic perspective that takes fully into account its incessant recomposition. Our book faithfully reflects this dynamic perspective.

The Importance of Labor

Labor economics is the study of the exchange of labor services for wages—a category that takes in a wide range of topics. The main ones are labor supply, labor demand, the impact of education on wages and employment, the influence of technological change, the influence of human migration, the role of unions, the labor contract, working conditions, job search by the unemployed, discrimination, the institutional framework in which hiring and firing take place, mandatory payroll contributions, and finally the impact of the levers used by policy makers to achieve income redistribution and stimulate (or protect) employment.

Developing a specific field for labor economics is justified by the importance of the exchange of labor services in modern economies. A large part of the population is made up of employees who are earning wages and others aspiring to become wage earners in the future if they have not yet left the educational system, or aspiring to become wage earners right now if they are looking for work. Figure I.2 tracks the path over time of employees as a proportion of the overall population of working age in 18 OECD (Organization for Economic Co-operation and Development) countries from 1970 to 2010. The proportion of employees is clearly high and has been rising over the last 40 years. It is influenced by the demand for labor, and that is an explanatory factor for its sharp drops during the 2008–2009 recession. It is also influenced by education and labor supply decisions that we analyze in detail in this book. Naturally employees make up the bulk of the category of all workers, which includes independent workers and employers. Their share has increased over the past 40 years and now ranges from 70% to over 90%, as shown in Figure I.3 for the same 18 countries. This explains the focus of labor economics on labor contracts and relations between firms and employees.

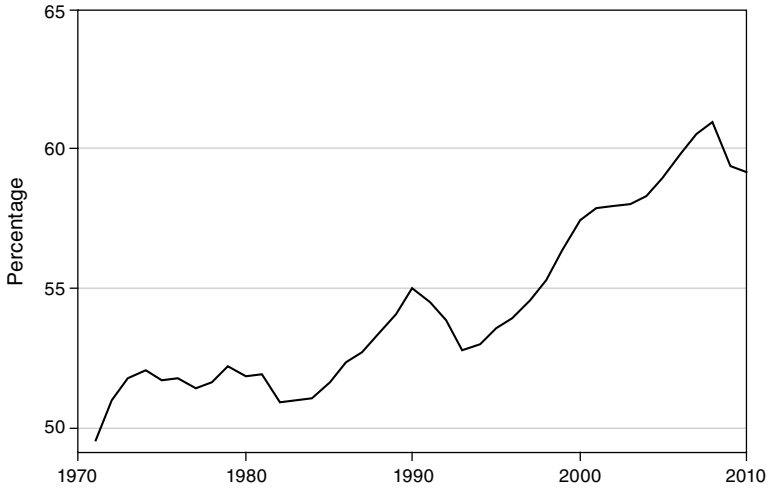


Figure I.2
Share of employees in the working-age population (15–64 years old) between 1970 and 2010 in 18 OECD countries.

Note: Nonweighted average of Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany, Ireland, Italy, Japan, Korea, Norway, Portugal, Spain, Sweden, United Kingdom, United States.

Source: OECD Annual Labor Force database.

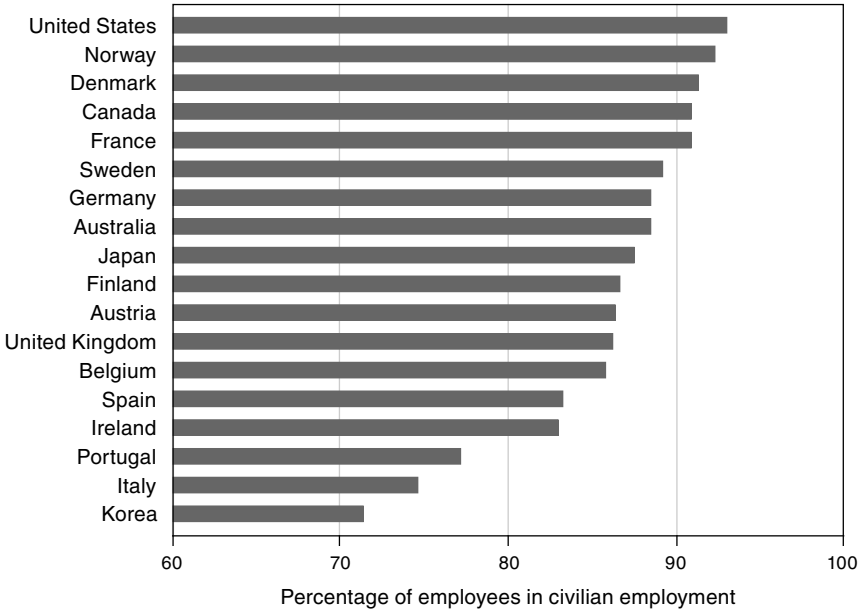


Figure I.3
Share of employees in civilian employment in 2010 in 18 OECD countries (Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany, Ireland, Italy, Japan, Korea, Norway, Portugal, Spain, Sweden, United Kingdom, United States).

Source: OECD Annual Labor Force database.

Models

Throughout this book we make plenty of room for facts. But we also present theories, most often in the form of mathematical models that have been built by labor economists in order to probe these facts and elicit their meaning. We also present techniques that enable them to compare the predictions of their models against the facts. Hence, we move back and forth between facts and data on one hand and theoretical reasoning on the other. For example, the study of labor supply includes descriptive material on the evolution of participation rates and the number of hours worked, as well as a model that explains individual choices on the basis of traditional hypotheses about individual rationality and scarcity of resources. Methods of assessing this model empirically, and the main empirical results, are then laid out. In this way we are able to understand, assess quantitatively, and predict the impact of changes in wages, the fiscal system, or social assistance on the labor supply in different contexts.

Economics makes use mainly of mathematical models. This textbook conforms to that rule. At least three reasons may be cited in justification.

The first, and by no means the least compelling, lies precisely in the quasi-monopoly held by this approach. The student owes it to himself or herself to become familiar with it if he or she wants to be able to read specialized journals in the field. But the domination of formalized economics is not the outcome of a random draw from among several possible equilibria. For one thing, economic analysis lends itself to mathematization, since it deals with quantified magnitudes. The questions put to economists generally demand answers in the form of numbers: Is wage inequality rising? Is competition from low-wage countries destroying jobs? Are mandatory contributions favorable to employment? In order to be precise and operational, the answers to questions like those have to be given in numbers, justified by a coherent chain of reasoning, with the underlying hypotheses made clear.

These requirements constitute another justification of mathematization. A mathematical model allows us to clearly establish a linkage between hypotheses and results. It proves particularly effective, indeed indispensable, when the mechanisms studied are complex and involve relations among a number of variables. Mathematization is entirely unavoidable if we want to understand strategic interactions, decisions made in uncertainty, situations of asymmetric information, and the dynamic choices of agents, to give a few examples.

Labor economics and all of economics have undergone a profound theoretical restructuring in recent decades, benefiting especially from advances made in the study of dynamics, strategic behavior, and decisions in uncertain environments. Both the analysis of labor supply, labor demand, wage formation, and the determinants of employment and unemployment, and the evaluation by labor economists of government interventions have been profoundly shaped by these advances. Our aim is to set this spectrum of developments before the reader within a unified didactic framework and to show that they have measurably improved our understanding of the functioning of the labor market. We have nevertheless taken great care to make our models as simple as possible. A mathematical appendix at the end of the book supplies the toolkit needed

to understand all the models utilized in the text. Finally, we have tried to articulate our theoretical and empirical lines of reasoning.

Econometrics and Evaluation

Readers should be aware that, beginning in the 1970s, labor economics has become the preferred arena in which to apply the most advanced econometric methods (microeconometrics in particular). The surveys by DiNardo and Lee (2011), French and Taber (2011), Keane, Todd, and Wolpin (2011), and List and Rasul (2011), and the book by Angrist and Pischke (2009) trace the development of empirical research in this field. The 1990s and 2000s have been particularly fruitful in this regard, and such empirical techniques as the experimental method, techniques for the estimation of structural models, and dynamic approaches have gained in importance. For example, the development of structural models of duration has permitted a better understanding of the behavior of those hunting for work and of the role unemployment insurance plays. Evaluative techniques that rely on natural experiments, field experiments, or laboratory experiments have likewise permitted some answers to some of the questions arising out of such controversial matters as discrimination and the real effect of certain policy levers on employment. Yet much remains to be done.

In this book we therefore make plenty of room for presentations of the empirical techniques used by labor economists, stressing the hypothetical underpinning required to bring to light a causal relation (when that is possible). Our presentations rely on one or several published articles to which we refer in each chapter and list as Further Readings. These articles have been chosen because they rely on methods that have risen to classic status and on clear strategies. Convinced that learning comes from doing, we present these reference articles in detailed fashion so that the reader may follow the methods employed step by step and learn to understand the conditions of validity and the limitations of their results. The website linked to this book, www.labor-economics.org, contains databases and programs that allow users to work through the main results of these articles for themselves. In addition to the short list of these essential articles, we supply extensive bibliographies for each chapter.

Road Map

This book is composed of four parts. Part One presents labor supply and labor demand behaviors. It shows how the interaction of supply and demand on competitive markets determines wages and employment. It also shows how the mechanisms of competition drive investments in education and training.

Chapter 1 presents consumption–leisure trade-off models and the theory of labor supply. Scrutiny of the trade-off between consumption and leisure is especially important for understanding fluctuations in the participation rates of different categories of the population and the choices people make about how much to work and when to retire. It includes a guide to the econometrics of labor supply and gives an example of

the identification of labor supply elasticities. Chapter 2 is dedicated to labor demand, first from a static perspective and then a dynamic one. Here we look at important questions like the impact of the costs of the factors of production on labor demand and substitution between capital and labor, the trade-off between workers and hours, and the effects of the adjustment costs of labor. Chapter 3 describes the basic competitive equilibrium model of the labor market. This model offers important insights into the problem of fiscal incidence. It also makes predictions on how wages and employment react to labor demand or supply shocks. As well the chapter provides examples of how to estimate labor demand elasticities. Extensions of this competitive model also predict how wage differentials should compensate for the laboriousness or danger of tasks, or how even minuscule differences in talent can translate into huge remuneration differentials. Chapter 4 presents decisions about education and their impact on individual performances in the labor market. This chapter specifies the determinants of individual choice about education and also the economic role played by education, which serves not just to transmit knowledge that improves productivity and socialization but also to select individuals within different productive sectors.

Part Two comprises four chapters devoted to imperfectly competitive labor markets. It shows how job search costs and asymmetric information shape workers' and employers' behaviors. It also provides insights on the role of unions and on the existence of discrimination.

Chapter 5 describes labor markets in the presence of job search costs. It explores the implications of costs arising from searching for a job when workers do not have cost-free access to perfect information about all the jobs available in the economy. The search model yields predictions that shed light on how the duration of unemployment depends on the characteristics of unemployed workers and how it is influenced by unemployment insurance. The chapter provides examples of how labor economists have gone about evaluating the impact of changes in unemployment insurance benefits. The search framework can also explain why identical workers can be paid differently and why small and large firms do not offer the same wages. Chapter 6 explores more deeply wage policies in situations of uncertainty and imperfect information, using agency and implicit contract models. These models throw interesting light on the logic of certain aspects of human resources management, like advancement by seniority and systems of promotion. The chapter explains under what circumstances firms and workers will have an interest in engaging in long-term relationships and why firms may have recourse to hierarchical promotions, internal markets, or other remuneration strategies to motivate workers. It also shows how social preferences and reputation effects may interact with incentives. Chapter 7 introduces collective bargaining, focusing on the behavior of unions and the manner in which we formalize the bargaining process. It analyzes the determinants of unionization and the impact of the bargaining power of workers on employment, profits, and productivity at the firm level; it also discusses strategies for the identification of these effects. The chapter looks as well at the opposition between employees with a steady job, the insiders, and workers who do not have this security, the outsiders, and shows that this opposition may be detrimental to employment and favor

the segmentation of the labor market. An example is given of the identification of the causal impact of labor unions in the United States. Chapter 8 shows how discrimination can arise and persist when labor markets are imperfectly competitive. It introduces several competing sources of discrimination known in the literature as "taste for discrimination" and "statistical discrimination," and it explains the role search frictions play in the persistence of discrimination. Several methods of estimating discrimination are presented in detail, notably for the cases of discrimination against blacks and women. Since discrimination can account for only a fraction of wage differences, other important "premarket" factors, such as ability and psychological attributes, are also discussed.

Part Three presents the contemporary perspective of labor economics on the phenomenon of unemployment and the role unemployment plays in the huge processes of job creation and job destruction that are going on at every moment in modern economies. These three chapters build on the job search and matching model and provide a detailed account of job and worker reallocations. This approach helps readers gain an understanding of the phenomenon of unemployment and the impact of technological progress and globalization on labor markets.

Chapter 9 reviews the main facts regarding unemployment in the OECD countries and uses search and matching models to study the determinants of employment and wages in a labor market in which jobs are ceaselessly destroyed and created and in which the reallocation of manpower is costly and takes time. In this chapter we diagnose the importance of frictional unemployment arising from the process of job destruction and creation. Chapter 10 studies the effects of technological progress on income inequality and unemployment. It recognizes the heterogeneity of manpower by distinguishing workers according to their skill levels. The chapter shows that technological progress had a significant impact on wage inequality and on the occupational structure of the workforce over the last century, and it analyzes in detail the phenomenon of wage and job polarization observed in the advanced economies over the recent decades. Chapter 11 turns to the effects of globalization (trade and migrations) on income inequality and unemployment. It shows that trade may have a positive impact on the productivity of firms and, as a result, can influence inequality and unemployment, depending on the capacity of the labor market to adjust. The chapter also shows that the impact of migration on wages and unemployment ought to be small even in the short term. The long-term trends of trade and migration are presented in detail, along with a number of empirical identification strategies.

Part Four contains three chapters devoted to public policy and the levers available to policy makers. The aim of this part is to subject the rationale and the impact of such policy levers to analysis. This analysis is conducted from an international perspective, highlighting the strong heterogeneity across countries.

Chapter 12 focuses on income redistribution policies and analyzes the impact of taxes and benefits on wages, employment, unemployment, labor market participation, and hours of work. The question of fiscal incidence is examined in detail. The chapter also presents the main features of minimum wages, a policy lever that is integrated into overall income policy in a number of OECD countries. The impact of the minimum

wage on labor market performance is analyzed in detail, as well as the empirical debates that have arisen concerning the issue, notably in the United States. Chapter 13 turns to insurance policies and employment protection legislation. It provides an overview of the unemployment insurance and employment protection systems in the OECD area. The principles of optimal unemployment benefits are characterized in various settings, and the impact of employment protection measures on wages, unemployment, productivity, and segmentation is set out for the reader. The chapter explores the potential interactions between employment protection and unemployment benefits. Chapter 14 surveys the variety of active labor market policies that have been implemented in the OECD countries to lower unemployment and analyzes their respective advantages and drawbacks in an equilibrium framework. It discusses the methodological principles that guide the evaluation of such labor market policy levers and provides an assessment of their respective impacts. The chapter provides detailed examples of identification strategies with respect to equilibrium effects and a synthesis of empirical results based on meta-analyses of hundreds of evaluation papers.

How This Book May Be Used

We deal with a wide range of topics in this book, and not all of them present the same degree of formal and conceptual difficulty. Those to whom they are taught may be studying for a degree at the level of bachelor, master, or doctor. The book's length dictates, moreover, that instructors using it to prepare courses in labor economics will assign selected readings. Here we offer examples of what we think are practical sequences.

- A course in basic labor economics, foregrounding competitive structures and behaviors in an essentially static environment

  1. Facts about labor supply (chapter 1, section 1), the basic model of labor supply and its various extensions (chapter 1, sections 2.1 and 2.2), and the econometric approach (section 3.1) followed by the empirical results (section 3.2). The econometric approach can also be presented using the evaluation of the impact of taxes on labor supply (chapter 12, section 1.3).

  2. The static theory of labor demand (chapter 2, section 1) as well as empirical estimates of the elasticities of labor demand (section 2).

  3. The competitive equilibrium (chapter 3, section 1.1), the question of tax incidence (section 1.2), and the adjustment to a shock on labor supply (section 3.3).

  4. Problems connected to education (chapter 4), including the factual elements (section 1), the theory of human capital (section 2.1), and the empirical assessment of the returns to education (sections 4.1 and 5.1).

5. Introduction of obstacles to competition, leading to discussion of monopsony (chapter 12, section 2.2.1) and theories of discrimination (chapter 8, section 2), and empirical work on compensating differentials (chapter 3, section 2.2), on discrimination (chapter 8, section 4), and the minimum wage (chapter 12, section 2.3.2).

6. The evolution of wage inequalities (chapter 10, section 2.1), taking into consideration the role of technological progress (section 2.2), international competition (chapter 11, section 1.2), and migratory flows (chapter 11, section 3.2).

7. The assessment of labor market policies (chapter 14, section 2), including elements of methodology (section 3.1) and the main empirical results (section 4).

- An in-depth course oriented toward microeconomics and dealing with wage formation and dynamic and informational problems

1. The intertemporal labor supply (chapter 1, section 2.3), with an example of the identification of elasticities (chapter 1, section 3.1).

2. Problems connected to education (chapter 4), bringing in the determinants of the duration of studies (sections 2.2 and 2.3), the signaling model (section 3), and the shift from the model of human capital to empirical identification, with the main results (sections 4 and 5).

3. Compensating wage differentials (chapter 3, section 2).

4. The effect of talent on wage distribution (chapter 3, section 3).

5. The assignment of skills to tasks (chapter 10, section 2.2).

6. The job search model and how it applies to wage formation (chapter 5).

7. Optimal unemployment insurance (chapter 13, section 1).

8. The dynamic theory of labor demand (chapter 2, section 3).

9. The labor contract in the presence of uncertainty and problems of incentive (chapter 6).

10. Collective bargaining (chapter 7).

11. Discrimination (chapter 8).

- A course in labor economics more focused on problems of unemployment and inequality

  1. Job search (chapter 5).
  2. The search and matching model (chapter 9).
  3. Technological progress and unemployment (chapter 10, section 1).
  4. Technological progress and inequality (chapter 10, section 2).
  5. International trade (chapter 11, sections 1 and 2).
  6. Migrations (chapter 11, section 3).
  7. Taxes and benefits (chapter 12, section 1).
  8. Minimum wage (chapter 12, section 2).
  9. Unemployment insurance (chapter 13, section 1).
  10. Employment protection (chapter 13, section 2).
  11. Training, employment subsidies, and job search assistance (chapter 14).

Data and Stata Programs and Teaching Material

For most of the facts set forth in this book, and also for the complete set of reference articles that are given detailed presentations, the relevant data are available, chapter by chapter, at the website www.labor-economics.org. In addition to databases, the site offers files in the .do format that make it possible to reproduce the figures and tables of estimations under Stata.

There follows a nonexhaustive list of the topics of the empirical reference articles for which the data are available at www.labor-economics.org.

- Assessing labor supply elasticities following tax changes, using grouping estimates (Blundell, Duncan, and Meghir, 1998; chapter 1, section 3.1.2).
- Assessing labor demand elasticities following a shock on labor supply, using instrumental variables (Acemoglu, Autor, and Lyle, 2004; chapter 3, section 1.3).
- Evaluating the impact of education on earnings, using instrumental variables (Angrist and Krueger, 1991; chapter 4, section 4.2.3).

- Evaluating the impact of unemployment benefits on the duration of unemployment, using difference-in-differences and duration models (Lalive, van Ours, and Zweimuller, 2006; chapter 5, section 3).
- Identification of discrimination, based on wage equations (Neal and Johnson, 1996; Lang and Manove, 2011; chapter 8, section 3.1).
- Identification of discrimination, based on the Blinder-Oaxaca decomposition method (O'Neill and O'Neill, 2006; chapter 8, section 3.2).
- Measuring the influence of technological progress on wage inequality between high- and low-skilled workers (Acemoglu and Autor, 2011; Autor and Dorn, 2013; chapter 10, section 2.3).
- Empirical evidence on the relationship between trade and unemployment, based on macroeconomic data and using the Arellano-Bond method (Dutt, Devashish, and Priya, 2009; chapter 11, section 2.1).
- The effect of migration on local labor markets, using spatial correlations and instrumental variables (Boustan, Fishback, and Kantor, 2010; chapter 10, section 3.3.1).
- Measuring the impact of tax credits on labor market participation and hours, using a difference-in-differences estimator (Eissa and Liebman, 1996; chapter 12, section 1.3.1).
- Exploring the origins of the United States/Europe difference in working hours, based on macroeconomic data (Prescott, 2004; chapter 12, section 1.3.2).
- Identifying the impact of minimum wage increases on employment, using a difference-in-differences estimator (Card and Krueger, 1994; chapter 12, section 2.3.2).
- Identifying the impact of targeted job placement programs for skilled youth, based on a randomized experiment (Crépon, Duflo, Gurgand, Rathelot, and Zamora, 2013; chapter 14, section 3.2).

REFERENCES

Acemoglu, D., & Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 4B, pp. 1043–1171). Amsterdam: Elsevier Science.

Acemoglu, D., Autor, D., & Lyle, D. (2004). Women, war and wages: The effect of female labor supply on the wage structure at mid-century. *Journal of Political Economy*, *112*(3), 497–551.

Angrist, J., & Krueger, D. (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, *106*, 976–1014.

Angrist, J., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Autor, D., & Dorn, D. (2013). The growth of low-skill service jobs and the polarization of the U.S. labor market. *American Economic Review*, *103*(5), 1553–1597.

Blundell, R., Duncan, A., & Meghir, C. (1998). Estimation of labour supply responses using tax policy reforms. *Econometrica*, *66*(4), 827–861.

Boustan, L., Fishback, P., & Kantor, S. (2010). The effect of internal migration on local labor markets: American cities during the great depression. *Journal of Labor Economics*, *28*(4), 719–746.

Card, D. (1990). The impact of the Mariel boatlift on the Miami labor market. *Industrial and Labor Relations Review*, *43*, 245–257.

Card, D., & Krueger, A. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review*, *84*, 772–793.

Crépon, B., Duflo, E., Gurgand, M., Rathelot, R., & Zamora, P. (2013). Do labor market policies have displacement effects? Evidence from a clustered randomized experiment. *Quarterly Journal of Economics*, *128*(2), 531–580.

DiNardo, J., & Lee, D. (2011). Program evaluation and research designs. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 4, no. 4, pp. 463–536). Amsterdam: Elsevier Science.

Dutt, P., Devashish, M., & Priya, R. (2009). International trade and unemployment: Theory and cross-national evidence. *Journal of International Economics*, *78*(1), 32–44.

Eissa, N., & Liebman, J. (1996). Labor supply responses to the earned income tax credit. *Quarterly Journal of Economics*, *112*(2), 605–607.

French, E., & Taber, T. (2011). Identification of models of the labor market. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 4A, pp. 537–617). Amsterdam: Elsevier.

Keane, M., Todd, P., & Wolpin, K. (2011). The structural estimation of behavioral models: Discrete choice dynamic programming methods and applications. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 4A, pp. 331–461). Amsterdam: Elsevier Science.

Lalive, R., van Ours, J., & Zweimuller, J. (2006). How changes in financial incentives affect the duration of unemployment. *Review of Economic Studies*, *73*, 1009–1038.

Lang, K., & Manove, M. (2011). Education and labor market discrimination. *American Economic Review*, *101*, 1467–1496.

List, J., & Rasul, I. (2011). Field experiments in labor economics. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 4A, pp. 103–228). Amsterdam: Elsevier Science.

Neal, D., & Johnson, W. (1996). The role of premarket factors in black-white wage differences. *Journal of Political Economy*, *104*, 869–895.

O'Neill, J., & O'Neill, D. (2006). What do wage differentials tell us about labor market discrimination? *Research in Labor Economics*, *24*, 293–357.

Prescott, E. (2004). Why do Americans work so much more than Europeans? *Federal Reserve Bank of Minneapolis Quarterly Review*, *28*(1), 2–13.

Acknowledgments

This second edition owes a great deal to our editor at MIT Press, John Covell. We would not have re-embarked on such an ambitious undertaking without John's support, his strong conviction, and the encouragement he has provided. It also owes much to the competence and efficiency of our translator, William McCuaig, who has been working with us for over ten years now on both versions of the text. At Books By Design in Cambridge, MA, our thanks go to project manager Nancy Benjamin and copyeditor Joan M. Flaherty.

What especially sets the second edition apart is the data sets made available to us by the authors of the empirical research papers whose methodologies we present in detail. We thank Richard Blundell, Leah Boustan, Pushan Dutt, Nicole Fortin, Rafael Lalive, Jeffrey Liebman, Roland Rathelot, and Philippe Zamora, with whom we have been in contact, and their co-authors, for their contribution. We likewise thank those, including David Autor, David Card, Steven Davis, Kevin Lang, and Thomas Lemieux, who have put the data and programs linked to their articles online, enabling us to present their results in detail. Florent Fremiggacci and Florian Guyot gave us precious assistance in organizing these data and the accompanying Stata programs for the website www.labor-economics.org.

This work took shape in courses given in Paris at the Ecole Nationale des Statistiques et de l'Administration Economique (ENSAE), the Ecole Polytechnique, Sciences Po, and the Université Paris 1 Panthéon-Sorbonne, as well as at the Norwegian School of Economics and the Université des Antilles et de la Guyane. We thank our students, whose curiosity and questioning have been a constant source of inspiration to us.

We also thank all the individuals, teachers, and researchers who have helped us with their observations, encouragement, assistance, and advice during the writing of the book in its different stages, in particular Yann Algan, Dominique Anxo, David Autor, Antoine d'Autume, Andrea Bassanini, Raicho Bojilov, Georges Bresson, Ana Cardoso, Jean-Louis Cayatte, Bart Cockx, Jacques Cremer, Huw Dixon, Manon Domingues Dos Santos, Christine Erhel, Patrick Fève, Rodrigo Fernandez, Gary Fields, François Fontaine, Pauline Fron, José Ignacio García Pérez, Christian Gianella, Pierre Granier, Pierre Hallier, Dan Hamermesh, Herwig Immervoll, Michel Juillard, Hubert Kempf, Sebastian Königs, Francis Kramarz, François Langot, Thierry Laurent, Etienne Lehmann, Olivier L'Haridon, Claudio Lucifora, Thierry Magnac, Franck Malherbet, Philippe Martin, Sebastien Martin, Jean-Baptiste Michau, Marie-Laure Michaud, Pierre Morin, Yann Nicolas, Philip Oreopoulos, Xavier Pautrel, Corinne Perraudin, Alain Pirotte, Erik Plug, Fabien Postel-Vinay, Julien Pratt, Ana Prieto, Corinne Prost, Muriel Pucci, Agnès Puynoyen, Monika Queisser, Linda Richardson, Jean-Marc Robin, Muriel Roger, Anne Saint-Martin, Stefano Scarpetta, Barbara Sianesi, Jean-Marc Tallon, Kostas Tatsiramos, Bruno van der Linden, Danielle Venn, and Yves Zenou.

PART ONE

Labor Supply and Demand Behaviors

CHAPTER 1

Labor Supply

In this chapter we will:

- See how people make choices between consumption, leisure, and household production
- Learn what the reservation wage is
- Understand why the shape of the labor supply curve results from the combination of substitution effects and income effects
- Learn what the wage elasticities of labor supply are
- Explore when and why people decide to retire
- Master the principles guiding the econometrics of labor supply and the main empirical results
- Apply these principles using data and programs that allow us to replicate the main results of the paper of Blundell, Duncan, and Meghir (1998) on the estimation of labor supply elasticities
- Provide an overview of the result of macro and micro empirical studies on labor supply elasticities

INTRODUCTION

In 1900, in the United States prime-age men (those between 25 and 54) used to work on average 50 hours per week at their jobs. Prime-age women worked only 8 hours but would do 50 hours of unpaid household work compared with only 4 hours for men. One hundred years later, in 2005, the situation had changed dramatically: prime-age men worked on average 37 hours and did more at home (17 hours per week), while women were now much more active in the labor market, being employed for 26 hours per week on average and working at home for 31 hours.[1] This change led overall to a significant increase in the labor force and is the result of the choices made by every single working-age person in the population regarding work hours, home duties, and leisure.

[1] The data are from Francis and Ramey (2009).

To hold a paid job, you must first have decided to do so. This is the starting point of the so-called neoclassical theory of the labor supply. It posits that each individual disposes of a limited amount of time, which he or she (henceforth we will switch randomly between these gendered pronouns in referring to unspecified individuals) chooses to allocate between paid work and leisure. Evidently the wage an individual can demand constitutes an important factor in the choice of the quantity of labor supplied, and taxes also play a role. But it is not the only factor taken into account. Personal wealth, income derived from sources outside the labor market, and even the familial environment also play decisive roles.

In reality the allocation of one's time depends on trade-offs more complex than a simple choice between work and leisure. In the first place, the counterpart of paid work is not simply leisure in the usual sense, for much of it consists of time devoted to "household production" (the preparation of meals, housekeeping, minor repairs and upkeep, the raising of children, etc.), the result of which substitutes for products available in the consumer goods market. This implies that the supply of wage labor takes into account the costs and benefits of this household production and that most often it is the result of planning, and even actual negotiation, within the family. The family situation, the number of children, the income a person enjoys apart from any wage labor (personal wealth, illegal work, spousal income, etc.) all weigh heavily in this choice. Decisions concerning labor supply also depend on trade-offs over the course of time that make the analysis of the agents' decisions richer and more complex.

Empirical studies on labor supply, which have multiplied in the course of the last 30 years, shed light on the determinants of labor supply. The development of these studies—reviewed in Blundell and MaCurdy (1999), Blundell et al. (2007), and Keane (2011)—has benefited from advances made in the application of econometric methods to individual data. It has also been driven by a need to evaluate public policies that attempt to influence labor supply directly, such as tax and benefit systems. A number of countries have set up programs explicitly aimed at increasing labor supply among the most disadvantaged, rather than park them on the welfare rolls. These "welfare-to-work" programs, sometimes abbreviated as workfare so as to contrast them with more traditional programs called simply welfare, have given a powerful incentive to empirical research on labor supply in the United States and Great Britain, as well as in certain European continental countries like Sweden and France. A better understanding of labor supply behaviors is also key for the design of tax systems in general. The more sensitive the labor supply to the net wage, the lower the optimal tax rate because high tax rates will then tend to shrink the source of taxable income. This sensitivity, also called elasticity, might vary substantially across gender, age, and income groups. Another motivation is the need to analyze fluctuations of employment over the economic cycle, which depend on how labor supply reacts to changes in wages.

The first section of this chapter presents some basic facts about labor supply over time and across countries. The second section lays out the principal elements of the neoclassical theory of labor supply. This approach is based on the traditional microeconomic model of consumer choice. The basic model explains the choice between the consumption of products available in the marketplace and leisure. This simple model is then extended in such a way as to take into account household production and intrafamilial decisions. The basic model is also enhanced into a "life-cycle" model that integrates the decisions taken by agents over the course of time. This enhancement is particularly important from the point of view of economic policy, for most employment

policy measures aim to modify the behavior of agents permanently. It also furnishes an adequate framework for analyzing decisions taken from the onset of a career until retirement. The third section of this chapter is devoted to empirical matters. It begins by laying out the main lines of the econometrics of labor supply, elucidates the principles that guide empirical studies in this area, and concludes with a review of the principal quantitative results arrived at by studies of labor supply.

1 FACTS ABOUT LABOR SUPPLY

The amount of time worked, the participation rates of men and women, and the part-time work of women have all undergone significant shifts over the last century.

1.1 Basic Definitions

The labor force (or active population) is made up of all persons who are either employed (whatever the duration of work, salaried or self-employed) or looking for a job (i.e., the unemployed). To be considered unemployed during a reference period, according to the standard ILO (International Labour Organization) definition, people must be (1) without work, that is, not in paid employment or self-employment, (2) currently available for work, and (3) seeking work.

The participation rate (or activity rate) is the ratio of the labor force to a reference population. Most often the reference population is the working-age (15–64) population. But other groups are often considered, such as persons 15 and older, which would then include people older than 64 and lead to lower levels of participation and slightly different dynamic patterns.

Note that the participation rate is not the sum of the employment and unemployment rates because according to the standard definitions, the employment rate is the ratio of the number of employed people to the working-age population, while the unemployment rate is the ratio of the number of unemployed people to the labor force.

1.2 The Trend in the Amount of Time Worked

The long-term trend in the amount of time worked illustrates certain important characteristics of labor supply. Table 1.1 shows that labor productivity, which over the long term shapes the trend of real wages, has not stopped growing since the 1870s, though at a pace that varies at different times and in different countries. Production per hour worked is around 15 times greater in 2000 than in 1870 in Germany, France, and Sweden. It has multiplied by (only) 6 in the United States, and 7 in the United Kingdom over the same period, since these two countries had much higher levels of productivity than the others at the end of the nineteenth century. In fact, before the agricultural and industrial revolutions, productivity had varied very little for several centuries. Likewise, down to the industrial revolution, the amount of time worked probably remained stable, coinciding more or less with the hours of daylight. Subsequently, the onset of the industrial revolution saw longer hours: in the factories we sometimes find that people were present at work for up to 17 hours per day. To work for 14 hours was normal, and a working day of 13 hours was considered short (Marchand and Thélot, 1997).

TABLE 1.1
Hours worked annually per person and real hourly wages in the manufacturing sector.

	Time worked per person per year (hr)				
	1870	1913	1938	1997	2011
Germany	2941	2584	2316	1507	1413
United States	2964	2605	2062	1850	1787
France	2945	2588	1848	1603	1476
United Kingdom	2984	2624	2267	1731	1625
Sweden	2945	2588	2204	1629	1644
	Wages per hour				
Germany	100	185	285	1505	1602
United States	100	189	325	586	603
France	100	205	335	1579	1890
United Kingdom	100	157	256	708	871
Sweden	100	270	521	1601	2011

Source: Maddison (1995) for 1870, 1913, and 1938, and OECD data for 1997 and 2011.

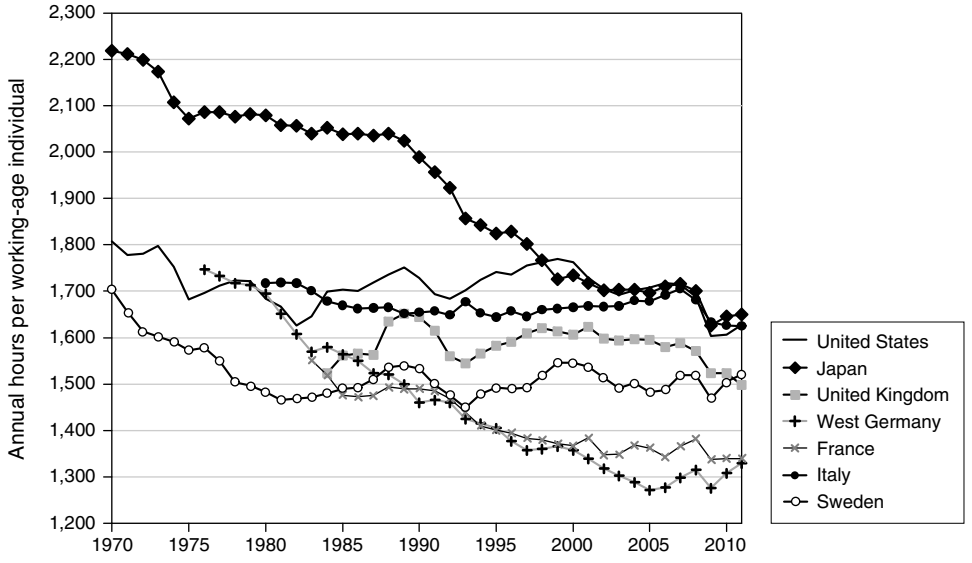


FIGURE 1.1
Amount of time worked annually in 7 OECD countries over the period 1970–2011 (total number of hours worked during the year divided by the average number of persons of working age).

Source: OECD Labor Force Statistics.

Nevertheless, hours worked have undergone shifts less marked, and have differed from one country to another, since the 1970s. In some countries the amount of time worked fluctuates, while in others it continues to shrink overall. Figure 1.1 shows that the annual amount of time worked was stable in Italy, the United States, and Sweden between 1980 and 2008, while it diminished in France, Japan, and Germany. These aggregate figures, which portray the global trend in the amount of time worked, are however difficult to interpret without further ado using the labor supply model (presented in the next section), inasmuch as they result from different composition effects owing to important changes in the structure of the labor force by age and sex that vary from country to country.

1.3 The Evolution of Participation Rates

Figures 1.2, 1.3, and 1.4 trace the evolution of total, male, and female participation rates in the labor markets of the United States, Europe (Germany, France, Italy, and the United Kingdom), and Japan since 1956, for the population aged 15 and older. It is apparent that the participation rate of men has markedly diminished since the beginning of the mid-1950s in Europe and the United States (figure 1.3). For example, it falls 20 points between 1956 and 2010 in the European countries and around 10 points in the United States. On the other hand, the participation rate for women has not stopped growing over the same period, having gained around 11 points in Europe and grown by more than 20 points in North America (figure 1.4). It should be noted that Japan forms an exception to the rule, inasmuch as its participation rates, both male and female,

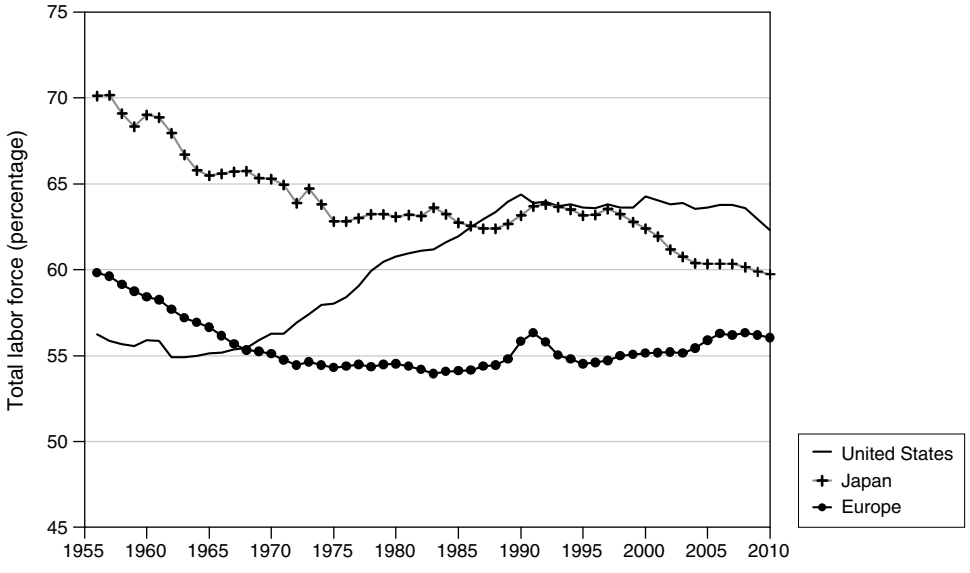


Figure 1.2
The evolution in civilian labor force participation rates in the United States, Europe, and Japan for persons 15 years of age and older, 1956–2010.

Source: OECD Annual Labor Force Statistics.

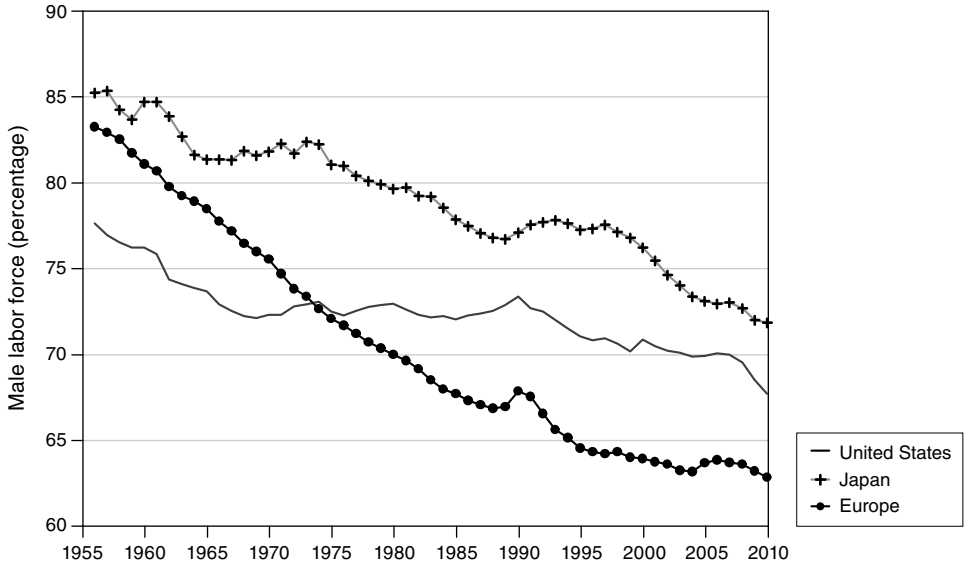


FIGURE 1.3
The evolution in civilian labor force participation rates of men in the United States, Europe, and Japan for persons 15 years of age and older, 1956–2010.

Source: OECD Annual Labor Force Statistics.

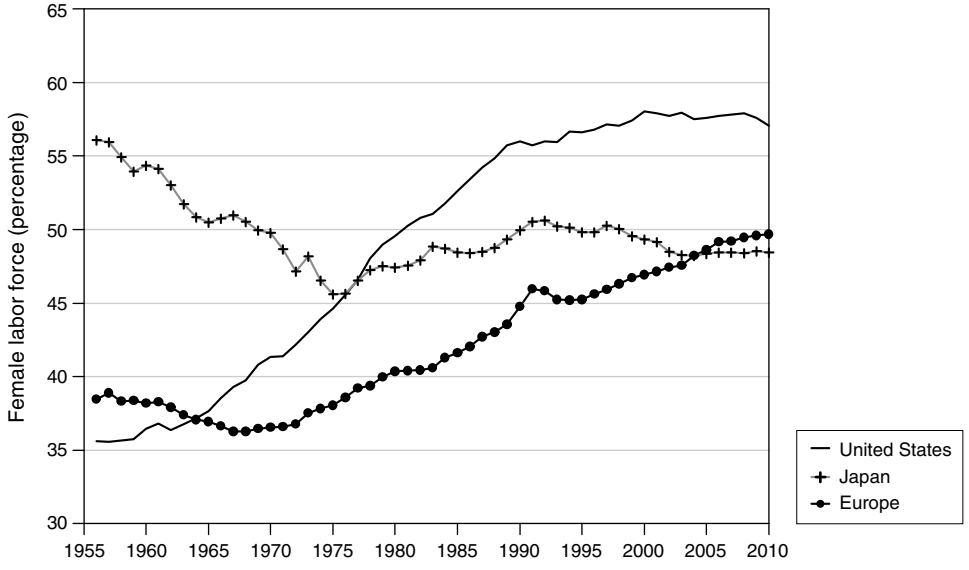


FIGURE 1.4
The evolution in civilian labor force participation rates of women in the United States, Europe, and Japan for persons 15 years of age and older, 1956–2010.

Source: OECD Annual Labor Force Statistics.

declined over this period. The male participation rate decreased by 14 points, while for women it decreased by 7 points, leading to an overall decline in participation rates. We also observe that for the European countries the contrary movements of the male and female participation rates approximately cancel each other out, with a slight decline of 4 points over the whole period. This observation does not apply to North America, where the very strong rise in the female participation rate caused the overall rate of participation to advance until the early 1990s.

Figure 1.4 brings out an important characteristic of the industrialized countries as a group, which is the continuing rise in the participation rate of women for the last several decades. This rise is surely explained by the profound changes in our way of life, but it also corresponds to a steep rise in the wages available to women, as shown in figure 1.5: in the 1970s, about 40% of women in the United States were paid low wages (defined as two thirds of the median gross wage in the economy), compared with 30% in 2010. The decrease is even more dramatic in the United Kingdom and in Japan. Other factors may play a part, such as a fall in the relative price of goods that can replace household work (washing machines, child care, etc.; see Greenwood and Vandenbroucke, 2008). All these factors interact, as we will see later when we present the model with household production.

The increase in participation rates has been particularly steep among married women, at least until the year 2000 in the United States, as shown in table 1.2. Married North American women tend to have a lower rate of participation in the labor market than do single women, even if the difference between these rates has a tendency to diminish over the long term. This can be explained by the fact that married women

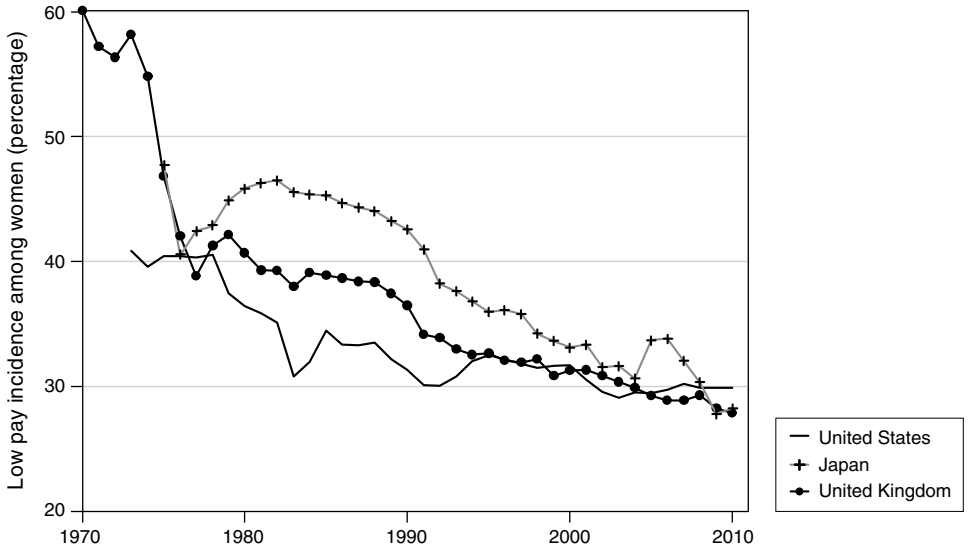


Figure 1.5
The incidence of low-paying jobs among women in the United States, Japan, and the United Kingdom. Low pay is defined as less than two thirds of the gross median earnings of all full-time workers.

Source: OECD Earnings Statistics.

TABLE 1.2
Civilian labor force participation rates of women aged 16 and older, classified by their marital status, in the United States.

	Single	Married
1900	45.9	5.6
1950	53.6	21.6
1988	67.7	56.7
2000	68.9	61.1
2010	63.3	61.0

Source: Ehrenberg and Smith (1994, table 6.1, p. 165) for 1900, 1950, and 1988 and Census Bureau for 2010.

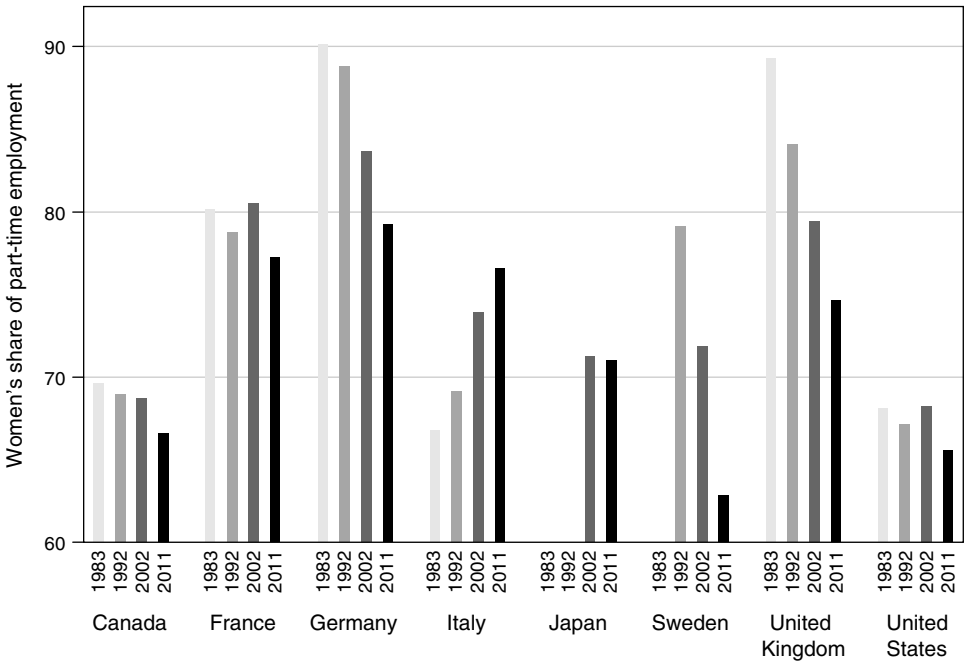


FIGURE 1.6
Women's share of part-time labor (in percentage terms) 1983–2011.

Source: OECD Labor Force Statistics.

also benefit from the earned income of their husbands. Empirical studies generally find that if a husband's income rises, his wife's labor supply falls off.

1.4 PART-TIME WORK BY WOMEN

The rising rate of female employment conceals a sizable difference between them and males: a majority of their jobs are part time. Figure 1.6 indicates that in the majority of the industrialized countries, women's share of part-time work often exceeds 70% (although with a declining trend in most cases). As we will see, several factors might explain this phenomenon. For one thing, for the same amount of work, women's wages are generally noticeably lower than men's. We may thus expect that the hours supplied

by women would tend to be fewer than for men. For another thing, married women have another source of income, which we will call "non-earned income," which often corresponds to the income of their husbands. In that case, one interpretation is that they can "afford" to supply fewer hours and hence will be found in part-time jobs more frequently than men. Of course, other factors come into play to explain this state of affairs—in our day, household chores and the raising of children are still most frequently the tasks of women—but the value of women's relative wage must not be left out of account, as we will see in the next section.

1.5 Leisure and Home Production

One might think that the marked decline, over more than 100 years, in the annual duration of work had released large amounts of time for other activities, such as leisure, and that the rise in the rate of female employment must have been accompanied by a profound reorganization of domestic production. Francis and Ramey (2009) have shown that in the United States the average increase in leisure time has in fact been limited. To start with, as figure 1.7 shows, the rise in the number of leisure hours per week for the whole of the population older than 14 years is indeed a reflection of the fall in the average number of hours worked per week. This concerns men primarily. Yet at the same time men have spent an increasing amount of time on domestic activities, which entails, as figure 1.8 shows, that the amount of time expended on domestic tasks for the whole of the population has not changed much over the century. As regards women, there has indeed been a rise in the average amount of time they spend at work, as manifested by their expanded participation in the labor market, which matches very closely the fall in the amount of time they spend on domestic tasks, as shown in figure 1.9. As well,

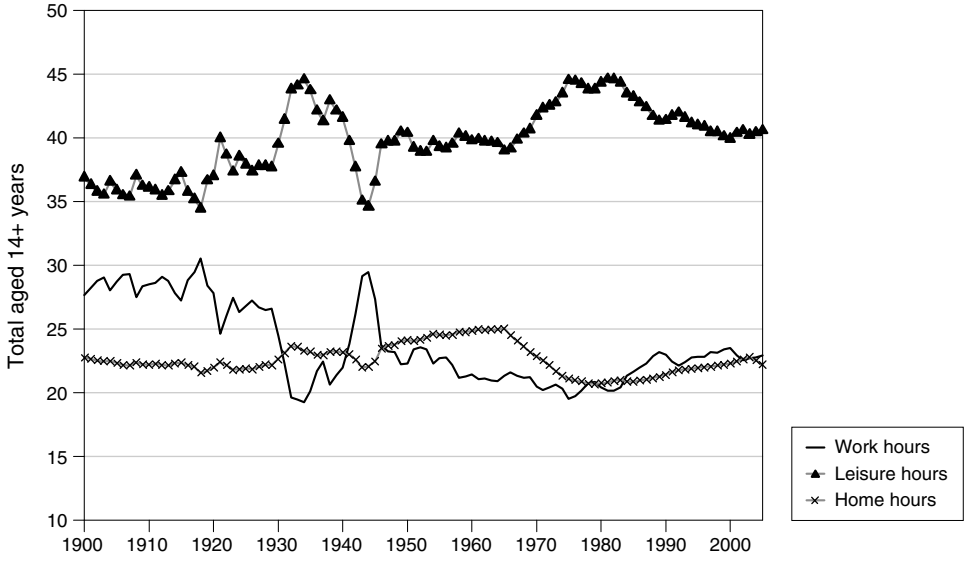


Figure 1.7
Work, leisure, and home hours per week in the United States 1900–2005.

Source: Francis and Ramey (2009).

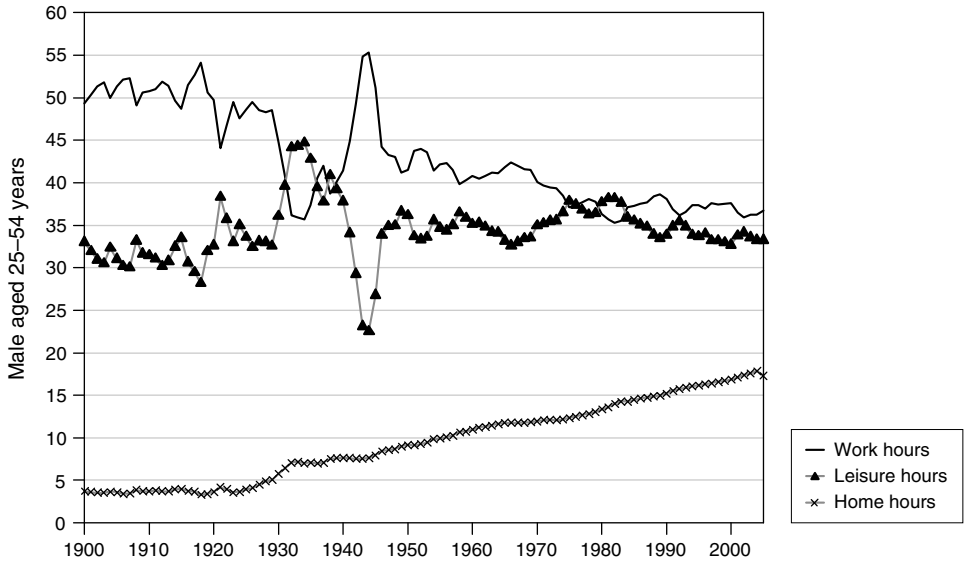


FIGURE 1.8
Work, leisure, and home hours per week of men in the United States 1900–2005.

Source: Francis and Ramey (2009).

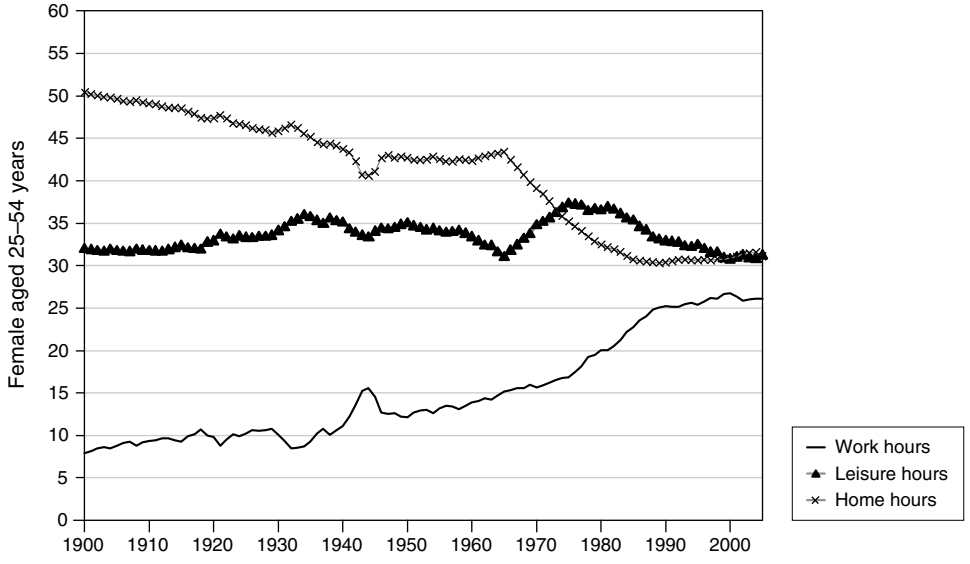


FIGURE 1.9
Work, leisure, and home hours per week of women in the United States 1900–2005.

Source: Francis and Ramey (2009).

hours devoted to education have on average multiplied fourfold over 100 years, which has absorbed much of the reduction in the lifetime duration of work. In total, the rise in life expectancy has certainly allowed men and women to devote more time to leisure over their lifetimes, but the average quantity of leisure per week and the total share of time spent on leisure over lifetimes have not increased substantially.

The model of labor supply we will now present furnishes plausible explanations for many of the factual observations adduced to this point.

2 THE NEOCLASSICAL THEORY OF LABOR SUPPLY

The theory of labor supply is grounded on the model of a consumer making a choice between consuming goods and consuming leisure. With it, we can elucidate the properties of labor supply and begin to understand the conditions of participation in the labor market. The model has been variously enhanced to make the theory of labor supply more precise and sometimes to modify it profoundly, principally by taking into account household production, the collective dimension of decisions about labor supply (most often within the family), and the life-cycle aspect of these decisions.

2.1 The Choice Between Consumption and Leisure

The basic model of a trade-off between consumption and leisure gives us the principal properties of the individual and aggregate supply of labor. In particular, it shows that labor supply is not necessarily a monotonic function of wages. It suggests that labor supply grows at the start, when the wage is low, and subsequently diminishes with the wage when the latter is sufficiently high. Further, the study of the trade-off between consumption and leisure makes it possible to grasp the factors that determine participation in the labor market.

2.1.1 The Basic Model

The traditional approach to labor supply arises, fundamentally, out of the idea that each of us has the possibility to make trade-offs between the consumption of goods and the consumption of leisure, this last being defined as time not spent at work. The analysis of this choice makes it possible to pinpoint the factors that determine labor supply, first at the individual, then at the aggregate, levels.

Preferences

The trade-off between consumption and leisure is shown with the help of a utility function proper to each individual, that is $U(C, L)$, where C and L designate, respectively, the consumption of goods and of leisure. Given that an individual disposes of a total amount of time, L_0, the length of time worked, expressed for example in hours h, is then

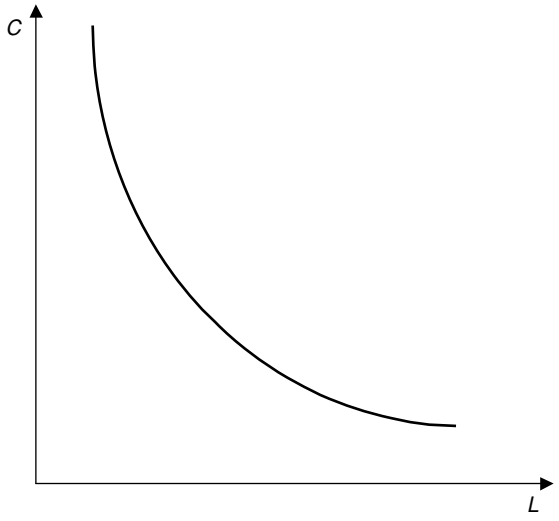


FIGURE 1.10
An indifference curve, where C = consumption of goods and L = leisure.

given by $h = L_0 - L$. It is generally supposed that an individual desires to consume the greatest possible quantity of goods and leisure; her utility function therefore increases with each argument. Moreover, the same individual is capable of attaining the same level of satisfaction with much leisure and few goods, or little leisure and many goods. The set of pairs (C, L) by which the consumer obtains the same level of utility $\overline{U}$, such that $U(C, L) = \overline{U}$, is called an *indifference curve*. A curve of this type is shown in figure 1.10. Its properties follow directly from those of the utility function (for more detail, consult Varian, 1992; Mas-Colell et al., 1995). In particular, the properties listed below will be useful for what follows:

- Each indifference curve corresponds to a higher level of utility, the farther out the curve is from the origin. Hence the consumer will prefer indifference curves situated farther out from the origin.
- Indifference curves do not intersect. If they did, the point of intersection would correspond to a combination of leisure and consumption through which the individual would have two different levels of satisfaction. Such incoherence in preferences is excluded.
- The increase of the utility function in relation to each of its components implies that the indifference curves are negatively sloped (see appendix 7.1 at the end of this chapter). The slope of an indifference curve at a given point defines *the marginal rate of substitution between consumption and leisure*. It represents the quantity of goods that a consumer must renounce in exchange for an hour of supplementary leisure for her level of satisfaction to remain unchanged.

- It is assumed that the individual is ready to sacrifice less and less consumption for an extra hour of leisure when the amount of time dedicated to leisure rises. This property signifies that the marginal rate of substitution between consumption and leisure diminishes with leisure time, or again that the indifference curves are convex, which is equivalent to the hypothesis of the quasi-concavity of the utility function (the relation between the shape of the indifference curves and the utility function is studied in appendix 7.1 at the end of this chapter).

Choices

An individual's income derives from his activity as wage earner and from his activity (or inactivity) outside the labor market. If we designate the real hourly wage by w, the income from wages totals wh. Investment income, transfer income, even gains deriving from undeclared or illegal activities are examples of what an individual may acquire outside the labor market. We will designate the set of these resources expressed in real terms by the single scalar R.

Note that for a married or cohabiting person, a part of her partner's income can be integrated into this set. Thus the budget constraint of the agent takes the form:

$$C \leq wh + R$$

This constraint is also expressed in the following manner:

$$C + wL \leq R_0 \equiv wL_0 + R \tag{1.1}$$

In this way we arrive at the standard concepts of the theory of the consumer. The fiction is that the agent disposes of a potential income R_0 obtained by dedicating his entire endowment of time to working, and that he buys leisure and consumer goods using this income. From this point of view, the wage appears to correspond equally to the *price* and the *opportunity cost* of leisure. The solution of the consumer's problem then follows the path of utility optimization subject to the budget constraint. We thus derive the functions of demand for consumer goods and leisure (for more details, see the microeconomics textbooks by, for example, Varian, 1992; Mas-Colell et al., 1995; Cowell, 2006). The program of the consumer is expressed:

$$\max_{\{C,L\}} U(C,L) \quad \text{subject to the budget constraint} \quad C + wL \leq R_0$$

We begin by studying the "interior" solutions, such as $0 < L < L_0$ and $C > 0$.

The Interior Solutions

For an interior solution, the consumer puts forth a strictly positive supply of labor. Using $\mu \geq 0$ to denote the Lagrange (or Kuhn and Tucker) multiplier associated with the budget constraint, the Lagrangian $\mathcal{L}$ of this program is:[2]

$$\mathcal{L}(C,L,\mu) = U(C,L) + \mu\,(R_0 - C - wL)$$

[2]Mathematical appendix A at the end of this book summarizes what it is necessary to know to solve a static optimization problem.

Designating the partial derivatives of the function U by U_L and U_C, the first-order conditions are expressed as:

$$U_C(C,L) - \mu = 0 \quad \text{and} \quad U_L(C,L) - \mu w = 0$$

On the other hand the complementary-slackness condition is expressed as:

$$\mu\,(R_0 - C - wL) = 0 \quad \text{with } \mu \geq 0$$

This relation and the hypothesis that the utility function increases with each of its components imply that the budget constraint is binding, since the first first-order condition is equivalent to $\mu = U_C(C,L) > 0$. Thus, the solution is situated on the budget line of equation $C + wL = R_0$. We obtain the optimal solution (C^*, L^*) by using this last equality and eliminating the Kuhn and Tucker multiplier μ of the first-order conditions, so that:

$$\frac{U_L(C^*,L^*)}{U_C(C^*,L^*)} = w \quad \text{and} \quad C^* + wL^* = R_0 \tag{1.2}$$

Figure 1.11 proposes a graphic representation of this solution. It shows that the optimal solution is situated at a tangency point between the budget line AB, whose slope is w, and the indifference curve corresponding to the level of utility obtained by the consumer. For the comparative statics of the model, it is worth noting that any

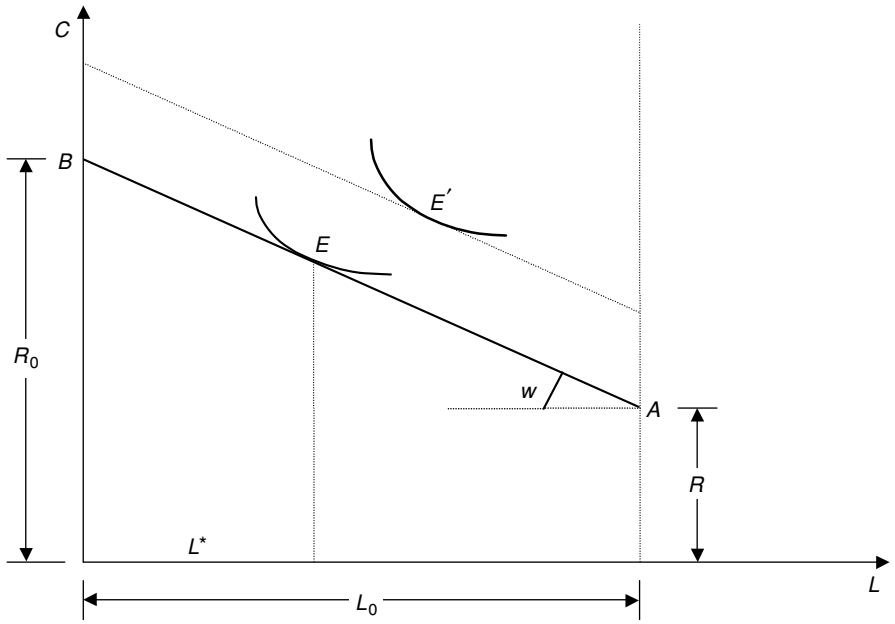


FIGURE 1.11
The trade-off between consumption C and leisure L.

increase in w results in a clockwise rotation of the line AB around point A, of abscissa L_0 and of ordinate R, and that a rise in non-earned income corresponds to an upward shift of this budget line.

The Reservation Wage

For relation (1.2) actually to describe the optimal solution of the consumer's problem, point E has to lie to the left of point A; otherwise labor supply is null ($L = L_0$). Now, the convexity of indifference curves implies that the marginal rate of substitution between consumption and leisure, U_L/U_C, decreases as one moves to the right along an indifference curve (see appendix 7.1 at the end of this chapter).

Since this marginal rate of substitution also represents the slope of the tangent to an indifference curve, an agent offers a strictly positive quantity of hours of work if and only if the following condition is met:

$$\left(\frac{U_L}{U_C}\right)_A < w$$

The marginal rate of substitution at point A is called the *reservation wage*. It is thus defined by:

$$w_A = \frac{U_L(R, L_0)}{U_C(R, L_0)} \tag{1.3}$$

According to this model, assuming that the allocation of time L_0 designates an invariable physical quantity, the reservation wage depends only on the form of the function U at point A and on the value R of non-earned income. It determines the conditions of *participation* in the labor market. If the current wage falls below it, the agent does not supply any hours of work; we then say that she is not participating in the labor market. The decision to participate in the labor market thus depends on the reservation wage. Hence its determinants deserve special attention. In this model, setting aside any change in the consumer's tastes, the only parameter capable of modifying the reservation wage is non-earned income R. If, with respect to this last variable, we derive the equation (1.3) that defines the reservation wage, we can easily verify that the latter rises with R if, and only if, leisure is a *normal*[3] good (one, that is, the consumption of which increases with a rise in income). In these conditions, an increase in non-earned income increases the reservation wage and thus has a *disincentive* effect on entry into the labor market.

2.1.2 The Properties of Labor Supply

The properties of the supply of individual labor result from the combination of a substitution effect and two income effects. The combination of these effects evidently leads to a nonmonotonic relation between wages and the individual supply of labor. We will see as well that by starting with individual decisions and taking into account the heterogeneity of individuals, we will be able to grasp the factors that determine the collective supply of labor.

[3] In deriving (1.3) with respect to R, we find that $\mathrm{d}w_A/\mathrm{d}R$ is of the sign of $(U_{LC}U_C - U_{CC}U_L)$. In appendix 7.2, we show that this expression is positive if and only if leisure is a normal good.

Substitution Effect and Income Effect

For an interior solution, the demand for leisure L^* is implicitly defined by relations (1.2). It is a function of the parameters of the model, which can conveniently be written in the form $L^* = \Lambda(w, R_0)$. The corresponding labor supply, $h^* = L_0 - L^*$, is often called the "Marshallian," or "uncompensated," labor supply. The impact of an increase in non-earned income R on time given over to leisure is indicated by the partial derivative of the function $\Lambda(w, R_0)$ with respect to its second argument, $\Lambda_2(w, R_0)$. It may be positive or negative. By definition, leisure is a *normal good* if its demand rises with R_0 (see appendix 7.2 to this chapter). In the opposite case, in which the time dedicated to leisure decreases with non-earned income, leisure is an *inferior good*. The consequences of an increase in non-earned income are represented in figure 1.11 by the shift from point E to point E'.

The impact of a variation in wages is obtained by differentiating function $\Lambda(w, R_0)$ with respect to w. Taking account of the fact that $R_0 = wL_0 + R$, we arrive at:

$$\frac{dL^*}{dw} = \Lambda_1 + \Lambda_2 \frac{\partial R_0}{\partial w} \quad \text{with} \quad \frac{\partial R_0}{\partial w} = L_0 > 0 \tag{1.4}$$

Figure 1.12 traces the movement of the consumer's equilibrium when wages go from a value of w to a value of $w_1 > w$. The partial derivative of the function Λ with respect to w, denoted Λ_1, corresponds to the usual compound of substitution

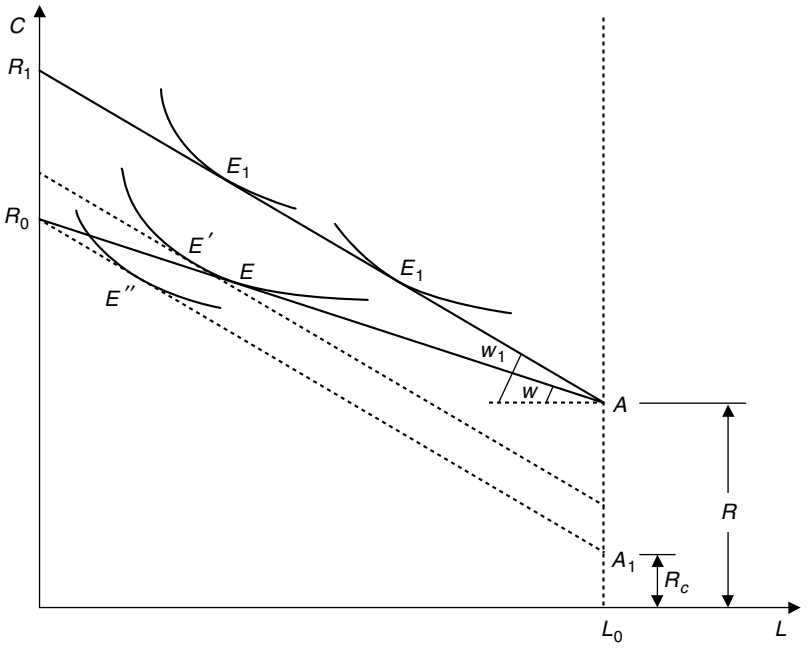


FIGURE 1.12
The effects of a wage increase.

and income effects in the theory of the consumer (the calculations are presented in appendix 7.2). To learn the sign of this derivative, it is best to reason in two stages. In the first stage, we suppose that the potential income R_0 does not change: the consumer then faces a new budget line A_1R_0. For him, it is as though his non-earned income had decreased from R to $R_c = R - (w_1 - w)L_0$. Income R_c is described as *compensated* income and the line A_1R_0 is called the *compensated* budget constraint. In the second stage, we assume that the potential income grows from R_0 to $R_1 = R + w_1L_0$.

Reckoning first with R_0 as a given, we discover the usual compound of substitution and income effects of the theory of the consumer. When the initial equilibrium lies at point E, the substitution effect moves it to point E' offering the same degree of utility as at E, but with the wage now worth w_1 (at point E' the tangent to the indifference curve is parallel to the budget line A_1R_0). The shift from point E to point E' corresponds to a "Hicksian," or "compensated," modification of the labor supply, obtained by minimizing the outlay of the consumer under the constraint of reaching a given level of utility. The substitution effect thus implies a reduction of leisure. Starting from point E' and assuming that the wage keeps the value w_1, the income effect shifts the equilibrium of the consumer to point E''. If leisure is a normal good, the shift from E' to E'' being the consequence of a fall in income, the demand for leisure must diminish. Thus, the substitution effect and the (indirect) income effect work to produce the same result: an increase in wage leads to a diminution of the time allotted to leisure, in other words, an increase in labor supply. Consequently, in relation (1.4) we will have $\Lambda_1 < 0$ if leisure is a normal good. Finally, the increase in potential income from R_0 to R_1 causes the equilibrium to shift from point E'' to point E_1. What we have is a *direct* income effect identified by the partial derivative Λ_2 of the demand for leisure with respect to R_0 in relation (1.4). If leisure is a normal good, then by definition Λ_2 is positive and any rise in wage leads to a rise in the consumption of leisure, and thus to a fall in labor supply. This direct income effect runs counter to the usual substitution and "indirect" income effects of the theory of the consumer. In sum, a wage increase has an ambivalent effect on labor supply. In figure 1.12 the abscissa of point E_1 can as easily lie to the left as it can to the right of that of E.

For convenience, we can aggregate the two income effects by retaining only the shift from E' to E, in which case we refer to the *global* income effect. This allows us to analyze a rise in the hourly wage with the help of only two effects. In the first place, there is an incentive to increase labor supply, since this factor is better remunerated (the substitution effect). But equally there is an opportunity to consume the same quantity of goods while working less, which motivates a diminution of labor supply (the global income effect) if leisure is a normal good.

Compensated and Noncompensated Elasticity of Labor Supply

Along with the Marshallian supply of labor h^* considered to this point, we can also make use of the Hicksian supply of labor; it is arrived at by minimizing the consumer's expenditure, given an exogenous minimal level of utility $\overline{U}$. The Hicksian supply of labor, denoted $\hat{h}$, is then the solution of the problem:

$$\min_{(L,C)} C + wL \quad \text{subject to constraint } U(C,L) \geq \overline{U}$$

The Marshallian supply depends on the wage and on non-earned income, whereas the Hicksian supply of labor depends on the wage and on the level of utility $\overline{U}$. The Hicksian elasticity of labor supply, defined by $\eta_H = (w/\hat{h})(\mathrm{d}\hat{h}/\mathrm{d}w)$, represents the percentage of variation of the Hicksian supply of labor that follows from a 1% rise in wage. It corresponds to the variation in labor supply for a shift from point E to point E' in figure 1.12. Hicksian elasticity is called "compensated" elasticity because it posits that the income of the consumer varies in order for him to stay on the same indifference curve. The Marshallian elasticity of labor supply, defined by $\eta_M = (w/h^*)(\mathrm{d}h^*/\mathrm{d}w)$, represents the percentage of variation of the Marshallian supply of labor that follows from a 1% rise in wage. It corresponds to the variation in the labor supply for a shift from point E to point E_1 in figure 1.12. Marshallian elasticity is also called noncompensated elasticity because it takes into account the variation in real income resulting from the variation in wages.

Marshallian and Hicksian elasticities are linked by the Slutsky equation, which is written thus:

$$\eta_M = \eta_H + \frac{wh^*}{R_0}\eta_{R_0} \tag{1.5}$$

A demonstration of this equality is presented in appendix 7.3 at the end of this chapter. The Slutsky equation shows that Marshallian elasticity is to be interpreted as the sum of two effects. The substitution effect, represented by the Hicksian elasticity η_H, is necessarily positive (the supply of labor increases with wages due to the substitution effect because the demand for leisure decreases when the wage increases). The (global) income effect, represented by the term $\frac{wh^*}{R_0}\eta_{R_0}$, is negative if leisure is a normal good (which means that the supply of labor decreases with wages due to the income effect).

The Shape of the Labor Supply Curve

Figure 1.13 shows a plausible graph of labor supply. When the hourly wage rises just above the reservation wage, the substitution effect prevails over income effects, and labor supply grows. But the global income effect swells with the wage, and it is reasonable to believe that when the latter reaches a certain level it will dominate the substitution effect. The supply of labor then begins to shrink. This is the reason that it is generally thought to turn down, as shown in figure 1.13.

Supplementary Constraints

The preceding analysis leaves out of account many elements that may play a part in the trade-off between work and leisure. For example, the budget constraint is actually piecewise linear, since on one hand overtime hours are not remunerated at the same rate as normal ones, and on the other hand income tax is progressive. This constraint may even present nonconvexities related to the ceilings on various social security contributions. Neither does the model hitherto presented take into account the fact that most often the decision to take a job entails a fixed cost independent of the number of hours worked, such as, for example, the purchase of a second vehicle or the cost of child care. All these elements pose serious problems for empirical assessment (see section 3.1).

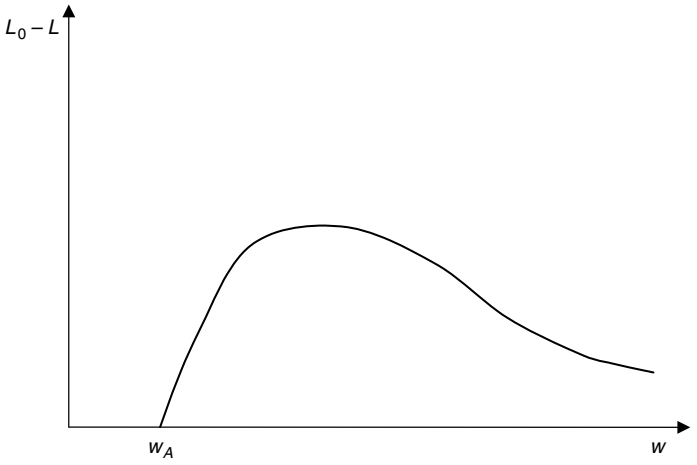


Figure 1.13
The individual labor supply.

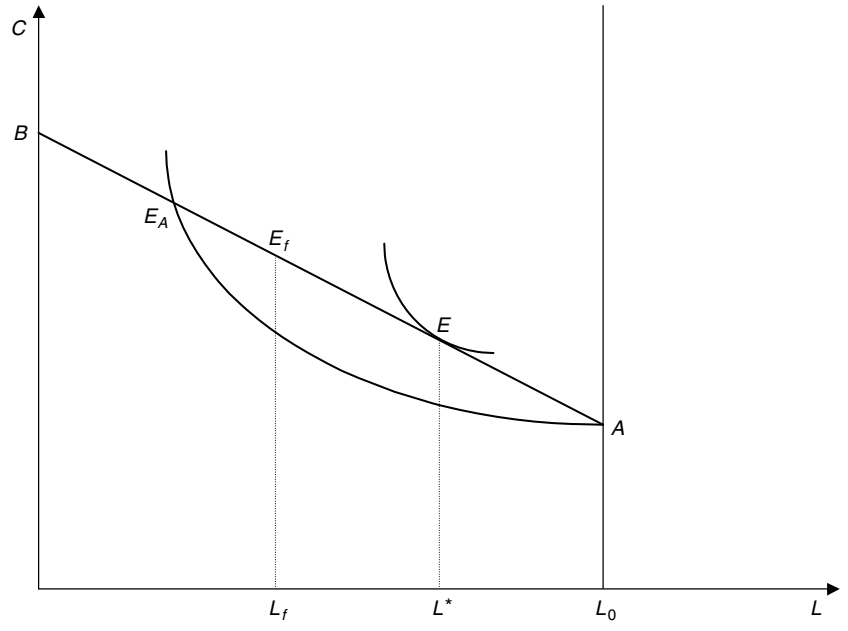


Figure 1.14
Constraint on hours of work.

Another element that may alter the foregoing analysis comes from the relative absence of freedom of choice in the number of hours worked. The majority of wage earners hold full-time employment, other workers hold part-time jobs, but the reality is always a far cry from a hypothetical complete flexibility in hours worked. To illustrate the effects of a rigidity constraint on hours worked, we present a situation in figure 1.14

in which the agent has a choice between working during a set period, represented by the abscissa point L_f, or not working at all.

Let us designate by E the nonconstrained optimum of the agent's problem. If this point is situated to the left of E_f, the agent agrees to furnish $(L_0 - L_f)$ hours of work; in this situation, she would simply have liked to work more. Conversely, when point E lies to the right of E_f, she agrees to work the quantity of fixed hours offered if, and only if, the point E_A—corresponding to the intersection of the indifference curve passing through A with the budget line—lies to the left of E_f. In this case, she obtains a level of utility superior to what she would have attained by not participating at all in the labor market. The agent then works more than she would have wished to (since $L^* > L_f$). On the other hand, if the point E_A were to lie to the right of E_f, she would choose not to participate, since she would have preferred to supply $(L_0 - L^*) > 0$ hours of work. This individual is in a situation that we may call "involuntary nonparticipation," since she does wish to supply a certain quantity of work at the current wage and faces constraints that keep her from supplying them. The abscissa and the ordinate of point E_f being equal respectively to L_f and $w(L_0 - L_f) + R$, the reservation wage—which we will still denote w_A—is defined by the equality:

$$U\left[R + w_A(L_0 - L_f), L_f\right] = U(R, L_0)$$

Extensive Margin, Intensive Margin, and Aggregate Labor Supply

We arrive at the aggregate labor supply, for a wage level of w, by adding up the total number of hours supplied by each individual. It is habitual to assume that the wage exerts two distinct effects on labor supply. In the first place, it influences the decision to work or not; this is called the extensive margin. In the second place, it determines the number of hours supplied by every person who does decide to work; this is called the intensive margin. To evaluate the sensitivity of aggregate labor supply to wages, these two margins both have to be taken into account: variations in the hours of persons who are working (intensive margin) and variations in the number of persons who are working (extensive margin). In order to grasp the impact of wages on choices at the extensive margin, let us consider a large population in which individuals have different non-earned incomes. Let us imagine that this diversity of nonlabor incomes, $R \in [0, +\infty)$, may be represented by a cumulative distribution function $\Phi(\cdot)$. Let us suppose that leisure is a normal good, such that the supply of labor, denoted $h(w, R)$, is a decreasing function of non-earned income. For every wage level w, there then exists a positive value of R, denoted $\bar{R}$ and defined by $h(w, \bar{R}) = 0$, such that only individuals whose non-earned income is inferior to $\bar{R}$ work. The others do not work, since their reservation wage is superior to w. If the size of the total population is normalized to unity, the aggregate labor supply is:

$$L_A(w) = \int_0^{\bar{R}} h(w, R) \mathrm{d}\Phi(R)$$

The derivative of the aggregate labor supply with respect to the wage w is:

$$\int_0^{\bar{R}} \frac{\partial h(w, R)}{\partial w} \mathrm{d}\Phi(R) + h(w, \bar{R})\Phi'(\bar{R}) \frac{\mathrm{d}\bar{R}}{\mathrm{d}w} \qquad \textbf{(1.6)}$$

The first term represents the changes in the intensive margin, which can be either positive or negative, depending on the relative importance of income and substitution effects. The second term, which represents the changes in the extensive margin, is necessarily equal to zero to the extent that, by definition, $h(w, \bar{R}) = 0$. This would signify that small variations in wages have an impact on the aggregate supply of labor, an impact felt solely through changes at the intensive margin. This proposition assumes that it is possible to alter the length of time worked at will. In that case, the contribution of the extensive margin, which concerns durations of work, is negligible. In actuality there exist, as noted above, indivisibilities in the supply of hours of work due to technological and institutional constraints. If persons who decide to work must supply a minimal duration of h_0 hours, then for the right-hand side of equation (1.6) we must substitute the term $h_0\Phi'(\bar{R})\frac{d\bar{R}}{dw}$, where $\bar{R}$ is defined by the relation $U\big(\bar{R} + wh_0, L_0 - h_0\big) = U(\bar{R}, L_0)$. The term $h_0\Phi'(\bar{R})\frac{d\bar{R}}{dw}$ is positive if leisure is a normal good because the definition of $\bar{R}$ implies that $d\bar{R}/dw$ is the inverse of the derivative of the reservation wage with respect to non-earned income.

We will see in the empirical part of this chapter how it is possible to estimate elasticities of labor supply at the intensive and at the extensive margins and how these two elasticities can be combined to estimate aggregate labor supply elasticities.

2.2 Labor Supply with Household Production and Within the Family

The basic model of a trade-off between consumption and leisure neglects numerous elements that may influence labor supply. In this subsection, we extend the model in two important directions. By assimilating time not dedicated to wage labor to leisure, the basic model fails to take account of production within households—production that represents a substitute for income from wages. Furthermore, decisions about labor supply frequently result from bargaining involving several members of the household.

2.2.1 Household Production

The dichotomy between leisure and wage labor masks an important part of the complexity of individual decisions concerning the allocation of time. In reality leisure is not the sole alternative to wage labor. Time devoted to household tasks is (generally) distinguished from leisure. Now these tasks are not always unavoidable. The bulk of the goods and services produced domestically can be purchased. It is possible, for example, to eat a meal that one has prepared oneself, or go to a restaurant, or telephone a caterer, or hire a cook. Clearly each alternative entails a different expense, and an individual's choice depends on his preferences, effectiveness at performing household chores versus doing paid work, income, and prices. We can analyze the consequences of time devoted to household production by modifying our basic model of labor supply at the margin.

The Consumer's Program

Individual preferences are always represented by the utility function $U(C, L)$. Goods consumed may be purchased, in quantity C_M, or produced domestically, in quantity C_D, with $C = C_D + C_M$. The total endowment of time available L_0 breaks down into paid

working time h_M, household working time h_D, and leisure L, hence $L_0 = h_M + h_D + L$. The efficiency of household tasks is represented by a "production function," $C_D = f(h_D)$, linking the amount of the good produced to the time spent on household work. This production function is increasing and concave, thus we will have $f' > 0$ and $f'' < 0$. Income is made up of wage earnings, wh_M, and non-earned income, R. The consumer must choose the quantities C_M, C_D, h_D, h_M, and L, which maximize his utility under the budget constraint $C_M \leq wh_M + R$. Let us further designate potential income as $R_0 = wL_0 + R$; since $h_M = L_0 - h_D - L$, the budget constraint is again written $C_M + wL \leq wh_D + R_0$. Taking into account the identity $C_M = C - f(h_D)$, the consumer's program then takes the following form:

$$\max_{\{C,L,h_D\}} U(C,L) \quad \text{subject to the budget constraint } C + wL \leq [f(h_D) - wh_D] + R_0$$

In this program the choice variables of the consumer are *total* consumption C, leisure L, and the time h_D given over to household production. Additionally, the budget constraint shows that the total income of the consumer is equal to the sum of the potential income R_0 and the "profit" derived from household activities. Since household production only comes into the consumer's program through the expression of this profit, its optimal value h_D^* is that which maximizes the value of this profit; hence it is defined by $f'(h_D^*) = w$. Given time h_D^* dedicated to household activities, the consumer's program becomes formally equivalent to that of the basic model, as long as we replace potential income R_0 by $\tilde{R}_0 \equiv R_0 + f(h_D^*) - wh_D^*$. The optimal solutions $C^* = C_M^* + f(h_D^*)$ and L^* are then defined by the equalities:

$$\frac{U_L(C^*, L^*)}{U_C(C^*, L^*)} = w = f'(h_D^*) \quad \text{and} \quad C^* + wL^* = \tilde{R}_0 \tag{1.7}$$

This result is close to the one described by equation (1.2) in the basic model. At the optimum, the marginal rate of substitution between consumption and leisure is equal to the wage. As previously, this condition describes the division between the consumption of goods and that of leisure. The equality $f'(h_D^*) = w$ shows that the allocation of working time between household and wage-earning activities is determined by the relative productivities of the two types of activity. Consequently, the wage reflects the individual productivity of wage labor. The agent thus has an interest in devoting his working time to household activities to the extent that the marginal productivity $f'(h_D)$ of an hour of this type of work is superior to an hour's wage. Therefore, he augments the length of time given to household work to the point where $f'(h_D^*) = w$.

Elasticity of the Labor Supply

The possibility of making trade-offs between household and wage-earning activities alters the elasticity of the labor supply curve. The system of equations (1.7) allows us to write the optimal demand for leisure in the form $L^* = \Lambda(w, \tilde{R}_0)$. Differentiating this equality with respect to w, we get:

$$\frac{dL^*}{dw} = \Lambda_1 + \Lambda_2 \frac{d\tilde{R}_0}{dw} \quad \text{with} \quad \frac{d\tilde{R}_0}{dw} = L_0 - h_D^*$$

As $f'(h_D^*) = w$ implies that $\mathrm{d}h_D^*/\mathrm{d}w = 1/f''(h_D^*)$, the identity $h_M^* \equiv L_0 - h_D^* - L^*$ entails:

$$\frac{\mathrm{d}h_M^*}{\mathrm{d}w} = -(\Lambda_1 + \Lambda_2 L_0) + \left[\Lambda_2 h_D^* - \frac{1}{f''(h_D^*)}\right] \tag{1.8}$$

The term $-(\Lambda_1 + \Lambda_2 L_0)$ represents the impact of a variation in the wage on the supply of wage labor for a given amount of household activity. It corresponds to the set of effects discussed in the basic model—see equation (1.2) above and the accompanying remarks. We have seen, in particular, that a change in the wage has an ambivalent impact on labor supply. The second term of the right-hand side of equation (1.8) is positive if leisure is a normal good (that is to say, if $\Lambda_2 > 0$). Consequently, the possibility of making trade-offs against household activity ought to increase the wage elasticity of the labor supply. This result might explain why empirical studies reveal that the wage elasticity of the supply of female labor is generally higher than that of the supply of male labor (see section 3.2.4). For men, the trade-off between household and waged activity is often marginal. An instructive limit case is that of an optimal "corner solution," with a null supply of domestic labor $h_D^* = 0$. This might be the case if the productivity of household work were far below the current wage. A high proportion of men would then trade leisure off against wage labor only, whereas many women, whose household productivity is high in relation to the wage that they could get, would trade off among leisure, household activity, and wage labor.

Taking household activity into consideration allows us to make the predictions of the basic model richer. It should be emphasized, however, that the model presented here remains very rudimentary. For one thing, it rests on the hypothesis of an identical disutility of work for waged and household activities. In reality, the inconvenience arising from these activities is different. A more general approach, proposed by Becker (1965), consists of taking into account the disutility (or the utility) associated with each activity by distinguishing the diverse kinds of work done in the home. Such an approach has the merit of analyzing the choices underlying the allocation of time among different activities with great precision (on this subject, see the syntheses of Gronau, 1986 and 1997).

2.2.2 INTRAFAMILIAL DECISIONS

The family has considerable influence on the behavior of its members. The supply of labor is not exempt from this rule, and the basic model has to be adapted so as to take into account the influence of family structures. The question bears an important empirical aspect, for numerous data (in particular those on consumption) only describe the behavior of the household, so we require a theory that goes beyond the basic individual frame of reference and gets us to a point where our estimates make some sense. The analysis of family choices has developed along two different lines. The first, known as the *unitary model*, starts from the principle that the family can be likened to a sole agent having its own proper utility function. The second, known generically as the *collective approach*, postulates that making choices is fundamentally something individuals do and that the family is no more than a particular framework that enlarges (or constrains) the range of choices of each member of it.

The Unitary Model

This approach extends, as simply as possible, the basic model proposed hitherto. Let us imagine a family composed of two persons: it postulates that the preferences of this entity are representable by a utility function $U(C, L_1, L_2)$, where C represents the *total* consumption of goods by the household and $L_i\,(i = 1, 2)$ designates the leisure of individual i.[4] This formalization assumes that the satisfaction attained through the consumption of a good depends solely on its total amount and not on the manner in which it is shared among the members. For agent i, let us denote her wage and non-earned income respectively as w_i and R_i; the optimal choices are then determined by maximizing utility under a single budget constraint. The program of the household is written:

$$\max_{\{C,L_1,L_2\}} U(C, L_1, L_2) \quad \text{subject to the constraint } C + w_1L_1 + w_2L_2 \leq R_1 + R_2 + (w_1 + w_2)L_0$$

Scrutiny of this program reveals that the unitary representation of the household implies that the *distribution* of non-earned incomes has no importance—the only thing that counts is their *sum* $R_1 + R_2$. This hypothesis, known in the literature as *income pooling*, signifies, for example, that it is not necessary to know which member of the couple is the beneficiary of transfer income. Now the fact is that empirical studies refute this hypothesis for large segments of the population. For example, Fortin and Lacroix (1997) find that the unitary model only fits couples with preschool-age children (see also Blundell and MaCurdy, 1999; Blundell et al., 2007). This invalidation is one of the reasons the unitary model of the household is not completely satisfactory and is giving way to the collective model for the purpose of describing decisions taken within a family.

The Collective Model

The most highly elaborated form of the collective model is due to Chiappori (1988, 1992). This model starts from the principle that household choices must arise out of individual preferences. In making the household the sole locus of decisions, the unitary model arbitrarily aggregates the preferences of its members and hence does not respect the basic principle of "methodological individualism." Conversely, if one does adhere to this principle, it appears natural to assume that decisions taken within a household are efficient in the Pareto sense, meaning that the possibility of mutually advantageous allocation does not occur. If we use $U_i(C_i, L_i)$, $i = 1, 2$, to designate the individual preferences of the persons composing the household, the efficient allocations will be the solutions of the following program:

$$\max_{\{C_1,C_2,L_1,L_2\}} U_1(C_1, L_1) \quad \text{subject to constraints:}$$

$$U_2(C_2, L_2) \geq \overline{U}_2$$

$$C_1 + C_2 + w_1L_1 + w_2L_2 \leq R_1 + R_2 + (w_1 + w_2)L_0$$

[4] A "public good" consumed by the household (children are usually given as the example, since both parents benefit from the presence of the children and the benefit obtained by one of the parents does not reduce that of the other parent as long as both parents live in the same household) is generally added to the arguments of the utility function. It is also possible to integrate the possibility of home production into this framework.

In this program the parameter $\overline{U}_2$ represents a given level of utility, and we may suppose that it depends on the parameters w_i and R_i. Chiappori (1992, proposition 1) then shows that the efficient allocations are also the solutions of *individual* programs in which each person would be endowed with a specific non-earned income and which would depend on the overall income of the household. More precisely, the program of agent i takes the following form:

$$\max_{\{C_i, L_i\}} U_i(C_i, L_i) \quad \text{subject to constraint } C_i + w_i L_i \leq w_i L_0 + \Phi_i$$

where Φ_i is a "sharing rule," depending on the parameters w_i and R_i, and such that $\Phi_1 + \Phi_2 = R_1 + R_2$. In other words, it is as if each member of the household received a fraction of the total non-earned income of the household. In a way, this approach reinforces the basic model of choice between the consumption of goods and leisure by specifying for the budget constraint of the individual the composition of his non-earned income. It is possible to expand the collective model by taking into account the "public" goods pertaining to the household and the household production of its members.

From the empirical point of view, the collective model has the advantage of not adopting the hypothesis of "income pooling" a priori; the latter is no more than a particular case of this model. Moreover, Chiappori (1992) shows that this formulation of the decision-making process within a household allows us to deduct individual consumption—which is not, for the most part, observable—using the individual supplies of labor and the total consumption of the household, which are observable entities. Hence the simple observation of the supplies of labor and individual incomes allows us to determine the sharing rules within households. Knowing these rules and using available data, it becomes possible to assess the consequences of public policies on each member of the household. In this context, Browning et al. (1994) have shown, on the basis of Canadian data, that differences of age and income among the members of households, as well as the wealth of households, appear to affect the sharing rules Φ_i. Lundberg et al. (1997) estimate that which spouse receives the child allowance affects household decisions. Oreffice and Quintana-Domeque (2012) find that relative physical attractiveness, measured by the body mass index, matters for the intrahousehold allocation of resources and, therefore, for the hours worked by both spouses (see Browning et al., 2012, for a survey of bargaining-power measures in collective models).

The Added Worker Effect

Models of intrafamilial choice throw a revealing light on decisions to participate in the labor market. Taking into account the familial dimension does indeed allow us to explain why certain members of the household specialize in household production while others offer their services on the market for wage labor. From whatever angle the household is viewed, the members' choices are interdependent, and an individual's fluctuations in income will have an impact on her own supply of labor, but also on that of the spouse or other members of the household, for example, working-age children. This interdependence of choices may lead an individual to increase her supply of labor when the household income declines. It might even motivate her to participate in the labor market if she was not already doing so before the income fell. In principle, a fall in wages may thus entail an increase in the labor force by spurring additional

workers to enter the market for the precise purpose of making up for their household's loss of income. From the empirical point of view, this *added worker effect* seems to have little weight (see, for example, Lundberg, 1985). It is interesting to note that the added worker effect implies a negative relationship between the participation rate and the average wage. When we constructed the aggregate supply of labor out of individuals making their decisions in isolation, we obtained a positive relationship between the average wage and the participation rate (see section 2.1.2). In practice, this second relationship turns out to be dominant, and we do in fact observe a positive correlation between wages and the participation rate.

2.3 Life Cycle and Retirement

The static models used to this point obviously do not allow us to understand how agents substitute for their consumption of leisure over time when their flow of income undergoes transitory or permanent shocks. Taking into explicit account a succession of periods does not markedly alter the conclusions of the static model, but it does provide an adequate framework to analyze dynamic behaviors, which is useful to understand how labor supply changes over business cycles. The decision to go into retirement, which is the definitive end of participation in the labor market, can also be analyzed suitably using a dynamic model of labor supply within which we have redefined the flow of income and legal constraints.

2.3.1 Intertemporal Labor Supply

The dynamic theory of labor supply gives a central role to the possibility of substituting for the consumption of physical goods and leisure over time. We highlight this possibility using a dynamic model in discrete time. This model likewise allows us to grasp the contrasting effects caused by a transitory change in wages or a permanent modification of the wage profile and, thus, to examine critically certain aspects of the theory of "real business cycles."

A Dynamic Model of Labor Supply

In a dynamic perspective, a consumer must make his choices over a "life cycle" represented by a succession of periods that start with an initial date, conventionally taken as equal to 0, and end with an independent terminal date, annotated T. Assuming that the period t unfolds between the dates $(t-1)$ and t, the succession of periods is then given by the index $t = 0, 1, 2, ..., T$. The index t is also used as an indicator of the age, professional experience, or seniority of an individual, according to the subjects under study. In a very general way, the preferences of the consumer must be represented by a utility function of the form $\mathcal{U}(C_0, ...C_t, ..., C_T; L_0, ..., L_t, ..., L_T)$ where C_t and L_t designate respectively the consumption of physical goods and the consumption of leisure for the period t. But this very general form does not permit us to obtain analytically simple and easily interpretable results. That is why it is often assumed that the utility function of the consumer is temporally separable, in which case it is written $\Sigma_{t=0}^{t=T} U(C_t, L_t, t)$. Under this hypothesis, the term $U(C_t, L_t, t)$ represents simply the utility obtained by the consumer in the course of period t. It is sometimes called the "instantaneous" utility

of the period t. We must bear in mind, however, that this representation of preferences is very restrictive: in particular, it does not allow us to take into account the inertia of habits of consumption, or habit persistence, that empirical studies reveal (see Hotz et al., 1988). To bring out this phenomenon, the influence of past consumption on the utility of the current period would have to be incorporated. Another important limitation of the model presented here has to do with the absence of decisions about training. Training increases the human capital of an individual and raises his wage-earning prospects, so there must be trade-offs among leisure, working time, and time dedicated to training (see Keane, 2011, and chapter 4, section 2 in this book).

In this dynamic model, we will assume that individuals have the opportunity to save, and we will use r_t to denote the real rate of interest between the periods $t-1$ and t. For each period, the endowment of time is an independent constant to which we will give the value 1 to simplify the notation. On this basis, the hours worked during a period t are equal to $(1-L_t)$. If we use A_t to designate the consumer's assets on date t and B_t to designate her income apart from wages and the yield on savings on the same date, with $A_{-1}=0$ so that B_0 stands for the initial wealth, the evolution of the assets of the consumer is described by:

$$A_t = (1+r_t)A_{t-1} + B_t + w_t(1-L_t) - C_t, \quad \forall t \geq 0 \tag{1.9}$$

This equation can easily be understood as follows: at each period t, the increase in wealth $A_t - A_{t-1}$ is due to income $w_t(1-L_t)$ from wage labor, to income r_tA_{t-1} from savings, and to other income B_t. Consumption C_t for the period has to be deducted from these gains. The non-earned income R_t for the period t is thus equal to $B_t + r_tA_{t-1}$.

Optimal Solutions and Demands in Frisch's Sense

The consumer attempts to maximize his intertemporal utility subject to the budget constraint described, on each date, by equation (1.9). If we use ν_t to denote the multiplier associated with this equation, the Lagrangian of the consumer's problem takes the form:

$$\mathcal{L} = \sum_{t=0}^{t=T} U(C_t, L_t, t) - \sum_{t=0}^{t=T} \nu_t \left[A_t - (1+r_t)A_{t-1} - B_t - w_t(1-L_t) + C_t\right]$$

The first-order conditions are obtained by equating the derivatives of this Lagrangian to zero with respect to variables C_t, L_t, and A_t. After a few simple calculations, we arrive at:

$$U_C(C_t, L_t, t) = \nu_t \quad \text{and} \quad U_L(C_t, L_t, t) = \nu_t w_t \tag{1.10}$$

$$\nu_t = (1+r_{t+1})\nu_{t+1} \tag{1.11}$$

Relations (1.10) imply $U_L/U_C = w_t$. The equality between the marginal rate of substitution and the current wage is thus maintained at every date, but this result is not general; it is a direct consequence of the hypothesis of the separability of the utility

function. Limiting ourselves to interior solutions, the optimal consumptions of physical goods and leisure are implicitly written in the following manner:

$$C_t = C(w_t, \nu_t, t) \quad \text{and} \quad L_t = L(w_t, \nu_t, t) \tag{1.12}$$

The supply of labor at date t is then defined by $h(w_t, \nu_t, t) = 1 - L(w_t, \nu_t, t)$.

Equation (1.12) shows that the supply of labor at date t depends on the current wage and the multiplier ν_t, which is the marginal utility of wealth.[5] Additionally, equation (1.11), which is known as the Euler equation, shows that the multiplier ν_t depends solely on the interest rate and on the initial value ν_0. More precisely, successive iterations of the logarithms of equation (1.11) entail:

$$\ln \nu_t = -\sum_{\tau=1}^{\tau=t} \ln(1 + r_\tau) + \ln \nu_0 \tag{1.13}$$

This way of writing the law of motion of ν_t proves extremely interesting from the empirical point of view, since it shows that ν_t can be broken down into a fixed individual effect ν_0 and an age effect $-\sum_{\tau=1}^{\tau=t} \ln(1 + r_\tau)$ common to all agents (see subsection 3.1 on the econometrics of the labor supply). Introducing uncertainty into this model, for example concerning wages, does not change the essential results markedly. We can verify that the first-order conditions (1.10) remain true, whereas the marginal utility of wealth ν_t becomes a random variable, following a stochastic process described by equation (1.13), with an error term with zero average appearing on the right-hand side of this equation (see Blundell and MaCurdy, 1999).

A priori, the value of ν_0 depends on all the wages received by an individual during his lifetime. If we want to estimate the effects of a modification of the wage profile—and not just those due to a change in the current wage—then we have to take into account the dependence of ν_0 on all wages. From this perspective, we see that this model is useful for distinguishing between the impact of *temporary* variations in the wage and the impact of *permanent* wage variations.

Frischian, Hicksian, and Marshallian Elasticities of Labor Supply

In the intertemporal model, three types of elasticity are most often distinguished. The Frischian elasticity represents the impact of a modification of the wage at date t on the supply of labor on the same date, assuming that the marginal utility of wealth, represented by the multiplier ν_t in the first-order conditions (1.10), remains constant. This elasticity thus describes the reaction to a change in the current wage, assuming that the marginal utility of wealth is constant. In this sense, Frischian elasticity measures a phenomenon of intertemporal substitution: it indicates by how much we are willing to alter the amount of time worked today when today's wage varies, knowing that the marginal utility of our wealth is unchanged. This elasticity is useful for measuring the impact of a transitory wage variation, which has a negligible impact on wealth.

Marshallian elasticity measures the total impact of a wage variation on labor supply, taking into account variability in the marginal utility of wealth. Finally, Hicksian

[5] The interpretation of the Lagrange multipliers is presented in appendix A3 at the end of this book.

elasticity measures the variation in labor supply, on the assumption that the level of intertemporal utility remains constant.

In the static case, the Marshallian (η_M) and Hicksian (η_H) elasticities are linked by the Slutsky relation (1.5). It is shown in appendix 7.4 that this relation holds good in the dynamic model, where the potential income R_0 [defined in equation (1.1)] of the static model is replaced by present intertemporal wealth, denoted Ω, which is defined by:

$$\Omega = \sum_{t=0}^{T} (1 + r_t)^{-t} (w_t + B_t)$$

Since the Slutsky relation (1.5) holds good in the dynamic model, we always have $\eta_H \geq \eta_M$. Appendix 7.4 also demonstrates that elasticities in the Frischian, Hicksian, and Marshallian senses are linked. So, for example, when the wage variation concerns only the current period, in other words when the wage at other dates remains unchanged,[6] the relation between the Marshallian and Frischian elasticities, denoted η_F, is written:

$$\eta_M = \eta_F + \frac{wh}{\Omega} \eta_\Omega (1 - \gamma \eta_\Omega) \tag{1.14}$$

In this expression, h and w represent respectively the labor supply and the wage for the current period, and Ω represents present intertemporal wealth. In this setting, η_M and η_Ω designate the elasticities of the labor supply for the current period with respect to the wage for that period and total present wealth, and η_F designates the Frischian elasticity of labor supply for the current period with respect to the current wage. Additionally, we denote $\gamma = -V_\Omega / \Omega V_{\Omega\Omega} > 0$, where V designates the indirect intertemporal utility function, and V_Ω its partial derivative with respect to Ω. Parameter γ corresponds to the elasticity of intertemporal substitution, equal to the inverse of Arrow-Pratt risk aversion. Relation (1.14) shows that the impact of a wage variation on labor supply may be broken down into an intertemporal substitution effect, measured by Frischian elasticity, which assumes a constant marginal utility of wealth, and a wealth effect represented by the term $\frac{wh}{\Omega} \eta_\Omega (1 - \gamma \eta_\Omega)$, which takes account of the impact of the wage variation on the marginal utility of wealth. This wealth effect may itself be broken down into two terms: the first term, $\frac{wh}{\Omega} \eta_\Omega$, comes from the variation in wealth Ω and the second term, $-\gamma \frac{wh}{\Omega} (\eta_\Omega)^2$, results from variation in the price of leisure which modifies the marginal utility of wealth.

It is possible to put the elasticities of Hicks, Marshall, and Frisch into ranked order. We know already that $\eta_H \geq \eta_M$. Using equations (1.5) and (1.14), we obtain a relation between the Frischian and Hicksian elasticities:

$$\eta_F = \eta_H + \gamma \frac{wh}{\Omega} (\eta_\Omega)^2$$

This relation demonstrates that Frischian elasticity is greater than Hicksian elasticity, and that Hicksian elasticity is larger than Marshallian elasticity: $\eta_F \geq \eta_H \geq \eta_M$.

[6] The formula in the case where the wage does vary in other periods is given in appendix 7.4 at the end of the chapter; see relation (1.59).

It is important, though, to note that the differences among these three elasticities arise solely from the existence of income effects. In the absence of income effect, the three elasticities are identical. Such is the case when preferences are quasi-linear, of the form:

$$U(C_t, L_t, t) = (1+\rho)^{-t}\left(C_t + \frac{\eta}{\eta-1}L_t^{\frac{\eta-1}{\eta}}\right) \quad \eta > 1, \quad \rho \geq 0 \tag{1.15}$$

In this case, conditions (1.10) and (1.11) entail that $L_t = w_t^{-\eta}$ and thus that hours worked depend only on the current wage. We will now consider a different case where Marshallian elasticity is null while Frischian elasticity is positive.

Transitory Wage Changes Versus Permanent Wage Changes

The difference, fundamental on the level of economic policy, between a modification of the wage profile and a change in a particular wage, emerges clearly with the help of the following example, taken from Blanchard and Fischer (1989, chapter 7, section 7.2). Let us suppose that the real interest rate is constant ($r_t = r, \forall t \geq 0$), that the consumer is receiving no exogenous income ($B_t = 0, \forall t \geq 0$), and that her instantaneous utility takes the explicit form:

$$U(C_t, L_t, t) = (1+\rho)^{-t}\left(\ln C_t + \frac{\eta}{\eta-1}L_t^{\frac{\eta-1}{\eta}}\right) \quad \eta > 1, \quad \rho \geq 0$$

The constant factor ρ represents the psychological discount rate. The Frischian demand functions are then written:

$$C_t = \frac{1}{(1+\rho)^t \nu_t} \quad \text{and} \quad L_t = \left[\frac{1}{(1+\rho)^t \nu_t w_t}\right]^{\eta}$$

We may note that the elasticity in Frisch's sense is equal, in absolute value, to the constant coefficient η. With a constant interest rate, the Euler equation (1.11) then gives $\nu_t = \nu_0/(1+r)^t$, and the demand functions are expressed as a function of ν_0 in the form:

$$C_t = \frac{1}{\nu_0}\left(\frac{1+r}{1+\rho}\right)^t \quad \text{and} \quad L_t = \left[\frac{1}{\nu_0 w_t}\left(\frac{1+r}{1+\rho}\right)^t\right]^{\eta} \tag{1.16}$$

To obtain an implicit equation giving the value of ν_0, we have to write the intertemporal budget constraint of the consumer. This constraint is arrived at by eliminating assets A_t through successive iterations of the accumulation equation (1.9). With $r_t = r$ and $B_t = 0$ for all $t \geq 0$, we arrive at:

$$\sum_{t=1}^{T}(1+r)^{-t}(C_t + w_t L_t) = \sum_{t=1}^{T}(1+r)^{-t} w_t \tag{1.17}$$

This expression generalizes the budget constraint (1.1) of the static model: it states that the discounted present value of expenditure for the purchase of consumer goods and leisure cannot exceed the discounted present value of global income.

The value of ν_0 is obtained by bringing the expressions of C_t and L_t given by (1.16) into the intertemporal budget constraint (1.17). It is implicitly defined by the following equation:

$$\sum_{t=1}^{T}(1+\rho)^{-t}\left\{1+\left[\left(\frac{1+r}{1+\rho}\right)^{-t}\nu_0 w_t\right]^{1-\eta}-\left(\frac{1+r}{1+\rho}\right)^{-t}\nu_0 w_t\right\}=0 \qquad (1.18)$$

It emerges clearly that the multiplier ν_0 depends on all wages over the individual's life cycle. For sufficiently large T this multiplier is affected very little by changes in a particular wage: what we have in that case is a transitory shock. On the other hand, it is affected by a change that affects all wages: what we have then is a modification of the wage profile, or a permanent shock. To grasp clearly the difference between these two types of shock, let us imagine that a permanent shock corresponds to a multiplication of all wages by a single positive quantity; relation (1.18) shows that ν_0 will be divided by this quantity. But relation (1.16) then indicates that the optimal level of leisure—and therefore that of hours worked—remains unchanged. In this model, a permanent shock has no influence on labor supply, since the income effect and the substitution effect cancel each other out. Let us now consider a transitory shock that causes only the wage w_t to change. This shock has only slight influence on the value of ν_0, and relation (1.16) shows that leisure at date t diminishes, while leisure at all other dates remains unchanged. This particular model thus succeeds in conveying the notion that the permanent component of the evolution of real wages has no effect on labor supply, whereas the transitory component affects the level of supply immediately through the optimal response of agents who adjust their supply of labor in response to *temporary* changes in the wage.

Labor Supply and the Business Cycle

Since the first publications of Lucas and Rapping (1969), a number of authors have studied changes in the labor supply as a function of movements in the real wage. The goal of these studies is to explain a striking fact of major importance, which is that aggregate employment fluctuates a great deal in the course of a cycle, while the transitory component of changes in the real wage proves limited in scope. At the outset, the theory referred to as that of "real business cycles" saw the mechanism of intertemporal substitution of leisure as the principal cause of fluctuations in the level of employment. According to this train of thought, the economy is always the object of multiple shocks (on technology or on preferences) that have repercussions on the remuneration of labor and capital; agents respond to these shocks in an optimal manner by instantaneously adjusting their supply of labor. More precisely, a favorable shock, one perceived as transitory, would motivate agents to increase their supply of labor today and to reduce it tomorrow when the shock has passed (for a comprehensive evaluation of the implications of the theory of real business cycles for the labor market, see Hall, 1999; Shimer, 2010). This theory is simple, even seductive, but it runs up against a sizable obstacle. If it is to agree with empirical findings, it must explain how *small* movements in the real wage could entail *large* variations in the level of employment.

Hence in its original version, the theory of real business cycles requires employment to be very sensitive to small changes in the wage. Relation (1.16) shows that this

will be the case if the absolute value of the intertemporal elasticity of substitution of leisure η is large. Now the majority of empirical studies arrive instead at small values. Hall (1980) estimates that a value of 0.4 might apply at the macroeconomic level; Pencavel (1986) suggests values even lower than that for men, while Blundell et al. (1993) find levels ranging from 0.5 to 1 for married women in the United Kingdom. In these circumstances, variations in the labor supply in response to transitory changes in the wage cannot serve as a sufficient basis for a theory of the business cycle. Relation (1.16) does indicate, however, that transitory shocks might influence the level of employment via interest rates. Since these variables are noticeably more volatile than wages, there would thus be another way to reproduce the stylized facts in question. This trail, however, also comes to a dead end. To demonstrate this, let us suppose that the intertemporal utility function of the consumer is temporally separable; the first-order conditions (1.10) then imply:

$$\frac{u_L(C_t, L_t, t)}{u_C(C_t, L_t, t)} = w_t \quad \forall t = 1, \ldots, T$$

If the wage does not change, it can easily be verified that this expression defines an increasing relation between consumption and leisure if these are normal goods. In this case, movements in labor supply supposedly due to the variability of interest rates alone would be accompanied by an *inverse* movement of consumption. Here too we run up against contradictory empirical observations, which show a positive correlation between levels of employment and consumption. Faced with this fresh setback, one might try out other modifications of the formulation of the problem of the trade-off over time between consumption and leisure, like, for example, giving up the hypothesis of separability or introducing fixed costs into the decision to participate. To this day, no way has really been found to escape the substantially negative verdict that hangs over explanations of variability in employment based on the sole mechanism of intertemporal substitution of leisure (see the discussion in section 3.2.3).

2.3.2 Economic Analysis of the Decision to Retire

Economic analysis of the process by which a person terminates his labor market participation fits well into the life-cycle model offered above, provided that legal constraints and the flow of income specific to retirement are brought into clear focus. In an uncertain environment, the process of making this decision can be analyzed with the help of the "option value" associated with the choice not to go into retirement today. Empirical studies show that workers generally react in a meaningful fashion to the financial incentives that accompany either early retirement or continued wage-earning.

Social Security and Private Pensions

Most countries in the OECD zone have put in place pension systems, public and private, enabling workers to receive income when they retire from the labor market. For example, the United States has a public system (Social Security) funded by mandatory contributions from employers, which gives a net benefit representing around 47 percent of her last net wage to the median worker retiring at age 66. This ratio is called the net replacement rate. Every individual has the opportunity to supplement this public retirement

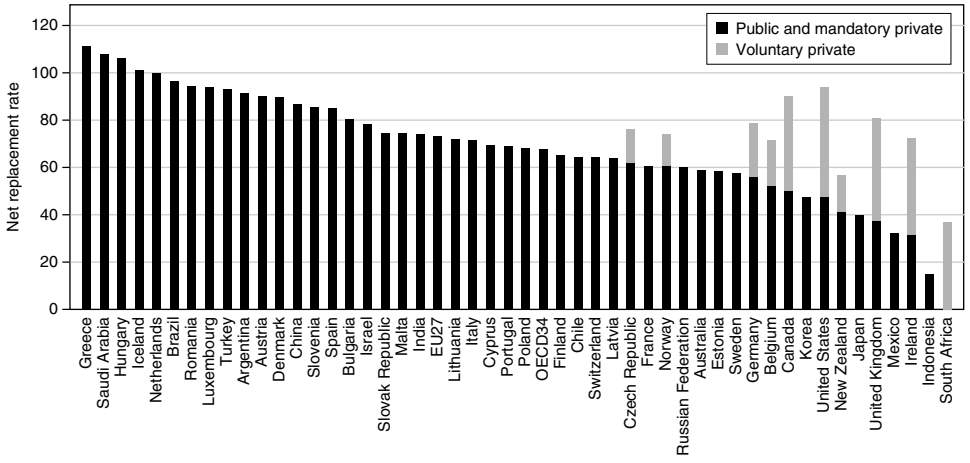


Figure 1.15
Net replacement rate from pensions (as a percentage of individual income), at the level of the median wage.

Source: OECD (2011, table II.2, p. 127).

payout with private pensions, contributions to which are negotiated between employer and employee at the moment the labor contract is signed. Taken as a whole, these contributions represent considerable financial accumulations—the celebrated pension funds—managed by specialized insurance companies that pay out retirement pensions to their members that vary according to the return their investments have made. In other countries like France and Sweden, the private system is practically nonexistent, and the net replacement rate offered by the public pensions is, in these two countries, on the order of 60 percent for a person who terminates his or her wage-earning activity (for a comparative international perspective, see OECD [2011] and figure 1.15, from which these isolated figures are taken).

The system of public and private pensions, to which we must add the tax system, creates incentives for workers to take their retirement earlier or later. Most retirement systems specify a legal age, sometimes called the "normal" retirement age, past which people can begin, if they wish, to draw full benefits without reduction for early retirement (for example, 66 in the United States, 65 in Germany, Denmark, and Japan). But every individual obviously has the right to retire before or after this legal age. As a general rule, she receives a smaller or larger pension the farther the age at which she ceases to work lies below or above the legal age. Hence the decision to retire brings into play a number of elements that emerge very clearly with the help of the life-cycle model, significantly modified.

Option Value in the Life-Cycle Model

Let us consider a person employed on date τ—this date represents, if you like, the age of this person—and let us suppose that this person decides to retire on date $s \geq \tau$. The evolution of his wealth starting from date τ is always given by equation (1.9), provided that we redefine certain variables of this equation. So, to simplify, we will suppose that the

agent does not work at all after date s; we will then have $L_t = 1$ for $t \geq s$. In practice, the process of ceasing to work can be gradual, and for that matter the legislation sometimes permits work to continue while the agent is receiving a retirement pension. We will use $B_t(s)$ to denote the income expected in the period $t \geq s$, composed of pension payments over the period t and other income that the agent may happen to have. Most often, this income is an increasing function of age s from career onset to retirement. To avoid confusion, we will use $B_t(0)$ to designate the non-earned income of the agent while he is still working, hence for $t < s$, and we will use C_{et} and C_{rt} respectively to designate his consumption of physical goods before and after retirement. For given s, the agent solves the following problem:

$$\max_{C_{et}, C_{rt}, L_t} \left[\sum_{t=\tau}^{s-1} U(C_{et}, L_t, t) + \sum_{t=s}^{T} U(C_{rt}, 1, t) \right]$$

subject to constraints

$$A_t = \begin{cases} (1 + r_t)A_{t-1} + B_t(0) + w_t(1 - L_t) - C_{et} & \text{if } \tau \leq t \leq s - 1 \\ (1 + r_t)A_{t-1} + B_t(s) - C_{rt} & \text{if } s \leq t \leq T \end{cases}$$

Let us designate the value of the welfare of the consumer at the optimum of this problem by $V_\tau(s)$, and finally let us denote the legal age of retirement by T_m, after which it is not possible to work any more. An agent age τ chooses the date s on which to end his working life by solving the following problem:[7]

$$\max_s V_\tau(s) \quad \text{subject to constraint } T_m \geq s \geq \tau \tag{1.19}$$

These problems never lend themselves to an explicit resolution and are generally solved numerically. In practice, we have to specify the utility function and the manner in which the replacement income is assembled to arrive at a model capable of being simulated or estimated empirically (one of the first attempts is found in Gustman and Steinmeier [1986]). Moreover, the decision to retire is made in an environment marked by numerous uncertainties (changes in one's professional and married life starting from date t, the chances of illness, changes in taste, retirement systems, etc.) that steadily subside as the legal age approaches. To simplify the explanation, we have written the agent's program without taking these uncertainties into account, but it is easy formally to introduce random factors into the utility function and into the equation for the evolution of wealth so as to obtain a stochastic model that fits reality more closely. In this case, $V_\tau(s)$ represents the intertemporal utility *expected* by an agent of age τ. Supplementary information may be acquired that will cause the decision made at age $(\tau + 1)$ to be different from the decision made at age τ. Let us denote by s^* the optimal solution of problem (1.19); for every period, the program (1.19) allows the agent to choose between two possibilities: retire today—the optimal solution to the agent's problem is a

[7] In this program, the terminal age $T \geq T_m$ must be interpreted as an indicator of anticipated length of life.

corner solution such as $s^* = \tau$—or continue to work until age $(\tau + 1)$ and reconsider his decision then, in which case the optimal solution is of the kind $s^* > \tau$.

This way of envisaging the process of ending one's working life leads us to examine the *option value* attached to the decision not to take retirement right now (Stock and Wise, 1990). Supposing that the decision to retire is irreversible, we have just shown that if $s^* = \tau$ the agent stops working immediately, and on the other hand if $s^* > \tau$ the agent continues to work and reconsiders his decision at age $(\tau + 1)$ in light of the new situation that he will be in when that date comes. The option value of not retiring today is thus equal to $V_\tau(s^*) - V_\tau(\tau)$. If it is positive the agent continues to work. If it is not, he goes into retirement. At the empirical level, this approach suggests that we estimate the probability of retirement at a given age by taking the option value as our principal explanatory variable. To obtain an indicator of this variable, we have to choose an explicit utility function, then estimate the option value tied to this utility function on the basis of a set of relevant variables, among which are income from public and private pensions and the wage outlook (readers may consult the survey of Lumsdaine and Mitchell [1999] for more detail). In general the indicator of the option value strongly influences decisions about retiring.

The Impact of Eligibility Rules

Empirical studies carried out in the United States show that changes made to the eligibility rules regarding Social Security pensions (the elimination of means testing, extension of the normal age for stopping work) have had little effect. The reason may be that private pension plans encourage workers to take their retirement starting at age 55, whereas Social Security only pays retirement income starting at age 62. If one looks only at private pensions, Gustman et al. (1994) show that individuals with the highest pensions are those who retire soonest. But this income effect is relatively feeble, since at age 60, a 10% increase in expected income over the entire (expected) duration of retirement reduces the length of working life by less than two months. Conversely, workers under financial pressure to put off their retirement do in fact extend their working lives. Here too the quantitative effects are faint: a 10% rise in expected income over the entire (expected) duration of retirement prolongs working life by less than six months. Using North American data from the Health and Retirement Survey (HRS), Coile and Gruber (2007) confirm these results but estimate that the effects may be quantitatively larger.

These results reveal the effects of retirement plans entered into at the time the worker was hired. But it is possible that, for reasons of productive efficiency, firms may offer pension plans that make it advantageous to take retirement sooner. Such firms will therefore attract workers who have a stronger inclination to retire early. In this case, the observed correlation between the financial incentives and the age at which retirement is taken does not reveal a causality; they simply show a property of an optimal contract between particular types of firms and particular workers. To eliminate this endogenous bias, numerous studies analyze the behavior of workers in the face of *unanticipated* changes in their retirement conditions. For example, Lumsdaine et al. (1990) studied a large American firm that, in 1982, offered a "window" to its employees over 55 who were enrolled in the pension plan, through which they could retire early; the financial bonus offered exceeded a year's worth of wages for certain categories of worker. By

definition, this window of opportunity was of limited duration and had not been anticipated by the employees. Lumsdaine et al. (1990) found that, in the case of the workers most advantaged by the new arrangement, the rate of leaving more than tripled. For the overall workforce, this study estimates that for a worker aged 50 employed in the firm, the likelihood of her retiring at age 60 was 0.77 under the new arrangement, whereas it was only 0.37 before it was put in place. These results are confirmed by Brown (1999), who systematically examined the effect of "windows" using data on the entire American population provided by the HRS.

The importance of financial incentives, whether direct, in the form of pension payments, or indirect, in the form of easier access to health insurance starting at a certain age, is confirmed by most research. The contribution of French and Jones (2011) estimates that pushing the age of eligibility for Medicare back from 65 to 67 years delays retirement by around 27 days between 60 and 69 years of age.

The effects of this type of financial incentive can also be studied through international comparisons. The studies of Gruber and Wise (1999, 2001, 2002) on a number of OECD countries show that financial incentives have, as a general rule, important impacts on the decision to retire. Gruber and Wise (2002), on the basis of data from 12 countries, calculate that for the average of these countries, a reform that lengthened by 3 years the period of eligibility required to retire on full pension would reduce the proportion of inactive men between 56 and 65 years of age by around 36% in the long term.

3 EMPIRICAL ASPECTS OF LABOR SUPPLY

The supply of labor is probably the area of labor economics in which the greatest number of empirical studies have been carried out in the course of the last 30 years. Advances in econometric methods have accompanied and made possible this increase. One of the main reasons for this trend is that for those whose job it is to plan employment policies or reforms of the fiscal system, the response of labor supply is a primary consideration.

One of the main problems confronting empirical analysis of labor supply is that the correlation between the pertinent financial incentives (principally the hourly wage) and the number of hours worked does not necessarily indicate a causal relation. It is possible that persons with little taste for leisure work longer and at the same time receive higher hourly wages because they are more motivated and so more efficient. In this case, it is not the high hourly wage that gave these people an incentive to supply more hours of work. On the contrary, it is conceivable that these same people, with their distaste for leisure, in the end receive lower *net hourly wages* (after tax) because of the progressivity of taxation. We would then observe a negative correlation between the hourly wage and the number of hours worked arising out of the progressivity of the fiscal system rather than any causal impact of the hourly wage on the volume of hours worked.

As a result of these problems of identification, researchers have turned to exogenous variations in income as a basis for estimating the impact of financial incentives on labor supply. The method frequently adopted is to compare the behavior of persons belonging to a "treated" group affected by an exogenous change in their income caused

by a change in the tax code, for example, with the behavior of other persons belonging to a "control" group who are unaffected by the same change. In this setting, two paths are possible. One approach evaluates the impact of a financial incentive, for example a tax credit, on hours worked, without estimating the parameters of a theoretical model of labor supply. This approach has the advantage of simplicity but makes it difficult to extrapolate the results to other contexts. It will be analyzed in further detail in chapter 12, which focuses on income redistributions and taxation. A second approach estimates the parameters of a model of labor supply, especially the different elasticities of Frisch, Marshall, and Hicks, so as to be able to extrapolate the results obtained to other situations. This presumes that one has good reason to believe that the model is a correct approximation of reality. In what follows we will present this approach before reviewing the empirical results.

3.1 Estimation of the Structural Parameters of Labor Supply Models

The estimation of the structural parameters of the labor supply model can be useful for evaluating the effects of numerous changes in the economic environment, the fiscal system for example, on behaviors and welfare. The estimation of structural parameters is thus a valuable aid to decision making in matters of public policy, since it has the power to *predict*, given well-defined (and possibly debatable) hypotheses, an order of magnitude for the consequences of different public initiatives.

It should be noted that the estimation of the structural parameters of labor supply models is today a domain of study in its own right, and we shall merely sketch the problems that arise within it and the principles that govern their resolution. A concrete example will show how these problems can be resolved in practice using data and programs that allow us to replicate the main results of the seminal contribution of the paper by Blundell et al. (1998). For a comprehensive account, the reader will profit from consulting the surveys of Blundell and MaCurdy (1999), Blundell et al. (2007), and Keane (2011).

3.1.1 Elasticities

The principal goal of empirical models of individual labor supply is to furnish an estimate of the wage elasticity of this supply. But the preceding theoretical analyses have taught us that there are several possible definitions of this elasticity, especially in the life-cycle model. On the empirical level, it is primarily the way an indicator of income from sources other than the current wage is constructed that permits us to discriminate among the definitions of elasticity. Also, wages are not fully exogenous variables for several reasons and they can only be observed by definition for those who supply a strictly positive number of hours. This is a challenge for the econometrician because labor supply behaviors have two margins: one is intensive and relates to the number of hours to be supplied, while the other is extensive and relates to the decision to participate or not in the labor market.

The Basic Equation and the Specification of Control Variables

As a general rule, estimates of labor supply equations are made on the basis of cross-section data (perhaps with temporal elements as well) produced by investigating a large

population, out of which a number of individuals or households are sampled. The empirical models which the econometrician tries to estimate almost always rest on a basic equation relating hours h_t worked by a given individual at hourly wage w_t at each date t. The following double-log-linear relation is a typical reduced form of this basic equation:

$$\ln h_t = \alpha_w \ln w_t + \alpha_R \mathcal{R}_t + \mathbf{x}_t \boldsymbol{\theta} + \varepsilon_t \tag{1.20}$$

In this expression, $\mathcal{R}_t$ is a measure of income other than the current wage (also called non-earned income, sometimes including income from assets or earned income from other household members), $\mathbf{x}_t$ is a vector of dimension $(1, n)$—one row and n columns—describing the n individual characteristics or control variables used, $\boldsymbol{\theta}$ is a vector of dimension $(n, 1)$ comprising n parameters to be estimated. The coefficients α_w and α_R are also parameters to be estimated, and finally, ε_t designates a random term reflecting individual heterogeneity that is not observed. Certain studies take h_t as a dependent variable rather than $\ln h_t$ and/or income w_t and $\ln \mathcal{R}_t$ rather than $\ln w_t$ or $\mathcal{R}_t$. These different specifications correspond to different restrictions on preferences (see Blundell and MaCurdy, 1999; Blundell, MaCurdy, and Meghir, 2007) that do not alter the principles guiding the estimation of equation (1.20). In order to fit theoretical models, for example the one in section 2.1.1, it is also possible to introduce a polynomial form of wage into the right-hand side of equation (1.20) so as to avoid postulating a priori that hours worked are a monotonic function of the hourly wage.

Parameter α_w measures the wage elasticity of labor supply. This elasticity can be interpreted in several ways according to the hypotheses made and the model utilized. This diversity of interpretation is presented here in the manner in which $\mathcal{R}_t$, indicating income apart from the current wage, is specified. The theoretical models taught us that individual labor supply at a given period was a function of the hourly wage for that period and other elements forming the expected wealth of an agent, such as her anticipated income from savings or work. If we limit ourselves to an equation of type (1.20), these elements have to be incorporated into variable $\mathcal{R}_t$. The important thing is to know how to carry out this incorporation in a way that is consistent with the life-cycle model. We will show that different ways to incorporate this variable allow us to estimate the different elasticities defined to this point.

Estimating Frischian Elasticity

The life-cycle model of section 2.3.1 has much to teach us when it comes to estimating Frisch elasticity, which measures variations in hours at time t for changes in wages at time t assuming that the marginal utility of wealth is constant. In particular, relations (1.12) and (1.13) defining its solutions, reveal that labor supply h_t depends on the current wage w_t and the marginal utility of wealth ν_t, so that $h_t = h(w_t, \nu_t, t)$. According to relation (1.13) of this model, the logarithm of ν_t breaks down into an individual fixed effect, independent of time, equal to $\ln \nu_0$, and an age effect $\sum_{\tau=1}^{\tau=t} \ln(1 + r_\tau)$, common to all agents and which may be written in the form ρt, supposing that r_τ is constant. We have

also seen that the value of ν_0 depends a priori on all the wages received by an individual during her lifetime. To obtain the elasticity of labor supply in Frisch's sense, we view the marginal utility of wealth ν_t as exogenously given. Following relation (1.13), we see that that amounts to supposing that $\ln \nu_0$ is also independent of the current wage but evidently does depend on individual characteristics. This property suggests substituting $\ln \nu_0 + \rho t$ for $\mathcal{R}$ in equation (1.20) to estimate Frischian elasticity. If we have longitudinal data available, we can eliminate individual fixed effects by taking this equation in first-differences, which is written:

$$\Delta \ln h_t = \rho + \alpha_w \Delta \ln w_t + \Delta \mathbf{x}_t \boldsymbol{\theta} + \Delta \varepsilon_t \tag{1.21}$$

This equation allows us to estimate the elasticity of labor supply in Frisch's sense, α_w, in a coherent manner, that is, the impact of a transitory change in the wage. It does not however allow us to evaluate the impact of a change in the overall wage profile, for a change of this type causes the marginal utility of wealth to vary a priori.

To estimate the impact of a change in the overall wage profile, another specification is necessary, one that can allow the marginal utility of wealth to vary with changes in wages.

Estimating Hicksian and Marshallian Elasticities

To estimate Hicksian and Marshallian elasticities, it is useful to look closely at some properties of the life-cycle model laid out in section 2.3.1. If the utility function is temporally separable, we have seen that the first-order condition (1.10) always implies equality between the marginal rate of substitution between consumption and leisure and the current wage at each date. This property suggests a two-stage resolution of this model, known in the literature as *two-stage budgeting* (MaCurdy, 1983; Blundell and Walker, 1986; Keane, 2011).

In the first stage, analogous to the basic static model, we define a potential income $\mathcal{R}_t$ for each period t, in such a way that the consumer's program consists of maximizing her instantaneous utility for the period t under a budget constraint, of which the nonearned income would be exactly $\mathcal{R}_t$. In the second stage, the consumer optimizes the series of $\mathcal{R}_t$, given the resources, present or anticipated, at her disposal. To arrive at such a program, we must first establish that the intertemporal budget constraint (1.9) of the life-cycle model may be rewritten in the following way:

$$C_t = \mathcal{R}_t + w_t h_t$$

where $\mathcal{R}_t = (1 + r_t)A_{t-1} + B_t - A_t$. The two-stage procedure by which the consumer resolves the program then emerges quite naturally. In the first stage, the consumer makes her choices for period t while maximizing instantaneous utility $U(C_t, 1 - h_t, t)$ subject to the static budget constraint $C_t = R_t + w_t h_t$, where R_t is considered given. The solution defines, as in the static model, the labor supply at date t as a function of the nonlabor income $\mathcal{R}_t$ and of the wage w_t.

At the conclusion of the first stage, the consumer attains a level of indirect utility $V(\mathcal{R}_t, t)$. In the second phase, she selects the optimal path for her assets A_t by solving

the program:

$$\max_{\{A_t\}} \sum_{t=0}^{T} V(\mathcal{R}_t, t) \quad \text{subject to } \mathcal{R}_t = (1 + r_t)A_{t-1} + B_t - A_t, \forall t$$

This two-stage procedure evidently yields the same solutions as the solution (in one stage) employed in section 2.3.1. On the empirical level, we should first note that the econometrician can know the values of $\mathcal{R}_t$ when he can observe the value of the consumption of physical goods C_t and the hours worked h_t at the same date, since $\mathcal{R}_t = C_t - w_t h_t$. If that is not the case, or if they cannot be known precisely enough, it is possible to estimate $\mathcal{R}_t$ by taking as explanatory variables the value A_{t-1} of assets at the outset of period t, the interest rate r_t, exogenous income B_t, all or part of the control variables of vector $\mathbf{x}_t$, and the expectation of all these independent variables (inasmuch as the value A_t of the assets at the end of the period t is not necessarily known, and depends on expectations of future resources).

Hence, we can substitute $\mathcal{R}_t = C_t - w_t h_t$ in equation (1.20) to obtain:

$$\ln h_t = \alpha_w \ln w_t + \alpha_R(C_t - w_t h_t) + \mathbf{x}_t \boldsymbol{\theta} + \varepsilon_t \tag{1.22}$$

From this equation, we can estimate the Marshallian elasticity of labor supply, $\partial \ln h / \partial \ln w = \alpha_w$, that is, the effect of a permanent wage change while non-earned income is held fixed (since the effect of non-earned income is accounted for separately in this equation). Recall from section 2.1.2 (equation (1.5)) that the Marshallian elasticity is the sum of the substitution effect, which is nothing other than the Hicksian elasticity (η_H), and of the (global) income effect ($\frac{wh}{\mathcal{R}}\eta_{\mathcal{R}}$):

$$\alpha_w = \eta_H + \frac{wh}{\mathcal{R}}\eta_{\mathcal{R}}$$

From equation (1.22), we deduce that the income effect corresponds to the coefficient of nonlabor income variable α_R multiplied by the wage: $\frac{wh}{\mathcal{R}}\eta_{\mathcal{R}} = \alpha_R wh$. Hence the Hicksian elasticity is equal to $\eta_H = \alpha_w - \alpha_R wh$, which is larger, in absolute value, than the Mashallian elasticity α_w if α_R is negative (i.e., if leisure is a normal good).

Let us now see how equations (1.21) and (1.22) can be estimated.[8]

3.1.2 An Instructive Example of a Life-Cycle-Consistent Approach

The first idea that may come to mind is to apply the method of ordinary least squares (OLS) to equations (1.21) or (1.22). Until the 1970s most studies proceeded in this way. But it is a flawed method, for it fails to take into account several potentially serious problems.

A first problem is the endogeneity of wages and non-earned income due to correlation with taste for work: applying the OLS to equations (1.20) or (1.22) relies on the assumption that wages and non-earned income are independent of the residual. This is obviously not the case if there are unobserved confounding variables that

[8] Data and programs are available on the book's website: www.labor-economics.org.

influence the hours of work and wages or nonlabor income. For instance, hours and wages may both depend positively on unobservable external factors, such as taste for work. In that case, the OLS coefficients of wages and non-earned income will be biased. The same problem applies to the relationship between hours and non-earned income. Pencavel (1986) reports a positive relationship between the level of assets and the level of hours worked, even after controlling for a number of observable characteristics, whereas we know that the income effect on labor supply should always be negative. This taste for work can be taken into account with individual fixed effects, as long as tastes are constant over time, and if the data have a panel dimension (i.e., persons are followed over several periods). In that case the error term in (1.22) is decomposed as $\varepsilon_t = \mu + \eta_t$ where μ is an individual fixed effect (time invariant) reflecting the person's taste for work and where η_t is an idiosyncratic individual taste shock (e.g., people may become unavailable for work in some particular period). However, this approach does not allow the econometrician to identify exogenous sources of changes in wages.

Another approach consists in finding situations with exogenous changes in incomes. This is the path followed by Blundell et al. (1998), who applied the life-cycle-consistent approach to married women in the United Kingdom from 1978 to 1992. During that time, the tax rates fell substantially over several periods. The fact that some working individuals have been exempt from any direct impact of these reforms due to the progressive nature of the tax system yields the opportunity to construct a suitable control group. This means that the data allow us to detect exogenous variations in net earned income after tax across groups, from which elasticities can be estimated. This is done by measuring the differences across groups and across periods in hours worked and wages. The core of their identification strategy lies here, for the decline in rates did indeed cause different cohorts to face different tax rate paths over their lifetime; those born in 1950, for instance, did not face the same tax profile over time as those born in 1960. Relative wages for groups with different education also changed markedly, since education is tightly linked to income levels and the change in tax schedules did not alter net income levels in the same way. Now, exogenous changes in tax schedules do induce exogeneity in a portion of the observed change in net wages but do not make all wage change exogenous, since gross wages (before tax rates are applied) can be linked to preferences and thus be endogenous to hours.

The basic idea of Blundell et al. in their article is to net out the endogenous changes from wage variations. The authors first group the individual data by cohort and education; that is, for each cohort/education level which constitutes one group, and for each period of change in the tax schedule, they construct group means of hours and wages. Separately, they calculate the means for each group over all periods and the means for each period over all groups. Then, they subtract these group and period means from the group means calculated in each period. The key assumption is that any residual variation in wages across groups (after taking out group and period means from the group-period means) is exogenous. Indeed, after this operation, unobserved time-invariant group factors that could influence wage levels and that could also be related to hour levels have been eliminated; unobserved time-variant factors common to all groups that could both influence wage levels and be related to hours levels have also been eliminated: the authors assume that unobserved confounding factors influencing hours and wages, like preference for work, might vary

across cohorts/education groups but do not vary over time within each cohort/education group.

The approach of Blundell et al. (1998) has been used by several contributions estimating parameters of the labor supply model. In particular, Devereux (2004) exploits the major changes in relative wages of both husbands and wives during the 1980s and the 1990s. Devereux treats national and regional changes in relative wages as an exogenous variation that can be used to identify labor supply responses. Blau and Kahn (2007) studied the labor supply of women from 1980 to 2000 in the United States using a similar approach.

Estimation of the Structural Parameter with Difference-in-Differences

To see better how to proceed, consider the basic semi-log equation which leaves aside non-earned income for notational simplicity:

$$h_{it} = \alpha + \alpha_w \ln w_{it} + \varepsilon_{it} \tag{1.23}$$

where w_{it} is the after-tax hourly wage of individual i at date t. Imagine that the tax reform implemented at date t affected two groups of individuals differently, group $g = T$ (for "treated") and group $g = C$ (for "control") and that the effect of the treatment (the applied policy change) transits through net wages. Let us assume that in the absence of the policy change, the means of hours would have evolved in the same way over time in both groups, or, in formal term, that:

$$\mathbb{E}[\varepsilon_{it}|g,t] = \eta_g + m_t \quad \text{for all } g \text{ and } t \tag{A1}$$

where η_g is a time-invariant group effect and m_t is a period effect common to all groups. This assumption is known as the *common trend assumption*. It assumes that the difference in average labor supply across groups, given the observables, remains unchanged over time. This assumption is a key *identifying assumption*, which means that the unobserved differences in average labor supply (given the wage, other income, and the demographics) are well accounted for by a permanent group effect and an additive time effect. In other words, unobserved factors, such as preference for work, can vary across groups or over time for all groups but cannot vary differently within groups over time; otherwise, the identification of wage elasticity would be impossible (the wage variations could stem from changing tastes for work and not just from the exogenous tax reform).

Using this assumption, and denoting by Δ the first difference operator (i.e., $\Delta x_t = x_t - x_{t-1}$), we get, from equation (1.23):

$$\Delta\mathbb{E}[h_{it}|T,t] = \alpha_w \Delta\mathbb{E}\ln[w_{it}|T,t] + \Delta m_t \tag{1.24}$$

$$\Delta\mathbb{E}[h_{it}|C,t] = \alpha_w \Delta\mathbb{E}\ln[w_{it}|C,t] + \Delta m_t \tag{1.25}$$

Then, assuming that the average change in after-tax wages before and after the reform is different for the treatment group and the control group, or, formally, that:

$$\Delta\mathbb{E}\ln[w_{it}|T,t] \neq \Delta\mathbb{E}\ln[w_{it}|C,t] \tag{A2}$$

the coefficient α_w can be deduced from the difference between equations (1.24) and (1.25):

$$\alpha_w = \frac{\Delta\mathbb{E}[h_{it}|T,t] - \Delta\mathbb{E}[h_{it}|C,t]}{\Delta\mathbb{E}[\ln w_{it}|T,t] - \Delta\mathbb{E}[\ln w_{it}|C,t]}$$

In this case the difference-in-differences estimator $\widehat{\alpha}_w$ of α_w, which measures the causal effect of the policy change on hours worked before and hours worked after the introduction of the new policy on those affected by this change (first difference) compared with those who were not affected by the reform (second difference), *conditional* on the fact that this impact transits through variation in wages, is:

$$\widehat{\alpha}_w = \frac{\Delta\overline{h}_t^T - \Delta\overline{h}_t^C}{\Delta\overline{\ln w}_t^T - \Delta\overline{\ln w}_t^C}$$

where the bar denotes sample average and the superscript denotes the group for which first differences are taken. This approach is valid if assumptions A1 and A2 are satisfied. A way to check the common trend assumption (assumption A1) is to check to see that differences in average labor supply across groups remain constant over time and that the composition of the groups does not change between before and after the reform. Assumption A2 is fulfilled if the groups are chosen to ensure that average changes in net wages are different across groups or, in other words, that the control and the treatment groups are indeed affected differently by the reform. For two groups and two periods, this estimator is equivalent to the difference-in-differences estimator. (See chapter 14, section 3.3, for a detailed presentation of the difference-in-differences approach.)

An advantage of this estimator is that it deals with measurement errors. This aspect is important to the extent that wages are usually measured with considerable error in microsurveys where individuals self-declare wages. In that case, OLS estimates of the coefficient of the wage variable tend to be biased toward zero, as is always the case when there are measurement errors (see, e.g., Wooldridge, 2013, chapter 15), leading to underestimates of labor supply elasticities. Another measurement error in wages arises when wage rates are constructed as the ratio of annual earnings to annual hours. If hours are also measured with some error this leads to "denominator bias." Such bias induces a negative correlation between measured hours and the ratio wage measure, biasing the wage coefficient in a negative direction (leading to the underestimation of elasticities). Grouping individuals is a way to reduce the bias induced by measurement error, assuming that the expected measurement errors are identical across groups.

Grouping Estimators

It is possible to generalize the difference-in-differences method to a situation with many groups and many periods. Using equations (1.23) and assumption A1, we can write:

$$\mathbb{E}[h_{it}|g,t] = \alpha + \alpha_w \mathbb{E}\ln[w_{it}|g,t] + m_t + \eta_g \qquad \textbf{(1.26)}$$

which can be written:[9]

$$D_h^{gt} = \alpha_w D_w^{gt} \tag{1.27}$$

where $D_h^{xt} = \mathbb{E}\left[x_{it} - \bar{x}|g,t\right] - \mathbb{E}\left[x_{it} - \bar{x}|g\right] - \mathbb{E}\left[x_{it} - \bar{x}|t\right], \mathbb{E}\left[x_{it}\right] = \bar{x}$.

Multiplying both sides of equation (1.27) by D_w^{gt} and summing over g and t, we get:

$$\sum_t \sum_g D_w^{gt} D_h^{gt} = \alpha_w \sum_t \sum_g \left(D_w^{gt}\right)^2 \tag{1.28}$$

Assuming that $\sum_t \sum_g D_w^{gt} D_w^{gt} \neq 0$, which is similar to assumption A2, meaning that average changes in wage differences across groups must be different, we get the expression of the elasticity of hours with respect to the wage:

$$\alpha_w = \frac{\sum_t \sum_g D_w^{gt} D_h^{gt}}{\sum_t \sum_g \left(D_w^{gt}\right)^2}$$

In practice, an estimator of this coefficient can be obtained by estimating the sample counterpart of equation (1.26) using OLS where each group is weighted by its relative size, or in other words, using weighted least squares. The equation that is estimated is:

$$\bar{h}_{gt} = \alpha + \alpha_w \overline{\ln w}_{gt} + m_t + \eta_g + \nu_{gt} \tag{1.29}$$

where $\bar{h}_{gt}$ and $\overline{\ln w}_{gt}$ are the sample group-period averages of hours and log wages, and ν_{gt} is an error term with zero mean.

Controlling for Participation

Wages are only observed by definition for those who work, and, more generally, labor supply behaviors might be different at the extensive margin (participating or not participating in the labor market) from those at the intensive margin (working more or fewer hours), especially when participating in the labor market entails some fixed cost. Failure to address this problem will lead to biased estimates of elasticities: if, for instance,

[9] Let us denote $\mathbb{E}\left[x_{it}\right] = \bar{x}$ and $\bar{m} = \sum_{t=1}^{T} m_t/T$, $\bar{\eta} = \sum_{g=1}^{G} \eta_g/G$, where T is the number of observed periods and G is the number of considered groups. Then, from equation (1.23) we have:

$$\begin{aligned}
\bar{h} &= \alpha + \alpha_w \bar{w} + \bar{m} + \bar{\eta} \\
\mathbb{E}\left[h_{it} - \bar{h}|g,t\right] &= \alpha_w \mathbb{E} \ln\left[w_{it} - \bar{w}|g,t\right] + m_t - \bar{m} + \eta_g - \bar{\eta} \\
\mathbb{E}\left[h_{it} - \bar{h}|g\right] &= \alpha_w \mathbb{E} \ln\left[w_{it} - \bar{w}|g\right] + \eta_g - \bar{\eta} \\
\mathbb{E}\left[h_{it} - \bar{h}|t\right] &= \alpha_w \mathbb{E} \ln\left[w_{it} - \bar{w}|t\right] + m_t - \bar{m}
\end{aligned}$$

Subtracting the last two rows from the second one, we get:

$$\mathbb{E}\left[h_{it} - \bar{h}|g,t\right] - \mathbb{E}\left[h_{it} - \bar{h}|g\right] - \mathbb{E}\left[h_{it} - \bar{h}|t\right] = \alpha_w \left(\mathbb{E}\left[w_{it} - \bar{w}|g,t\right] - \mathbb{E}\left[w_{it} - \bar{w}|g\right] - \mathbb{E}\left[w_{it} - \bar{w}|t\right]\right)$$

the higher the wage, the higher the probability to participate in the labor market, then people we see working positive hours despite relatively low wages probably also have a high taste for work; this induces a negative correlation between w_{it} and ε_{it} among the subpopulation of workers, even if w_{it} is exogenous in the population as a whole. The question that faces the econometrician is then: given a sample of individuals, how to take into account persons who do not work (or episodes during which an agent has not worked if the data are also longitudinal)? Certain studies subsequent to the 1970s simply set $h_{it} = 0$ for these persons. In other words, these studies took the view that certain workers choose exactly $h_{it} = 0$ just like any other value of h_{it}, which entails that equation (1.23) holds for any wage value of h_{it} and w_{it}. This assumption is false. Equation (1.23) is only valid for wages *above* the reservation wage, and for *all other* wages labor supply is null. Making do with equation (1.23) and setting $h_i = 0$ for episodes of nonwork thus leads to specification errors. An alternative solution was simply to exclude the unemployed, and nonparticipants in the labor market, from the sample. But in that case the econometrician commits a selection bias, forgetting that not to supply any hours of work is a decision in the same way that supplying them is. The fact that this type of decision is not described by equation (1.23) does not authorize us to set it aside purely and simply. The solution is either to employ an empirical model which, like the basic model of section 2.1.1, describes participation and hours decisions *jointly*, or to apply a sample selection correction term to equation (1.23) (see appendix 7.5 to this chapter). However, in the literature on male labor supply, this issue is often ignored on the grounds that the large majority of adult nonretired men do participate in the labor market, so the extensive margin would be minor. This is not the case when it comes to the labor supply of women, especially married women, and in the related literature the treatment of participation has become central.

In the context of the present model, this problem is important to the extent that we have implicitly assumed, so far, that the composition effects from changes in participation on the mean of the error term ε_{it} in equation (1.23) can be fully accounted for by the additive group and period effects. This is not a realistic assumption. First, changes in periods will cause individuals to enter and leave the labor market. Second, a tax policy reform in itself may lead to changes in participation. This problem might be particularly true for women: a higher wage may induce women with higher taste for leisure to enter the market within groups, even after controlling for specific characteristics of the groups and the periods. In more technical terms, this means that the mean of the residual of equation (1.29) may depend on participation decisions, represented by a variable denoted by P_{it}, that takes the value 1 if the individual is employed and zero otherwise. Therefore, $\mathbb{E}(\varepsilon_{it}|P_{it}, g, t)$ may vary over time, whereas it should be constant. Hence, OLS estimates could be biased if we forget to take participation into account.

To deal with the compositional effects of changes in participation rates on the mean of the error term in the labor supply equation, it is possible to use a two-step estimation method originally proposed by Heckman (1976) (see appendix 7.5 at the end of this chapter) and often called the Heckit method. Heckman's insight is that sample selection can be viewed as a form of omitted-variables bias, which can be corrected by adding a term to the estimated equation to replace this omitted variable. This term, denoted $\lambda(x) = \Phi'(x)/\Phi(x)$, where $\Phi(x)$ is the standard normal cumulative distribution function (cdf), is the inverse Mills ratio or a "selection hazard": we can think of it as

measuring the amount of selection in the data—the higher λ, the more drastic the selection arising from nonparticipation.

To implement this method, we must first (step 1) formulate a model, based on economic theory, for the probability of participating. The canonical specification for this relationship is a probit regression (see appendix 7.5 at the end of this chapter) of the form $\Pr[P_{it} = 1|\mathbf{z}] = \Phi(\mathbf{z}\boldsymbol{\gamma})$, where $\mathbf{z}$ is a vector of explanatory variables (including at least one significant variable excluded from the hours equation), $\boldsymbol{\gamma}$ is a vector of unknown parameters, and Φ is the cdf of the standard normal distribution. Of course $\boldsymbol{\gamma}$ is estimated with the probit using the *entire* population in the sample (of those participating and those not participating). Then (step 2), we can compute the inverse Mills ratio for each individual (here for each group-period cell). Denoting by $\widehat{L}_{gt} = \mathbf{z}_{gt}\widehat{\boldsymbol{\gamma}}$ the estimated proportion of individuals participating in group g in period t, then $\lambda^P_{gt} = \lambda(\widehat{L}_{gt}) = \Phi'(\widehat{L}_{gt})/\Phi(\widehat{L}_{gt})$. Finally, we add λ^P_{gt} to the initial equation and run it on the individual sample of those participating using weighted least squares:

$$\bar{h}_{gt} = \alpha + \alpha_w \overline{\ln w}_{gt} + m_t + \eta_g + \delta_P \lambda^P_{gt} + \nu_{gt} \tag{1.30}$$

To estimate this equation, we must of course assume that after taking out any variation induced by changes in the sample composition (due to participation decisions), wages must still vary differently across groups over time. If $\widehat{\delta}_p$ is not significantly different from 0, this means there is no significant participation bias.

Blundell et al. (1998) also remark that there are two important kinks in the tax schedule that must be controlled for: one at the level of earnings beyond which social contributions must be paid, and another beyond which income tax must be paid, leading to drops in income at those points. Not controlling for these kinks would bias downward the wage effect, since for people on the kink we would attribute their inertia to preferences rather than to the structure of the budget constraint. A simple way to overcome this problem is to drop these observations close to the kinks (by 5 hours) and to correct this potentially endogenous selection around the kinks by adding an additional inverse Mills ratio, denoted λ^T_{gt} (estimated using a probit model that models the probability to belong to this subgroup).

Adding Non-earned Income

The life-cycle approach includes non-earned income in the labor supply equation. As explained above in the discussion of the two-stage budgeting procedure, the inclusion of income effects is important for interpreting the estimated wage elasticity as a Marshallian elasticity and also for computing Hicksian elasticities for the purpose of evaluating the welfare effects of tax reforms. If we include a non-earned income variable in equation (1.30), then the same operation must be performed as with wages: either we include group means for this variable in the equation or we include the residual from a first-stage equation where nonlabor income is regressed over group and period dummies. This procedure allows us to deal with the endogeneity of non-earned income and to identify the elasticity of hours based on exogenous variations only. Thus the labor supply equation that Blundell et al. (1998) estimate has the form:

$$\bar{h}_{gt} = \alpha + \alpha_w \overline{\ln w}_{gt} + \alpha_R[\bar{C}_{gt} - \overline{w}_{gt}\overline{h}_{gt}] + m_t + \eta_g + \delta_P \lambda^P_{gt} + \delta_T \lambda^T_{gt} + \nu_{gt} \tag{1.31}$$

Here the second term is the virtual nonlabor income allocated to period t, η_g and m_t are the group and time dummies, and λ_{gt}^P and λ_{gt}^T are the inverse Mills ratios to control for participation rates and the conditioning out of observations close to the tax kinks. The authors add the possibility for α_w and α_R to vary with demographic groups, simply by interacting $\ln w$ and $(C - wh)$ with the demographic characteristics.

Estimation of (1.31) is by weighted least squares. Again, the identifying assumption is that all the unobservable confounding factors that influence both wages and hours are accounted for by time-invariant education-cohort group effects (η_g) and time-varying effects (m_t), which are the same across groups. Once these two types of effect are taken into account, it is assumed that there are no time-varying confounding factors that differ across groups.

Implementation and Main Results

To implement this procedure, the authors group the Family Expenditure Survey (FES) into two education groups (legal minimum versus additional education) and four cohorts (people born in 1930–39, 1940–49, 1950–59, and 1960–69), giving eight groups followed from 1978 to 1992. Education and age cohort are the "grouping instruments." The authors screen the data to include only 20- to 50-year-old women with employed husbands, over 15 financial years. This gives 24,626 women of whom 16,781 work. Only workers are used to estimate (1.31) while the full sample is used to form the λ_{gt}^P. As already mentioned, 2,970 of these women are within a few hours of a kink point in the tax schedule. Hours are "usual weekly hours, including usual overtime," and the pretax wage is built by dividing "usual weekly earnings, including usual overtime pay" by the hours. Consumption is measured as nondurable household consumption. The authors find that group/time interactions are highly significant in the wage and non-earned income equations.

The estimates of (1.31) imply a Marshallian (uncompensated) wage elasticity at the mean of the data of 0.17 and a Hicks (compensated) elasticity of 0.20. Table 1.3 presents the elasticities implied by the estimation of equation (1.31) with the possibility for α_w and α_R to vary with demographic groups. The Frisch elasticity cannot be recovered with the equation used because it does not incorporate a measure of lifetime wealth (lifetime wealth is sometimes proxied with consumption but here we only use

TABLE 1.3

Elasticities for married women in the United Kingdom using education and age cohorts as grouping instruments.

			Uncompensated	Compensated	Other	Group means		
	α_w	α_R	wage	wage	income	Hours	Wage	Income
No children	4.493	0.000	0.140	0.140	0.000	32	2.97	88.63
Youngest child 0–2	4.105	−0.028	0.205	0.301	−0.185	20	3.36	129.69
Youngest child 3–4	6.686	−0.022	0.371	0.439	−0.173	18	3.10	143.64
Youngest child 5–10	2.777	−0.014	0.132	0.173	−0.102	21	2.86	151.13
Youngest child 11+	3.260	−0.011	0.130	0.160	−0.063	25	2.83	147.31

Source: Blundell et al. (1998, table IV, p. 846, and table V, p. 848).

non-earned income). The Marshall elasticities range from 0.130 to 0.371, highest for women with children at preschool age, as we might expect. The income effects are all negative except for women with no children, where it is zero. As a result, the Hicksian elasticities are all positive. To calculate these elasticities at the group means, in the model used by the authors, the Marshall elasticity is given by $\partial \ln h / \partial \ln w = \alpha_w / \overline{h}$, the elasticity of other income is given by $\partial \ln h / \partial \ln .(\overline{C - wh}) = \alpha_R.(\overline{C - wh})/\overline{h}$, the income effect is given by $\alpha_R.\overline{w}$. The Hicks elasticity is the difference between the Marshall elasticity and the income effect.

3.2 Main Results in the Literature

The econometric methods laid out above, and some other alternative procedures, have allowed researchers to discern the properties of labor supply with greater precision.

3.2.1 Form of Labor Supply

Does an individual's supply of labor take the form of a hump-shaped curve, as depicted in figure 1.13? The study by Blundell et al. (1992) suggests that it does. Using data from research on the expenditures of British families, these authors focus on a sample of single mothers, whose weekly supply of labor they estimate, distinguishing between those who have non-earned income R greater than the median of the sample and those for whom non-earned income is less than the median. The results of this study are represented in figure 1.16.

Scrutiny of this graph confirms, in the first place, that the hypothesis that leisure is a normal good is well founded. We see that for practically all values of hourly wage, individuals in the sample who dispose of a non-earned income exceeding the median work less than the others. This graph also shows that the labor supply curve can indeed present a maximum (and even local maxima). Excluding wage values that are too low, we see that the labor supply curve for individuals whose non-earned income is less than the median strongly resembles the theoretical form of figure 1.13. For other individuals in the sample, the resemblance is less marked, but the essential point remains: for low hourly wages (on the order of £1 to £1.5), there is little supply and the substitution effect prevails, whereas for higher wages (from around £3 on up), the global income effect overrides the substitution effect.

3.2.2 Extensive and Intensive Margin Elasticities

The wide range of methods and samples used to estimate the elasticity of labor supply leads to a wide spread of results (Keane, 2011; Keane and Rogerson, 2012). Many studies have found that extensive-margin labor supply elasticity is larger than intensive-margin labor supply elasticity. In essence, two reasons explain this result: indivisible labor supply and optimization frictions.

Indivisible Labor Supply

The existence of indivisibilities in labor supply may lead to an elasticity at the extensive margin greater than the elasticity at the intensive margin (Hansen, 1985; Rogerson, 1988; Rogerson and Wallenius, 2007; Keane and Rogerson, 2012). In this case, changes in tax

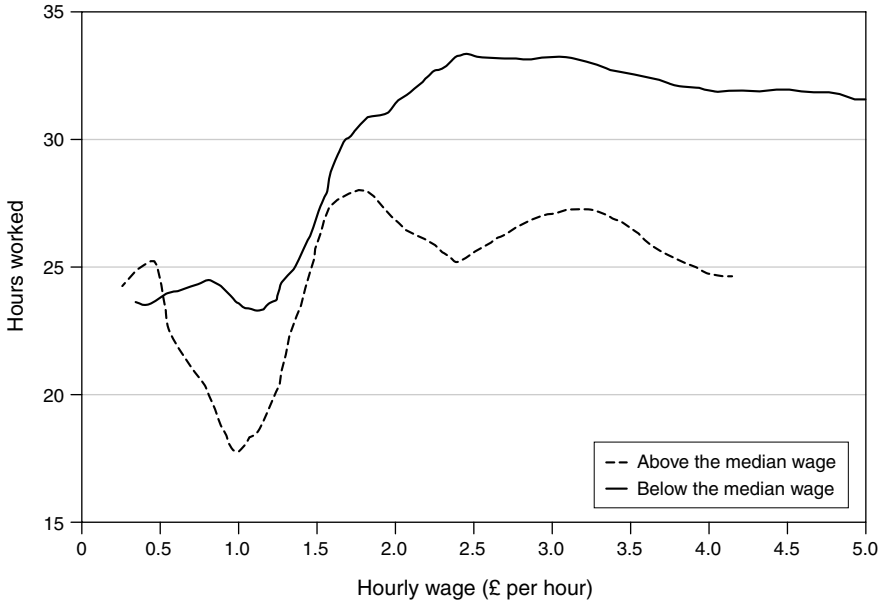


Figure 1.16
The labor supply of single mothers.

Source: Blundell et al. (1992).

or wage rates are compatible with large extensive-margin responses even if they have little effect on hours conditional on employment. To show this result, let us take the case envisaged previously (section 2.1.2), in which each agent has the choice between working for a fixed length of time $\bar{h} = L_0 - L_f$ and not working at all. Let us imagine that the diversity of reservation wages $w_A \in [0, +\infty)$, that may come from heterogeneous preferences and non-earned income may be represented by a cumulative distribution function $\Phi(\cdot)$. By definition, the quantity $\Phi(w)$ represents the participation rate, that is to say, the proportion of working-age individuals in the population whose reservation wage is below the current wage w. Since the function Φ is increasing, the participation rate climbs as the wage increases. The extensive-margin elasticity is equal to the elasticity of Φ with respect to w: $\mathrm{d}\ln\Phi/\mathrm{d}\ln w$. Hence, in this example, the extensive-margin elasticity can be large even if, at the intensive margin, the elasticity is equal to zero.

Optimization Frictions

A complementary explanation of the difference between estimations of elasticities at the intensive and extensive margins relies on the existence of fixed adjustment costs of labor supply at its optimal value (Chetty, 2012). This eventuality may arise from organizational constraints internal to the firm, which make the adjustment of hours costly, or create the cost of finding another job better adapted to the worker's desired timetable, if it is not possible to adjust the hours in his current job. Such adjustment costs may lead to underestimating the elasticity of labor supply at the intensive margin, for the gains from adjusting hours of work in the wake of small wage variations are

highly likely to be smaller than the adjustment costs, to the extent they are second-order gains. Conversely, when the variations are at the extensive margin, *and if working entails a fixed cost*, the gains are of the first order.

To show this, let us revert to the static model at the beginning of this chapter in which preferences are represented by the utility function $U(C,L)$ and the budget constraint by the relation $wL + C = wL_0 + R$. We will assume that working entails a fixed cost, denoted $F > 0$, for example outlays on transportation or suitable workplace attire. In this case, an individual who works attains a level of utility $U[w(L_0 - L) + R - F, L]$. The optimal duration of leisure, denoted $L(w)$, is the value of $L \leq L_0$ that verifies the first-order condition $U_L - wU_C = 0$.

Let us suppose that the wage goes from level w to level w_1 and that it is optimal, for this individual, to adjust his duration of work at the intensive margin (i.e. to work more). His new optimal duration of leisure is $L(w_1) < L_0$. The gain from adjusting his hours is:

$$G_I = U[w_1(L_0 - L(w_1)) + R - F, L(w_1)] - U[w_1(L_0 - L(w)) + R - F, L(w)]$$

The approximation of this gain by a first-order Taylor expansion around point $[w_1(L_0 - L(w_1)) + R - F, L(w_1)]$ gives:

$$\begin{aligned} G_I = [L(w_1) - L(w)]\,[&U_L[w_1(L_0 - L(w_1)) + R - F, L(w_1)] \\ &- w_1 U_C[w_1(L_0 - L(w_1)) + R - F, L(w_1)]] \end{aligned}$$

This approximation of the first-order gain is null, for $U_L - wU_C = 0$ at point $[w_1(L_0 - L(w_1)) + R - F,\ L(w_1)]$. This result is illustrated in figure 1.17, where we see that a wage increase from level w to level w_1 induces a shift from point A to point B if

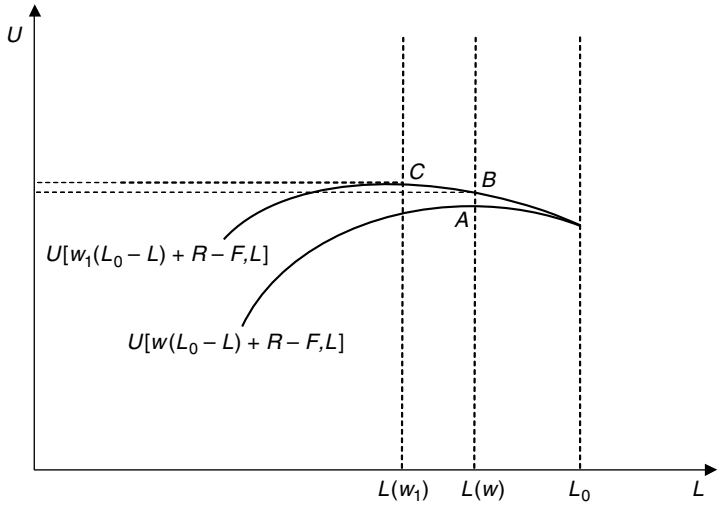


FIGURE 1.17
The consequence of a wage increase from w to w_1 on the intensive-margin labor supply.

the duration of work remains unchanged at its initial level $L_0 - L(w)$. In the shift from A to B, utility increases, since the wage for the hours initially worked rises. The gain from adjusting hours worked (represented on the vertical axis) corresponds to a shift from point B to point C, which induces a weak increment of utility, null in the second order, since point B is situated on the curve $U[w_1(L_0 - L) + R - F, L]$, close to the maximum of this curve.

In contrast to the adjustment at the intensive margin, the gains from adjusting hours at the extensive margin are not null in the first order. Assuming that the individual has an interest in not working at wage w and an interest in working at wage $w_1 > w$, the gain from the shift from nonwork to work is given by:

$$G_E = U[w_1(L_0 - L(w_1)) + R - F, L(w_1)] - U(R, L_0)$$

A first-order Taylor expansion around point $[w_1(L_0 - L(w_1)) + R - F, L(w_1)]$ gives:

$$G_E \approx FU_C[w_1(L_0 - L(w_1)) + R - F, L(w_1)]$$

Figure 1.18 illustrates this result by presenting a situation where it is not of interest to work at wage w, while it is of interest to furnish a quantity of labor $L_0 - L(w_1) > 0$ at wage w_1. Readers will observe that the gain (shown on the vertical axis) from shifting from initial point A, where no labor is furnished, to point C, corresponding to the optimal duration of work for wage w_1, may be greater than the second-order gain that would be obtained by shifting from point B to point C, corresponding to the gain realized through an adjustment at the intensive margin. The gains from adjusting the duration of labor at the extensive margin are of the first order, for in that situation the individual does not benefit from the wage rise if he does not adjust his hours of work. Conversely, an individual who is working and making decisions at the intensive margin benefits

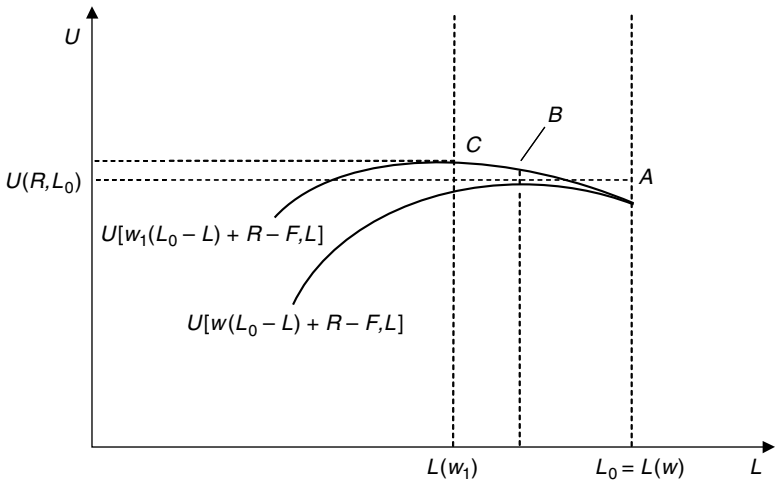


FIGURE 1.18
The consequence of a wage increase from w to w_1 on the extensive-margin labor supply.

from the wage rise even if he does not adjust his hours of work. This is why the gains from adjusting hours worked are greater at the extensive margin than at the intensive margin.[10]

The existence of fixed costs can explain why small variations in the wage have a null impact on the variation of hours at the intensive margin and a significant impact on the variation of hours at the extensive margin. Thus, the estimation of elasticity at the intensive margin, in the presence of small variations in the wage, caused for example by fiscal reforms of a minor kind, may lead to a null value, while elasticity in a context without friction, if estimated in the presence of large-scale wage variations, would be significantly different from zero. From this perspective, Chetty shows that many micro-econometric studies of labor supply are uninformative about intensive-margin elasticities because they cannot reject large values of this elasticity with adjustment costs of even 1% of earnings in choosing labor supply. Combining estimates from several studies, Chetty (2012) argues that point estimates of "structural" (i.e., frictionless) Hicksian elasticities are 0.33 at the intensive margin and 0.25 at the extensive margin. Hence, Chetty finds that structural elasticity at the intensive margin is even greater than at the extensive margin. This result suggests that structural intensive- and extensive-margin elasticities are likely not very different. Chetty also finds that Frisch elasticities cannot be much larger than Hicksian elasticities, given plausible income effects.

3.2.3 Micro and Macro Elasticities

Macro elasticities of labor supply measure the elasticity of aggregate labor supply at the country level. The "aggregate hours elasticity" can be computed from the extensive and intensive elasticities. This can be understood with a simple example where all individuals are identical except for their reservation wage and the size of the population is equal to one. Let us denote the wage by w. In this context, the aggregate number of hours is equal to the individual number of hours conditional on working, $h(w)$, times the participation rate $\Phi(w)$, where $\Phi(w)$ stands for the cumulative distribution function of the reservation wages in the population. Therefore, the aggregate hours elasticity is equal to the sum of the extensive and intensive elasticities.[11]

Chetty et al. (2011b) have summarized the micro and macro evidence on the extensive and intensive margins. Their results are displayed in table 1.4. Each cell in this table shows a point estimate of the relevant elasticity based on the analysis by the authors of many existing micro and macro studies. Micro estimates are identified from quasi-experimental studies, usually exploiting variations in taxation over time. Micro estimates correspond to structural elasticities (without friction, as stipulated in the previous subsection). Macro estimates are identified from cross-country variation in tax rates (Hicksian elasticities) and derived to match business cycle fluctuations (Frisch

[10] The importance of introducing a fixed cost of working F into this line of reasoning will be evident. It can easily be verified that the gain from adjusting hours of work at the extensive margin is of the second order if $F = 0$, for in that case, the situation is strictly identical to adjustment at the intensive margin.

[11] If $L_A(w) = h(w)\Phi(w)$, we can write $\ln L_A(w) = \ln h(w) + \ln \Phi(w)$. Deriving this equation with respect to w, we get:

$$\frac{w}{L_A(w)}\frac{\mathrm{d}L_A(w)}{\mathrm{d}w} = \frac{w}{h(w)}\frac{\mathrm{d}h(w)}{\mathrm{d}w} + \frac{w}{\Phi(w)}\frac{\mathrm{d}\Phi(w)}{\mathrm{d}w}$$

elasticities). It should be noted that Chetty et al. (2011b) consider *structural* micro elasticities to match micro and macro estimates because they assume that changes in aggregate net wages, induced by countrywide tax reforms, for instance, should entail negligeable friction. The reason is that individuals should be able to coordinate when there are large and countrywide changes (Chetty et al., 2011a). Chetty et al. (2011b) focus on the Hicksian elasticity, which is relevant to determine the impact of taxes in steady-state if government revenues are returned to the consumer as a lump sum, as commonly assumed in representative-agent macro models. If revenues are not returned to consumers, tax changes have income effects and the Marshallian elasticity becomes the relevant parameter. Since leisure is a normal good, the Hicksian elasticity is an upper bound (in absolute value) of the Marshallian elasticity.

Table 1.4 shows that structural micro estimates of Hicksian elasticities and macro estimates of Hicksian elasticities match when both the extensive and the intensive margins are considered together. Accordingly, Hicksian micro and macro elasticities are consistent with the observed differences in aggregate hours across countries with different tax systems.

But Frisch elasticities do not match: estimates are small based on micro evidence but large based on macro studies. Macro studies do not always decompose Frisch elasticities into extensive and intensive elasticities. Therefore, the estimates in brackets show the values implied by the macro aggregate hours elasticity if the intensive Frisch elasticity is chosen to match the micro estimate of 0.54. Macro models calibrate the Frisch aggregate hours elasticities to match business cycle data, especially employment fluctuations. They find on average an intertemporal elasticity of 2.84, more that 3 times larger than the one based on micro studies (0.82). This means that extensive labor supply elasticities identified in micro studies are not large enough to explain the large fluctuations of employment observed over the business cycle (as compared to hours), even when fixed costs of adjustments at the participation margin are accounted for.

Hence theorists and empirical researchers are left with two possibilities: either the micro estimates are based on models that overlook important factors that could increase elasticities, or macroeconomists of the business cycle do not have the right model available, one that could describe economic fluctuations consistently with observed agents'

Table 1.4

Micro vs. macro labor supply elasticities. Each cell shows a point estimate of the relevant elasticity based on meta-analyses of existing micro and macro evidence. Micro estimates are identified from quasi-experimental studies; macro estimates are identified from cross-country variation in tax rates (steady-state elasticities) and business cycle fluctuations (intertemporal substitution elasticities). The aggregate hours elasticity is the sum of the extensive and intensive elasticities. Macro studies do not always decompose intertemporal aggregate hours elasticities into extensive and intensive elasticities. Therefore, the estimates in brackets show the values implied by the macro aggregate hours elasticity if the intensive Frisch elasticity is chosen to match the micro estimate of 0.54.

		Intensive margin	Extensive margin	Aggregate hours
Steady-state (Hicksian)	micro	0.33	0.26	0.59
Steady-state (Hicksian)	macro	0.33	0.17	0.50
Intertemporal substitution (Frisch)	micro	0.54	0.28	0.82
Intertemporal substitution (Frisch)	macro	[0.54]	[2.30]	2.84

Source: Chetty et al. (2011b, table 1, p. 2).

behavior. This is an ongoing debate and research field (see the surveys of Keane, 2011; Keane and Rogerson, 2012). One line of research attempts to estimate life-cycle models where the accumulation of human capital is endogenous. If work experience builds human capital, then the current labor supply decision also affects future wages, notably for higher-wage earners. This can lead to higher estimates of Frisch elasticities (Imai and Keane, 2004).

3.2.4 The Elasticity of Labor Supply of Men and Women

It is generally acknowledged that labor supply elasticities for men are very small and often not significantly different from zero, whereas labor supply elasticities for women are somewhat larger (though less than one), especially the labor supply elasticity of married women, which is demonstrably positive and greater than that of their spouses (Blundell and MaCurdy, 1999). If we turn to theoretical models, these results indicate that within the household fiscal reforms affect principally the participation decisions of women, since on average they have access to wages lower than those of men and in all likelihood possess a comparative advantage when it comes to household production.

The difference between the elasticities of the labor supplied to the market by men and by married women is explained by the fact that women's labor is regarded as more substitutable for domestic work than that of men, especially when the woman is less qualified than her spouse. In this regard, it is interesting to note that there exists a relation between the trend of the rate of participation of women in the labor market and the trend of the elasticity of their labor supply. The labor force participation of women increased steeply throughout the 20th century, although at a slower pace since the 1990s. This is notably the case in the United States, as shown in figure 1.4 and in table 1.1. Blau and Kahn (2007) have analyzed this slowdown for married women between 1980 and 2000. The interpretation of Blau and Kahn is that the labor supply function shifted sharply to the right in the 1980s, implying a higher availability for work at any given wage, but that no such shift happened in the 1990s. This would account for the more rapid growth of female labor supply in the 1980s than in the 1990s. Additionally, over the two decades, married women's labor supply elasticity was halved (the labor supply slope became steeper), which is a very important change also identified by Heim (2007), and the labor supply of women also became less responsive to their husbands' wages. Overall, at the end of the century and after decades of increasing participation, women's labor supply behavior grew closer to men's as preference for work changed, probably reflecting the fact that women became economically more independent of men and became more oriented towards their own careers. This trend also means that public policies aimed at increasing further the labor supply of women will be less effective in the future.

3.2.5 The Cost of Leisure and the Productivity of Home Production

To this point, we have focused on contributions dealing with the estimation of the elasticities of labor supply with respect to labor and nonlabor incomes. However, the labor supply model predicts that other variables, such as the cost of leisure activities and the productivity of home production, also influence labor supply. Taking these variables

into account proves useful for explaining the strong increase in female labor supply observed in many countries over the last century (see section 1.3).

The Cost and the Utility of Leisure

The labor supply depends on the cost of leisure. In the basic models, this cost is measured simply as an opportunity cost, that is forgone wages. Direct cost, like the price of leisure activities, can also be considered in measuring the elasticity of labor supply. González-Chapela (2007) estimated the intertemporal elasticity of labor supply by using not just wage variations but also the trend in the price of various leisure-related goods (sports equipment, games, hobbies, magazines and books, etc.) and services (club membership, training, lessons, etc.) on hours worked by men in 27 large American cities between 1976 and 1993. In these cities, not only is there a spread in the price of goods and services related to leisure but their trend has diverged over the course of time. By using this source of variation, Gonzáles-Chapela estimates the intertemporal elasticity of substitution. He obtains an intertemporal elasticity of 0.16 to the price of leisure goods and 0.25 to wages for working-age men. This means, for a man working 2,000 hours per year, that a fall of 1% in the price of leisure goods would prompt a fall of 3.2 hours in the length of time worked annually. Now, the relative prices of leisure goods and services have risen from the 1980s until at least 1993 in the United States, whereas they have remained constant or have continued to trend downward in certain European countries after a steep decline already registered during the 1970s. This might help to explain the differential in the trend in length of time worked between the United States and Europe over that period.

Another approach is to look at changes in the utility derived from leisure. An exogenous source of variation is quite simply the weather outside. When the weather is fine, the opportunity cost of working is greater than it is on days when it rains or is otherwise inclement. Hence variations in weather conditions are also a potential source of variation in hours worked annually. For the United States, Connolly (2008) used data from surveys carried out in 2003 and 2004 on time use (which measured the amounts of time invested in work, domestic activity, and leisure activity), with which she combined meteorological data for the same period. She finds that men reduce their investment in leisure time by 30 minutes on rainy days so they can work longer.

Explaining the Change in Female Labor Supply

Some studies have focused on the impact of children on the working lives of women. These studies generally bring out a negative effect of parenthood on labor supply by women. For instance, Bloom et al. (2009) estimate the effect of fertility on female labor force participation in a panel of countries using abortion legislation as an instrument for fertility. They find that removing legal restrictions on abortion significantly reduces fertility and estimate that, on average, a birth reduces a woman's labor supply by almost 2 years during her reproductive life. Moreover, they argue that behavioral change, in the form of increased female labor supply, contributes significantly to economic growth during the demographic transition when fertility declines. However, it turns out that the effect of parenthood on labor supply by women depends on things like part-time opportunities, child care, optional parental leave, and child allowances, which are different across countries (Del Boca et al., 2009).

The ability to control birth timing is one important factor that probably bolstered female participation. Using state-level changes from 1960 to 1976 in the United States that progressively expanded the legal rights to contraception of individuals aged 18 to 21, Bailey (2006) shows that access to the pill before age 21 significantly reduced the likelihood of a first birth before age 22 by 14 to 18%, increased the number of 26- to 30-year-old women in the paid labor force by approximately 8%, and raised their number of annual hours worked by 68.

The availability of child care is another possible factor that may help to explain the high female employment rates in Scandinavian countries, as suggested by a natural experiment that occurred in Canada. In 1997 the government of Quebec introduced a new set of family policies, including a universal child care program to provide regulated child care spaces to all children aged 0–4 years in Quebec with a parental contribution of only C$5.00 per day, whether or not the parents were working. Baker et al. (2008) evaluated the impact of these subsidies using a difference-in-differences approach that compares Quebec to other provinces of Canada where no such change occurred in 2000 or later relative to 1997 or earlier. The Quebec policy induced a shift into new child care use, although approximately one-third of the newly reported use appears to come from women who previously worked and had informal arrangements. Despite this crowding-out effect, Baker et al. find an increase in employment of women in Quebec, relative to the rest of Canada, of 7.7 percentage points, or 14.5% of baseline participation (about half as large as the impact of the program on child care utilization).

4 SUMMARY AND CONCLUSION

- According to the neoclassical theory of labor supply, every individual trades off between consuming a good and consuming leisure. The supply of individual labor is positive if the current wage exceeds the *reservation wage*, which depends on preferences and non-earned income. If labor supply is positive, the marginal rate of substitution between consumption and leisure is equal to the hourly wage.
- The relation between the individual supply of labor and the hourly wage is the result of combined substitution and income effects. The substitution effect implies an increasing relation between the wage and labor supply, while the income effect works in the opposite direction if leisure is a normal good. The supply of labor generally rises with the wage at low wage levels (the substitution effect prevails) and falls off when the wage reaches higher levels (the income effect prevails).
- In the neoclassical theory of labor supply, the labor force participation rate corresponds to the proportion of individuals whose reservation wage is less than the current wage.
- When an individual has the opportunity to devote a part of her endowment of time to household production, at the optimum the hourly wage is equal to the marginal productivity of household work. The possibility of substituting household production for wage work increases the elasticity of the individual supply of wage work.

- As a general rule, the mechanism of substitution of leisure over time implies that the permanent component of the evolution of real wages has a smaller effect on labor supply than the transitory component.
- The Hicksian elasticity is about 0.3 at the intensive margin and about 0.2 at the extensive margin. This implies that the Hicksian elasticity of aggregate supply of hours is about 0.5.
- The Frisch elasticity is about 0.5 at the intensive margin and about 0.3 at the extensive margin. This implies that the Frisch elasticity of aggregate supply of hours is about 0.8.
- The elasticity of labor supply by women is, in general, greater than that of men, which is generally small, although this difference diminishes over time.

5 RELATED TOPICS IN THE BOOK

- Chapter 3, section 1.2: The question of tax incidence
- Chapter 3, section 1.3: The effect of a shock on labor supply
- Chapter 4, section 2: The theory of human capital
- Chapter 5, section 2.1.3: The choice between nonparticipation, job search, and employment
- Chapter 9, section 1.1: Unemployment, employment, and participation
- Chapter 11, section 3.1: The characteristics of migrations
- Chapter 12, section 1.2: The effects of taxes on the labor market
- Chapter 12, section 1.3: What empirical studies tell us
- Chapter 12, section 2.2: Minimum wage and employment
- Chapter 13, section 1: Unemployment insurance

6 FURTHER READINGS

Blundell, R., & MaCurdy, T. (1999). Labor supply: A review of alternative approaches. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 3A, chap. 27). Amsterdam: Elsevier Science.

Heckman, J. (1974). Shadow prices, market wages and labor supply. *Econometrica*, *42*(4), 679–694.

Keane, M. (2011). Labor supply and taxes. *Journal of Economic Literature*, *49*(4), 961–1075.

Lumsdaine, R., & Mitchell, O. (1999). New developments in the economic analysis of retirement. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 3, chap. 49, pp. 3261–3307). Amsterdam: Elsevier Science.

7 APPENDICES

7.1 PROPERTIES OF INDIFFERENCE CURVES

If we suppose that the satisfaction of an agent increases with leisure and consumption, so that $U_C(C,L) > 0$, and $U_L(C,L) > 0$, the indifference curves are then negatively sloped. Consequently the indifference curve associated with level of utility $\overline{U}$ is composed of the set of couples (C,L) satisfying $U(C,L) = \overline{U}$. This equality implicitly defines a function $C(L)$, which satisfies $U[C(L),L] = \overline{U}$. Differentiating this last expression with respect to L, we get:

$$C'(L) = -\frac{U_L(C,L)}{U_C(C,L)} < 0 \tag{1.32}$$

The indifference curves are indeed negatively sloped. We observe that the absolute value of the slope $C'(L)$ of an indifference curve is equal to the marginal rate of substitution U_L/U_C between consumption and leisure.

The hypothesis of the convexity of indifference curves is equivalent to the property of quasiconvexity of the utility function. Indifference curves are convex if and only if $C''(L)$ is positive. This second derivative is calculated using the equality $U(C,L) = \overline{U}$ and equation (1.32). We thus get:

$$C''(L) = \frac{U_L\left[2U_{CL} - U_{LL}\left(\frac{U_C}{U_L}\right) - U_{CC}\left(\frac{U_L}{U_C}\right)\right]}{(U_C)^2} \tag{1.33}$$

Since $C''(L)$ is of the sign of the term between square brackets of the numerator of the right-hand side of equation (1.33), the quasiconcavity of the utility function corresponds to the condition:

$$U(C,L) \text{ quasiconcave} \iff 2U_{CL} - U_{LL}\left(\frac{U_C}{U_L}\right) - U_{CC}\left(\frac{U_L}{U_C}\right) > 0 \tag{1.34}$$

7.2 THE PROPERTIES OF THE LABOR SUPPLY FUNCTION

For an interior solution, relations (1.2) allow us to obtain the demand for leisure L^*. We thus have:

$$wU_C(R_0 - wL^*, L^*) - U_L(R_0 - wL^*, L^*) = 0 \tag{1.35}$$

This equation implicitly defines L^* as a function of $R_0 = wL_0 + R$ and of w. We denote this function $\Lambda(w, R_0) = L^*$. Its partial derivatives are obtained by differentiating equation (1.35), which implies:

$$\mathrm{d}L^*\left(-w^2U_{CC} + 2wU_{CL} - U_{LL}\right) + \mathrm{d}w\left[U_C - L\left(wU_{CC} - U_{CL}\right)\right] + \mathrm{d}R_0\left(wU_{CC} - U_{CL}\right) = 0 \tag{1.36}$$

By replacing the value w defined by (1.35), so that $w = U_L/U_C$ in (1.36), we get the expressions of the partial derivatives of function Λ:

$$\Lambda_1 = \frac{\partial L^*}{\partial w} = \frac{-L\left(\frac{U_{CL}U_C - U_{CC}U_L}{U_L}\right) - U_C\left(\frac{U_C}{U_L}\right)}{\left[2U_{CL} - U_{LL}\left(\frac{U_C}{U_L}\right) - U_{CC}\left(\frac{U_L}{U_C}\right)\right]} \tag{1.37}$$

$$\Lambda_2 = \frac{\partial L^*}{\partial R_0} = \frac{\frac{U_{CL}U_C - U_{CC}U_L}{U_L}}{\left(2U_{CL} - U_{LL}\left(\frac{U_C}{U_L}\right) - U_{CC}\left(\frac{U_L}{U_C}\right)\right)} \tag{1.38}$$

According to relation (1.34), the quasiconcavity of the utility function implies that the denominator of the right-hand side of equations (1.37) and (1.38) is positive. Λ_2 is then of the sign of $U_{CL}U_C - U_{CC}U_L$. It is positive if and only if leisure is a normal good (L^* then grows with R_0). If Λ_2 is negative, leisure is an inferior good. Scrutiny of equation (1.37) shows that an increase in the wage entails an income effect (which we have described as indirect) and a substitution effect corresponding to the first and second terms in square brackets of the numerator of the right-hand side. If leisure is a normal good, $U_{CL}U_C - U_{CC}U_L > 0$, the two effects work in the same way and Λ_1 is negative. If leisure is an inferior good, Λ_1 has an ambiguous sign.

7.3 COMPENSATED AND NONCOMPENSATED ELASTICITY

The Hicksian demand functions of leisure and of consumption goods are obtained by minimizing the expenditures of the consumer under the constraint of a minimal exogenous level of utility, denoted $\bar{U}$. They are thus solutions of the problem:

$$\min_{(L,C)} C + wL \quad \text{subject to constraint } U(C,L) \geq \bar{U} \tag{1.39}$$

Let us use $\hat{L}(w,\bar{U})$ and $\hat{C}(w,\bar{U})$ to designate the solutions of this problem; the expenditure function, denoted $e(w,\bar{U})$, is defined by the identity $e(w,\bar{U}) = \hat{C}(w,\bar{U}) + w\hat{L}(w,\bar{U})$. By construction, the Hicksian and Marshallian demand functions, respectively $\hat{L}(w,\bar{U})$ and $L^* = \Lambda(w,R_0)$—given by the equation (1.2)—satisfy the identity $\Lambda\left[w, e(w,\bar{U})\right] = \hat{L}(w,\bar{U})$. If we derive this identity with respect to w, we get:

$$\Lambda_1\left[w, e(w,\bar{U})\right] + e_1(w,\bar{U})\Lambda_2\left[w, e(w,\bar{U})\right] = \hat{L}_1(w,\bar{U}) \tag{1.40}$$

We may point out that function $d(w) \equiv \hat{C}(x,\bar{U}) + w\hat{L}(x,\bar{U}) - e(w,\bar{U})$ reaches a minimum for $w = x$, which implies $d'(w) = 0$ for $w = x$, and thus $e_1(w,\bar{U}) = \hat{L}(w,\bar{U})$. To simplify these notations, let us simply use L and $h = L_0 - L$ to designate the solutions of problem (1.39). Multiplying both sides of relation (1.40) by w/h, we get:

$$\frac{w}{h}\Lambda_1 + \frac{wL}{h}\Lambda_2 = \frac{w}{h}\hat{L}_1 \tag{1.41}$$

Moreover, since $L^* = \Lambda(w, R + wL_0)$ and $\hat{L} = \hat{L}(w, \bar{U})$, the Marshallian and Hicksian elasticities of labor supply are respectively defined by:

$$\eta_M = -\frac{w}{h}\frac{\partial L^*}{\partial w} = -\frac{w}{h}(\Lambda_1 + L_0\Lambda_2) \quad \text{and} \quad \eta_H = -\frac{w\hat{L}_1}{h} \tag{1.42}$$

Comparing (1.41) and (1.42), we finally arrive at the equality:

$$\eta_M = \eta_H + \frac{wh}{R_0}\eta_{R_0} \tag{1.43}$$

In this expression $\eta_{R_0} = h\Lambda_2/R_0$ represents the Marshallian elasticity of labor supply with respect to potential income. Identity (1.43) is the Slutsky equation. It links the Hicksian elasticity η_H (also called compensated elasticity) to the Marshallian elasticity η_M (also called noncompensated elasticity).

7.4 Frischian, Hicksian, and Marshallian Elasticities of Labor Supply

This appendix presents definitions of the elasticities in the Hicksian, Marshallian, and Frischian senses in the intertemporal model of labor supply, as well as the relations among these three elasticities.

To simplify the calculations a little without prejudice to the generality of the results, we assume that $r_t = r$ for all $t \geq 0$. In that case, the intertemporal labor supply is the solution of the maximization of intertemporal utility $\sum_{t=0}^{t=T} U(C_t, L_t, t)$ under the intertemporal budget constraint. Assuming that the endowment of time at each date is equal to 1, this constraint is written:

$$\sum_{t=0}^{T}(1+r)^{-t}(C_t + w_t L_t) = \sum_{t=0}^{T}(1+r)^{-t}(w_t + B_t)$$

In this intertemporal model the different elasticities are defined with respect to the present values of wages. Hence $(1+r)^{-t}w_t$ represents the wage of date t actualized to date $t = 0$. The intertemporal budget constraint may be written $\sum_{t=0}^{T}\left[(1+r)^{-t}C_t + w_t^a L_t\right] = \sum_{t=0}^{T} w_t^a$, where $w_t^a = (1+r)^{-t}w$. So as to lighten the notations and facilitate comparison with the static model, we henceforth assume that w_t represents the present value of the wage of date t; the intertemporal budget constraint then simplifies to:

$$\sum_{t=0}^{T}\left[(1+r)^{-t}C_t + w_t L_t\right] = \Omega \quad \text{where } \Omega = \sum_{t=0}^{T}\left[w_t + (1+r)^{-t}B_t\right] \tag{1.44}$$

Let us denote by λ the multiplier associated to this constraint; the first-order conditions are written:

$$U_C(C_t, L_t, t) = \lambda(1+r)^{-t} \quad \text{and} \quad U_L(C_t, L_t, t) = \lambda w_t, \quad \text{for } t = 0, \ldots, T \tag{1.45}$$

This system of equations associated with the budget constraint defines the optimal solutions as functions of vector of wages $\mathbf{w} = (w_0, w_1, ..., w_T)$ and initial wealth Ω. We may thus denote them $C(\mathbf{w}, \Omega, t)$, $L(\mathbf{w}, \Omega, t)$, and $\lambda(\mathbf{w}, \Omega)$. As in the static case, the indirect utility function $V(\mathbf{w}, \Omega)$ is given by:

$$V(\mathbf{w}, \Omega) = \sum_{t=0}^{t=T} U\left[C(\mathbf{w}, \Omega, t), L(\mathbf{w}, \Omega, t), t\right] \tag{1.46}$$

Without loss of generality, we shall focus exclusively on the demand for leisure (and thus on labor supply) in the first period (the reasoning is the same for any period t). Denoting the wage for the first period w (rather than w_0) and omitting the time index $t = 0$, this demand for leisure is written $L(\mathbf{w}, \Omega)$, with $\mathbf{w} = (w, w_1, ..., w_T)$. It is often characterized as Marshallian.

7.4.1 ROY'S IDENTITY

Let us designate by V_{w_t} and V_Ω respectively the partial derivatives of the function $V(\mathbf{w}, \Omega)$ with respect to the wage w_t for the period t and with respect to wealth Ω.

Since by definition:

$$V(\mathbf{w}, \Omega) = \max_{\{C_t, L_t, \lambda\}} \sum_{t=0}^{t=T} U(C_t, L_t, t) + \lambda \left(\Omega - \sum_{t=0}^{T} \left[(1 + r)^{-t} C_t + w_t L_t\right] \right)$$

we obtain, using the envelope theorem:[12]

$$V_{w_t}(\mathbf{w}, \Omega) = -\lambda L_t(\mathbf{w}, \Omega) \tag{1.47}$$

$$V_\Omega(\mathbf{w}, \Omega) = \lambda \tag{1.48}$$

The elimination of the multiplier λ between relations (1.47) and (1.48) yields an equation known as Roy's identity. It is written:

$$L_t(\mathbf{w}, \Omega) = -\frac{V_{w_t}(\mathbf{w}, \Omega)}{V_\Omega(\mathbf{w}, \Omega)} \tag{1.49}$$

7.4.2 THE SLUTSKY EQUATION

As with the static case studied in appendix 3, Hicksian demands may be defined as demand functions that minimize the total cost under the constraint of a minimal exogenous level of utility, denoted $\bar{U}$. They are thus solutions of the problem:

$$\min_{(L_t, C_t)} \sum_{t=0}^{T} \left[(1 + r)^{-t} C_t + w_t L_t\right] \quad \text{subject to} \quad \sum_{t=0}^{t=T} U(C_t, L_t, t) \geq \bar{U}$$

[12] Let us denote $\mathcal{L}(C_1, ..., C_T, L_1, ..., L_T, \lambda, \mathbf{w}, \Omega)$ the right-hand side of this last relation; by the envelope theorem, we have:

$$V_{w_t} = \frac{\partial \mathcal{L}}{\partial w_t} \quad \text{and} \quad V_\Omega = \frac{\partial \mathcal{L}}{\partial \Omega}$$

Let μ be the multiplier associated with the constraint of this problem; the first-order conditions are written:

$$(1+r)^{-t} = \mu U_C(C_t, L_t, t) \quad \text{and} \quad w_t = \mu U_L(C_t, L_t, t) \tag{1.50}$$

We will denote the Hicksian demands by $C^H(\mathbf{w}, \bar{U}, t)$ and $L^H(\mathbf{w}, \bar{U}, t)$. The expenditure function $e(\mathbf{w}, \bar{U})$ is then defined by $e(\mathbf{w}, \bar{U}) = \sum_{t=0}^{T} \left[(1+r)^{-t} C^H(\mathbf{w}, \bar{U}, t) + w_t L^H(\mathbf{w}, \bar{U}, t)\right]$. Deriving the two members of this last equality with respect to wage w_t, we get:

$$e_{w_t}(\mathbf{w}, \bar{U}) = \sum_{s=0}^{T} \left[(1+r)^{-s} \frac{\partial C_s^H}{\partial w_t} + w_s \frac{\partial L_s^H}{\partial w_t}\right] + L^H(\mathbf{w}, \bar{U}, t) \tag{1.51}$$

By deriving with respect to w_t the budget constraint at equilibrium, $\sum_{t=0}^{t=T} U(C_t^H, L_t^H, t) = \bar{U}$, and then utilizing the first-order conditions (1.50) and relation (1.51), we arrive at the identity:

$$e_{w_t}(\mathbf{w}, \bar{U}) = L^H(\mathbf{w}, \bar{U}, t)$$

As with the static case set out in appendix 3, this relation indicates that the Hicksian demand for leisure for date t corresponds to the derivative of the expenditure function with respect to the wage for the period.

Additionally, the Marshallian and Hicksian demands for leisure satisfy the relation:

$$L(\mathbf{w}, e(\mathbf{w}, \bar{U}), t) = L^H(\mathbf{w}, \bar{U}, t)$$

Deriving with respect to w_t, we get:

$$\frac{\partial L_t}{\partial w_t} + e_{w_t} \frac{\partial L_t}{\partial \Omega} = \frac{\partial L_t^H}{\partial w_t}$$

Since $e_{w_t}(\mathbf{w}, \bar{U}) = L^H(\mathbf{w}, \bar{U}, t) = L(\mathbf{w}, e(\mathbf{w}, \bar{U}), t) \equiv L_t$, we arrive at the Slutsky equation:

$$\frac{\partial L_t}{\partial w_t} = \frac{\partial L_t^H}{\partial w_t} - L_t \frac{\partial L_t}{\partial \Omega}$$

If we shift to supplies of labor $h_t = 1 - L_t$, the Slutsky equation becomes:

$$\frac{\partial h_t^H}{\partial w_t} = \frac{\partial h_t}{\partial w_t} + (1 - h_t) \frac{\partial h_t}{\partial \Omega}$$

Thus we return to equation (1.5) from the static model, which was written using the elasticities.

7.4.3 MARSHALLIAN AND HICKSIAN ELASTICITIES

Similar to the procedure followed in appendix 7.3 for the static model, we will now decompose the impact of a wage variation on the labor supply for the first period, distinguishing between a substitution effect and a wealth effect. Because the model is intertemporal, it is necessary to specify whether the wage variation affects solely the current wage or the wage at other dates. We will begin by taking the case of a permanent variation in the wage, where all the wages of vector $\mathbf{w}$ are modified by an amount $\mathrm{d}\mathbf{w}$. With that as our point of departure, it will be easy to shift our focus to the situation where only the wage for the current period varies.

By definition, the Hicksian elasticity of the labor supply in the initial period, denoted η_H, then corresponds to the quantity $\frac{w}{h}\frac{\partial h^H}{\partial w}$. Now, when all wages vary by the same amount $\mathrm{d}w$ we have $\eta_H = \sum_{t=0}^{T} \frac{w_t}{h}\frac{\partial h^H}{\partial w_t}$. Using the Slutsky equation, the Hicksian elasticity is given by:

$$\eta_H = \sum_{t=0}^{T} \frac{w_t}{h}\frac{\partial h}{\partial w_t} + \frac{\sum_{t=0}^{T} w_t(1-h_t)}{\Omega}\eta_\Omega \tag{1.52}$$

Here $\eta_\Omega = \frac{\Omega}{h}\frac{\partial h}{\partial \Omega}$ designates the elasticity of the labor supply in the first period with respect to wealth. Under the hypothesis that leisure is a normal good, the result is $\eta_\Omega < 0$.

The Marshallian elasticity, denoted η_M, assesses the total effect of a permanent variation in the wages vector $\mathbf{w}$ on the (Marshallian) supply of labor $h(\mathbf{w}, \Omega)$. It is thus defined by $\eta_M = \frac{w}{h}\frac{\mathrm{d}h}{\mathrm{d}w}$. Now, $\frac{\mathrm{d}h}{\mathrm{d}w} = \sum_{t=0}^{T} \frac{\mathrm{d}h}{\mathrm{d}w_t}\frac{\mathrm{d}w_t}{\mathrm{d}w} = \sum_{t=0}^{T} \frac{\mathrm{d}h}{\mathrm{d}w_t}$ and therefore $\eta_M = \sum_{t=0}^{T} \frac{w_t}{h}\frac{\mathrm{d}h}{\mathrm{d}w_t}$. We thus have:

$$\frac{\mathrm{d}h}{\mathrm{d}w_t} = \frac{\partial h}{\partial w_t} + \frac{\partial h}{\partial \Omega}\frac{\partial \Omega}{\partial w_t}$$

Since $\frac{\partial \Omega}{\partial w_t} = 1$, we arrive at:

$$\eta_M = \sum_{t=0}^{T} \frac{w_t}{h}\frac{\partial h}{\partial w_t} + \frac{\sum_{t=0}^{T} w}{\Omega}\eta_\Omega \tag{1.53}$$

By eliminating $\sum_{t=0}^{T} \frac{w_t}{h}\frac{\partial h}{\partial w_t}$ between (1.52) and (1.53), we find a relation between the Marshallian and Hicksian elasticities. It is written:

$$\eta_M = \eta_H + \frac{\sum_{t=0}^{T} w_t h_t}{\Omega}\eta_\Omega \tag{1.54}$$

If only the current-period wage varies, this relation is identical to that of the static model:

$$\eta_M = \eta_H + \frac{wh}{\Omega}\eta_\Omega$$

Since $\eta_\Omega < 0$ (taking the habitual case in which leisure is a normal good), we have $\eta_M < \eta_H$. Thus we obtain the relation between the Marshallian and Hicksian elasticities when all wages vary by an amount $\mathrm{d}w$. This equation demonstrates that the impact

of a wage variation on labor supply may be broken down, as in the static model, into a substitution effect with a constant intertemporal utility that is positive (it is represented by the Hicksian elasticity η_H) and a wealth effect that is negative if leisure is a normal good (it is represented by the term $\frac{wh}{\Omega}\eta_\Omega$).

7.4.4 Frischian Elasticity

It is also possible to break the impact of a wage variation down into an intertemporal substitution effect, with the marginal utility of wealth held constant, and a wealth effect. The intertemporal substitution effect is defined on the basis of the Frisch function, which assumes that the marginal utility of wealth—which is simply the multiplier λ according to relation (1.48)—remains constant when wages vary. As regards the Frischian demand for leisure in the first period, the first-order conditions (1.45) show that it depends only on the current wage w and on λ, so we may denote it $L^F(w,\lambda)$. This is the setting in which we will obtain a relation among the Frischian, Marshallian, and Hicksian elasticities.

Frischian labor supply is given by $h^F(w,\lambda) = 1 - L^F(w,\lambda)$. For all values of λ, thanks to relation (1.48), a Frischian expenditure function, denoted $e^F(\mathbf{w},\lambda)$, may be defined such that:

$$V_\Omega\left[\mathbf{w}, e^F(\mathbf{w},\lambda)\right] = \lambda \tag{1.55}$$

Frischian labor supply and Marshallian labor supply are then linked by the relation:

$$h^F(w,\lambda) = h\left[\mathbf{w}, e^F(\mathbf{w},\lambda)\right] \tag{1.56}$$

It is possible to calculate the relation between the Frischian and Marshallian elasticities if we suppose that all the wages of vector $\mathbf{w}$ vary by an amount $\mathrm{d}w$. Differentiating equation (1.56) gives us an expression of the Frischian elasticity, denoted η_F, for labor supply in the first period:

$$\eta_F = \frac{w}{h}\frac{\mathrm{d}h^F}{\mathrm{d}w} = \sum_{t=0}^{T}\frac{w}{h}\frac{\partial h}{\partial w_t} + \frac{w}{h}\frac{\partial h}{\partial \Omega}\sum_{t=0}^{T} e^F_{w_t} \tag{1.57}$$

In this last relation $e^F_{w_t}$ is the partial derivative of the expenditure function with respect to wage w_t. We can obtain an expression of $e^F_{w_t}$ by deriving (1.55) with respect to w_t, which takes the form:

$$e^F_{w_t}(\mathbf{w},\lambda) = -\frac{V_{\Omega w_t}\left[\mathbf{w}, e^F(\mathbf{w},\lambda)\right]}{V_{\Omega\Omega}\left[\mathbf{w}, e^F(\mathbf{w},\lambda)\right]}$$

And so:

$$\sum_{t=0}^{T} e^F_{w_t}(\mathbf{w},\lambda) = -\frac{\sum_{t=0}^{T} V_{\Omega w_t}\left[\mathbf{w}, e^F(\mathbf{w},\lambda)\right]}{V_{\Omega\Omega}\left[\mathbf{w}, e^F(\mathbf{w},\lambda)\right]} \tag{1.58}$$

Moreover, we can write Roy's identity as follows: $V_{w_t}(\mathbf{w}, \Omega) = -L_t(\mathbf{w}, \Omega)V_\Omega(\mathbf{w}, \Omega)$. Deriving this last equality with respect to Ω, we find:

$$V_{\Omega w_t} = -V_R \frac{\partial L_t}{\partial \Omega} - L_t V_{\Omega\Omega} = V_\Omega \frac{\partial h_t}{\partial \Omega} - (1 - h_t)V_{\Omega\Omega}$$

If we carry this value of $V_{\Omega w_t}$ into (1.58), we get:

$$\sum_{t=0}^{T} e^F_{w_t} = -\frac{V_\Omega}{V_{\Omega\Omega}} \sum_{t=0}^{T} \frac{\partial h_t}{\partial \Omega} + \sum_{t=0}^{T}(1 - h_t)$$

Finally, if we carry this value of $\sum_{t=0}^{T} e^F_{w_t}$ into (1.57), the result is:

$$\eta_F = \sum_{t=0}^{T} \frac{w}{h}\frac{\partial h}{\partial w_t} + \frac{w}{h}\frac{\partial h}{\partial \Omega}\sum_{t=0}^{T}(1 - h_t) - \frac{V_\Omega}{V_{\Omega\Omega}}\sum_{t=0}^{T}\frac{w}{h}\frac{\partial h}{\partial \Omega}\frac{\partial h_t}{\partial \Omega}$$

The sum of the first two terms of the right-hand side of this equation exactly matches the Hicksian elasticity; see (1.52). For the third term of the right-hand side, we note that:

$$\sum_{t=0}^{T} \frac{w}{h}\frac{\partial h}{\partial \Omega}\frac{\partial h_t}{\partial \Omega} = \frac{1}{\Omega}\sum_{t=0}^{T}\frac{wh_t}{\Omega}\eta^h_\Omega \eta^{h_t}_\Omega$$

In the expression of η_F we then see the opposite of the elasticity of intertemporal substitution, that is $\gamma = -\frac{\Omega V_\Omega}{V_{\Omega\Omega}}$, which is also the inverse of the index relating to the risk aversion of Arrow and Pratt. With the notation we have used to this point, $\eta_\Omega = \eta^h_\Omega$, we finally have:

$$\eta_F = \eta_H + \gamma\eta_\Omega \sum_{t=0}^{T} \frac{wh_t}{\Omega}\eta^{h_t}_\Omega \tag{1.59}$$

On the assumption that leisure is a normal good, then $\eta_\Omega \eta^{h_t}_\Omega \geq 0$, and we thus have $\eta_H \leq \eta$. As we have seen that $\eta_M \leq \eta_H$, we find that the different elasticities are ranked in the following order:

$$\eta_M \leq \eta_H \leq \eta_F$$

It is of interest to note that in the absence of income effect, that is if $\eta_\Omega = \eta^{h_t}_\Omega = 0$ for all t, then $\eta_M = \eta_H = \eta_F$.

One can also obtain, on the basis of equations (1.54) and (1.59), a breakdown of the Marshallian elasticity that measures the total impact of the wage variation on the labor supply:

$$\eta_M = \eta_F - \gamma\eta_\Omega \sum_{t=0}^{T} \frac{wh_t}{\Omega}\eta^{h_t}_\Omega + \frac{\sum_{t=0}^{T} wh_t}{\Omega}\eta_\Omega$$

When the wage variation affects only the current period, that is, when the wages for the other dates are held constant, this relation is:

$$\eta_M = \eta_F + \frac{wh}{\Omega}\eta_\Omega(1 - \gamma\eta_\Omega)$$

7.5 Sample Selection

Labor market participation decisions imply problems of sample selection: we observe hours and wages for a subset of the population, but the sample is truncated because it depends on another variable, namely, participation. This raises two key questions: (1) What market wage distribution should be used for nonparticipants, and (2) Are labor supply behaviors at the extensive margin (participation) fundamentally different from behavior at the intensive margin (hours of work)? Among the most compelling reasons for separating these two margins is the existence of the fixed costs of participating in the labor market: entering the market requires looking for a job, finding solutions for child care, reorganizing household activities, and so on. Two possibilities are available: a joint estimation of extensive and intensive margins or a correction of the sample selection in estimating the intensive margin (see Blundell and MaCurdy, 1999; Blundell et al., 2007; Wooldridge, 2010, 2013).

7.5.1 Joint Estimation

Suppose individual heterogeneity in tastes for work so that:

$$h = \begin{cases} >0 & \text{if } w > w_A \\ 0 & \text{otherwise} \end{cases}$$

where w_A is the reservation wage. Assume that the utility of a consumer will then take the form $C^{1-\beta}L^{\beta}$, $1 > \beta > 0$, while the budget constraint continues to be written $C + wL = wL_0 + R$. The preference for work is expressed by the coefficient β according to the linear form $\beta = \mathbf{x}\boldsymbol{\theta} + \varepsilon$, where ε is a normally distributed random term. Following the static model of section 1.1.1, we know that the reservation wage w_A is equal to the marginal rate of substitution U_L/U_C taken at point (R, L_0) and that the maximization of utility subject to the budget constraint gives the optimal value of leisure. After several simple calculations, we find that:

$$w_A = \frac{\beta}{1-\beta}\frac{R}{L_0} \quad \text{and} \quad L = \begin{cases} \beta\left(L_0 + \frac{R}{w}\right) & \text{if } w \geq w_A \\ L_0 & \text{if } w < w_A \end{cases}$$

Since the coefficient β is a function of the random term ε, the inequality $w \geq w_A$ is equivalent to an inequality of the values of ε, which is written:

$$w \geq w_A \Longleftrightarrow \varepsilon \leq \frac{wL_0}{R + wL_0} - \mathbf{x}\boldsymbol{\theta}$$

In conclusion, decisions concerning labor supply $h = L_0 - L$, and participation may be summed up in this fashion:

$$h = \begin{cases} L_0 - (\mathbf{x}\boldsymbol{\theta} + \varepsilon)\left(L_0 + \frac{R}{w}\right) & \text{if } \varepsilon \leq \frac{wL_0}{R+wL_0} - \mathbf{x}\boldsymbol{\theta} \\ 0 & \text{if } \varepsilon > \frac{wL_0}{R+wL_0} - \mathbf{x}\boldsymbol{\theta} \end{cases} \tag{1.60}$$

This expression of labor supply is related, as regards the interior solution, to the basic equation (1.20). But we see that taking into account participation decisions constrains the variations of the random term, making them depend on explanatory variables. In these circumstances, the use of ordinary least squares is seen to be inadequate. Let us now suppose that in the available sample of individuals, N in size, individuals $i = 1, ..., J$ have worked h_i hours and that individuals $i = J + 1, ..., N$ have not worked. Let us denote by $\Phi(.)$ and $\phi(.)$ respectively the cumulative distribution function and the probability density of the random term ε. It is then possible to write the likelihood of the sample. Following rule (1.60) giving the optimal decisions of an agent, when an individual i has worked h_i hours, that means that the random term has taken the value $\varepsilon_i = w_i(L_0 - h_i)/(R_i + w_iL_0) - \mathbf{x}_i\boldsymbol{\theta}$. In this case its contribution to the likelihood of the sample is equal to $\phi(\varepsilon_i)$. If agent i has not worked, that means that the random term is bounded above by the value $\tilde{\varepsilon}_i = [w_iL_0/(R_i + w_iL_0)] - \mathbf{x}_i\boldsymbol{\theta}$. In this case, its contribution to the likelihood of the sample is given by $\Pr\{h_i = 0\} = 1 - \Phi(\tilde{\varepsilon}_i)$. Setting $\overline{\Phi} = 1 - \Phi$, the likelihood function of the sample is written in logarithmic form:

$$\mathcal{L} = I(h>0)\,\mathbb{E}[\varepsilon | w \geq w_A] + I(h=0)\mathbb{E}[\varepsilon | w < w_A] \tag{1.61}$$

$$= \sum_{i=1}^{i=J} \ln \phi \left[\frac{w_i(L_0 - h_i)}{R_i + w_iL_0} - \mathbf{x}_i\boldsymbol{\theta} \right] + \sum_{i=J+1}^{i=N} \ln \overline{\Phi} \left[\frac{w_iL_0}{R_i + w_iL_0} - \mathbf{x}_i\boldsymbol{\theta} \right] \tag{1.62}$$

where I is an indicator function and $\boldsymbol{\theta}$ is the unknown parameters of preferences. In a linear specification, when ε is i.i.d. and normally distributed, this is equivalent to the Tobit censored regression estimation (Blundell, MaCurdy and Megher, 2007, p. 4679).

The expression (1.61) of the likelihood function also highlights a delicate problem. By definition, the econometrician does not observe the wages of individuals $i = J + 1, ..., N$ who do not work. However, relation (1.61) shows that it is necessary to attribute a fictitious wage to these individuals if we want to maximize the likelihood function. We thus need to be able to assign a quantity to the (unobserved) wage notionally offered to an individual, which she has refused. The most common solution at present consists of deducing the wage of a nonparticipant using the wage received by participants with similar characteristics in terms of educational qualification, experience, age, and so on. In practice we can explain the wages of individuals participating in the labor market by a regression of the type $w_i = \mathbf{y}_i\boldsymbol{\theta}_p + u_i$ in which the vector $\mathbf{y}_i$ represents the characteristics of an individual i participating in the labor market and $\boldsymbol{\theta}_p$ designates the vector of the parameters to be estimated. Let us use $\hat{\boldsymbol{\theta}}_p$ to denote the vector of the estimates of $\boldsymbol{\theta}$; we can then use this vector $\hat{\boldsymbol{\theta}}_p$ to calculate the wage w_k of a nonparticipant k, using the vector $\mathbf{y}_k$ of her characteristics and setting $w_k = \mathbf{y}_k\hat{\boldsymbol{\theta}}_p$. This simple technique unfortunately presents a *selection bias*, since it assumes that the regression equation $w_i = \mathbf{y}_i\boldsymbol{\theta}_p + u_i$ also applies to the fictitious wages of nonparticipants.

This hypothesis is highly likely to be erroneous, inasmuch as participants in the labor market must on average have unobserved characteristics that allow them to demand wages higher than those that nonparticipants can demand. Formally, this means that the distribution of the random disturbance u_i should not be the same for participants and nonparticipants. The distribution that applies to participants ought to weight the high values of the random factor more strongly than the one that applies to nonparticipants and consequently the estimation procedure described previously will *overestimate* the fictitious wage attributable to a nonparticipant. One way to correct this bias is to make simultaneous estimations of equations explaining wages and decisions to supply labor (see Heckman, 1974, for an application).

7.5.2 Sample Selection Correction

A simpler technique for coping with sample selection is provided by the Heckman (1976, 1979), or *Heckit*, method. The idea is to estimate the intensive margin equations (1.20) or (1.22) after adding a variable that measures the degree of truncation of the sample. Consider the system of equations which defines the hours of work h_{it} and the participation decision p_{it} of individual i at date t:

$$h_{it} = \mathbf{x}_{it}\boldsymbol{\alpha}_x + \varepsilon_{it} \tag{1.63}$$

$$p_{it} = \mathbb{I}(\mathbf{z}_{it}\boldsymbol{\alpha}_z + \upsilon_{it}) \tag{1.64}$$

with $\mathbb{E}(\varepsilon|\mathbf{x}) = 0$, $\mathbb{E}(\upsilon|\mathbf{z}) = 0$, and $\mathbb{I}(y)$ as an indicator function equal to 1 if $y > 0$ and equal to zero otherwise. In this context, $p_{it} = 1$ if individual i at time t has positive hours $h_{it} > 0$ and $p_{it} = 0$ otherwise. Hence, participation depends on a set of observed variables $\mathbf{z}$. We assume further that $\mathbf{z}$ contains $\mathbf{x}$, that is, that there are some variables in $\mathbf{z}$ not in $\mathbf{x}$ and that all variables in $\mathbf{x}$ are also in $\mathbf{z}$. This is called the *exclusion restriction*. It is assumed that:

$$\mathbb{E}(\varepsilon|\mathbf{x}, \mathbf{z}) = 0$$

We also assume that υ has a normal distribution with zero mean and variance equal to 1. We can easily see that correlation between ε and υ generally causes a sample selection problem. To understand why, assume that ε and υ are independent of $\mathbf{z}$. Then, taking the expectation of (1.63), conditional on $\mathbf{z}$ and υ, and using the fact that $\mathbf{x}$ is a subset of $\mathbf{z}$, gives:

$$\mathbb{E}(h|\mathbf{z}, \mathbf{x}, \upsilon) = \mathbf{x}\boldsymbol{\alpha}_x + \mathbb{E}(\varepsilon|\mathbf{z}, \upsilon) = \mathbf{x}\boldsymbol{\alpha}_x + \mathbb{E}(\varepsilon|\upsilon)$$

where $\mathbb{E}(\varepsilon|\mathbf{z}, \upsilon) = \mathbb{E}(\varepsilon|\upsilon)$ because ε and υ are independent of $\mathbf{z}$. Now, assuming that $\mathbb{E}(\varepsilon|\upsilon) = \rho\upsilon$ for some parameter ρ (which is the case if ε and υ are jointly normal), we get:

$$\mathbb{E}(h|\mathbf{z}, \mathbf{x}, \upsilon) = \mathbf{x}\boldsymbol{\alpha}_x + \rho\upsilon$$

We do not observe v, but we can use this equation to compute $\mathbb{E}(h|\mathbf{z}, p)$ and then take the case where $p = 1$:

$$\mathbb{E}(h|\mathbf{z}, \mathbf{x}, p) = \mathbf{x}\boldsymbol{\alpha}_x + \rho\mathbb{E}(v|\mathbf{z}, p)$$

Because p and v are related by (1.64), and v has a standard normal distribution, which cumulative density function is denoted by Φ, we have:

$$\mathbb{E}(v|\mathbf{z}, p = 1) = \lim_{\Delta \to 0} \frac{\Pr(\mathbf{z}\boldsymbol{\alpha}_z \leq v \leq \mathbf{z}\boldsymbol{\alpha}_z + \Delta)}{\Pr(v > -\mathbf{z}\boldsymbol{\alpha}_z)} = \frac{\Phi'(\mathbf{z}\boldsymbol{\alpha}_z)}{\Phi(\mathbf{z}\boldsymbol{\alpha}_z)}$$

The ratio $\Phi'(\mathbf{z}\boldsymbol{\alpha}_z)/\Phi(\mathbf{z}\boldsymbol{\alpha}_z)$ is known as the inverse Mills ratio, denoted $\lambda(\mathbf{z}\boldsymbol{\alpha}_z)$. This leads to the equation:

$$\mathbb{E}(h|\mathbf{z}, \mathbf{x}, p = 1) = \mathbf{x}\boldsymbol{\alpha}_x + \rho\lambda(\mathbf{z}\boldsymbol{\alpha}_z) \tag{1.65}$$

Equation (1.65) shows that the expected value of h, given the determinants of participation, which also include the determinants of hours, $\mathbf{z}$, and conditional on the observability of hours, is simply a function of $\mathbf{x}$ plus a correction factor that depends on the inverse Mills ratio evaluated at $\mathbf{z}$. Hence we can estimate without bias $\boldsymbol{\alpha}_x$ using just the truncated sample if we include the inverse Mills ratio $\lambda(\mathbf{z}\boldsymbol{\alpha}_z)$. If $\rho = 0$, this means that there is no issue of sample selection bias. This happens when ε and v are uncorrelated. Using the OLS, this will show whether the estimated coefficient of the inverse Mills ratio is significantly different from zero. On the contrary, if $\rho \neq 0$ it means that there is an issue, and we would omit a variable if we did not include $\lambda(\mathbf{z}\boldsymbol{\alpha}_z)$. How then to calculate the inverse Mills ratio? From the assumptions we have made, p given $\mathbf{z}$ follows a probit model:

$$\Pr(p = 1|\mathbf{z}) = \Phi(\mathbf{z}\boldsymbol{\alpha}_z) \tag{1.66}$$

and we can estimate $\boldsymbol{\gamma}$ using the entire sample (of those participating and not participating), and then calculate λ for each participating individual.

To summarize, the Heckit method is implemented as follows:

Step 1: Obtain the probit estimates $\hat{\boldsymbol{\alpha}}_z$ from the model (1.66) using all observations.

Step 2: Compute the estimated inverse Mills ratio $\lambda(\mathbf{z}\hat{\boldsymbol{\alpha}}_z) = \Phi'(\mathbf{z}\hat{\boldsymbol{\alpha}}_z)/\Phi(\mathbf{z}\hat{\boldsymbol{\alpha}}_z)$.

Step 3: Estimate $\hat{\boldsymbol{\alpha}}_x$ and $\hat{\rho}$ from the OLS estimation of equation (1.65).

REFERENCES

Arrow, K. (1965). The theory of risk aversion. In *Aspects of the theory of risk bearing*, by Y. Saatio, Helsinki. Reprinted in *Essays in the theory of risk bearing* (pp. 90–102). Chicago, IL: Markham Publ. Co., 1971.

Baker, M., Gruber, J., & Milligan, K. (2008). Universal child care, maternal labor supply, and family well-being. *Journal of Political Economy*, *116*(4), 709–745.

Bailey, M. (2006). More power to the pill: The impact of contraceptive freedom on women's life cycle labor supply. *Quarterly Journal of Economics*, *121*(1), 289–320.

Becker, G. (1965). A theory of the allocation of time. *Economic Journal*, *75*, 493–517.

Blanchard, O., & Fischer, S. (1989). *Lectures on macroeconomics*. Cambridge, MA: MIT Press.

Blau, F., & Kahn, L. (2007). Changes in the labor supply behavior of married women: 1980–2000. *Journal of Labor Economics*, *25*(3), 393–438.

Bloom, D., Canning, D., Fink, G., & Finlay, J. (2009). Fertility, female labor force participation, and the demographic dividend. *Journal of Economic Growth*, *14*(2), 79–101.

Blundell, R., Chiappori, A., Magnac, T., & Meghir, C. (2007). Collective labour supply: Heterogeneity and nonparticipation. *Review of Economic Studies*, *74*, 417–447.

Blundell, R., Duncan, A., & Meghir, C. (1992). Taxation and empirical labour supply models: Lone parents in the UK. *Economic Journal*, *102*, 265–278.

Blundell, R., Duncan, A., & Meghir, C. (1998). Estimation of labour supply responses using tax policy reforms. *Econometrica*, *66*(4), 827–861.

Blundell, R., & MaCurdy, T. (1999). Labor supply: A review of alternative approaches. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 3A, chap. 27). Amsterdam: Elsevier Science.

Blundell, R., MaCurdy, T., & Meghir, C. (2007). Labor supply models: Unobserved heterogeneity, nonparticipation and dynamics. In J. J. Heckman & E. Leamer (Eds.), *Handbook of econometrics* (vol. 6A, chap. 69). New York, NY: Elsevier.

Blundell, R., Meghir, C., & Neves, P. (1993). Labour supply and intertemporal substitution. *Journal of Econometrics*, *59*(1–2), 137–160.

Blundell, R., Meghir, C., Symons, E., & Walker, I. (1988). Labour supply specification and the evaluation of tax reforms. *Journal of Public Economics*, *36*, 23–52.

Blundell, R., & Walker, I. (1986). A life cycle consistent empirical model of labour supply using cross section data. *Review of Economic Studies*, *53*, 539–558.

Brown, C. (1999). Early retirement windows. In O. Mitchell, B. Hammond, & A. Rappaport (Eds.), *Forecasting retirement needs and retirement wealth*. Philadelphia: University of Pennsylvania Press.

Browning, M., Bourguignon, F., Chiappori, P.-A., & Lechène, V. (1994). Income and outcomes: A structural model of intrahousehold allocation. *Journal of Political Economy*, *102*, 1067–1096.

Browning, M., Chiappori, P.-A., & Weiss, Y. (2012). *Family economics*. New York, NY: Cambridge University Press.

Chetty, R. (2012). Bounds on elasticities with optimization frictions: A synthesis of micro and macro evidence on labor supply. *Econometrica*, *80*(3), 969–1018.

Chetty, R., Friedman, J., Olsen, T., & Pistaferri, L. (2011a). Adjustment costs, firm responses, and micro vs. macro labor supply elasticities: Evidence from Danish tax records. *Quarterly Journal of Economics*, *126*(2), 749–804.

Chetty, R., Guren, A., Manoli, D., & Weber, A. (2011b). Are micro and macro labor supply elasticities consistent? A review of evidence on the intensive and extensive margins. *American Economic Review, Papers and Proceedings*, *101*, 471–475.

Chetty, R., Guren, A., Manoli, D., & Weber, A. (2013). Does indivisible labor explain the difference between micro and macro elasticities? A meta-analysis of extensive margin elasticities. *NBER Macroeconomics Annual*, *27*, 1–56.

Chiappori, P.-A. (1988). Rational household labor supply. *Econometrica*, *56*(1), 63–89.

Chiappori, P.-A. (1992). Collective labor supply and welfare. *Journal of Political Economy*, *100*, 437–467.

Coile, C., & Gruber, J. (2007). Future social security entitlements and the retirement decision. *Review of Economics and Statistics*, *89*(2), 234–246.

Connolly, M. (2008). Here comes the rain again: Weather and the intertemporal substitution of leisure. *Journal of Labor Economics*, *26*(1), 73–100.

Cowell, F. (2006). *Microeconomics: Principles and Analysis*. New York, NY: Oxford University Press.

Del Boca, D., Pasqua, S., & Pronzato, C. (2009). Motherhood and market work decisions in institutional context: A European perspective. *Oxford Economic Papers*, *61*, 147–171.

Devereux, P. (2004). Changes in relative wages and family labor supply. *Journal of Human Resources*, *39*(3), 696–722.

Ehrenberg, R., & Smith, R. (1994). *Modern labor economics: Theory and public policy* (5th ed). New York, NY: HarperCollins.

Fortin, B., & Lacroix, G. (1997). A test of neoclassical and collective models of household labor supply. *Economic Journal*, *107*, 933–955.

Francis, N., & Ramey, V. (2009). Measures of per capita hours and their implications for the technology-hours debate. *Journal of Money, Credit and Banking*, *41*(6), 1071–1097.

French, E., & Jones, J. (2011). The effects of health insurance and self-insurance on retirement behavior. *Econometrica*, *79*(3), 693–732.

González-Chapela, J. (2007). On the price of recreation goods as a determinant of male labor supply. *Journal of Labor Economics*, *25*(4), 795–824.

Greenwood, J., Seshadri, A., & Yorukoglu, M. (2005). Engines of liberation. *Review of Economic Studies*, *72*(1), 109–133.

Greenwood, J., & Vandenbroucke, G. (2008). Hours worked (long-run trends). In L. Blume & S. Durlauf (Eds.), *The new Palgrave dictionary of economics* (2nd ed., vol. 4, pp. 75–81). New York, NY: Palgrave Macmillan.

Gronau, R. (1986). Home production. In O. Ashenfelter & R. Layard (Eds.), *Handbook of labor economics* (vol. 1, chap. 4). Amsterdam: North-Holland.

Gronau, R. (1997). The theory of home production: The past ten years. *Journal of Labor Economics*, *15*(2), 197–205.

Gruber, J., & Wise, D. (Eds.) (1999). *Social security and retirement around the world*. Chicago, IL: University of Chicago Press.

Gruber, J., & Wise, D. (2001). An international perspective on policies for an aging society (Working Paper No. 8103). National Bureau of Economic Research, Cambridge, MA.

Gruber, J., & Wise, D. (2002). Social security programs and retirement around the world: Micro estimation (Working Paper No. 9407). National Bureau of Economic Research, Cambridge, MA.

Gustman, A., Mitchell, O., & Steinmeier, T. (1994). The role of pensions in the labor market. *Industrial and Labor Relations Review*, *47*(3), 417–438.

Gustman, A., & Steinmeier, T. (1986). A structural retirement model. *Econometrica*, *54*(3), 555–584.

Hall, R. (1980). Labor supply and aggregate fluctuations. In K. Brunner & A. Meltzer (Eds.), *On the state of macroeconomics*, Carnegie-Rochester Conference Series on Public Policy. Amsterdam: North-Holland.

Hall, R. (1999). Labor market frictions and employment fluctuations. In J. Taylor & M. Woodford (Eds.), *Handbook of macroeconomics* (vol. IB, chap. 17, pp. 1137–1170). Amsterdam: North-Holland.

Hansen, G. (1985). Indivisible labor and the business cycle. *Journal of Monetary Economics*, *16*, 309–337.

Heckman, J. (1974). Shadow prices, market wages and labor supply. *Econometrica*, *42*(4), 679–694.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, *5*(4), 475–492.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*, 153–161.

Heckman, J., Lalonde, R., & Smith, J. (1999). The economics and econometrics of active labor market programs. In A. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 3, chap. 31, pp. 1865–2097). Amsterdam: Elsevier Science.

Heim, B. (2007). The incredible shrinking elasticities: Married female labor supply, 1978–2002. *Journal of Human Resources*, *42*(4), 881–918.

Hotz, J., Kydland, F., & Sedlacek, G. (1988). Intertemporal substitution and labor supply. *Econometrica*, *56*, 335–360.

Imai, S., & Keane, M. (2004). Intertemporal labor supply and human capital accumulation. *International Economic Review*, *45*, 601–642.

Keane, M. (2011). Labor supply and taxes. *Journal of Economic Literature*, *49*(4), 961–1075.

Keane, M., & Rogerson, R. (2012). Micro and macro labor supply elasticities: A reassessment of conventional wisdom. *Journal of Economic Literature*, *50*(2), 464–476.

Lucas, R., & Rapping, L. (1969). Real wages, employment and inflation. *Journal of Political Economy*, *77*, 721–754.

Lumsdaine, R., & Mitchell, O. (1999). New developments in the economic analysis of retirement. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 3, chap. 49, pp. 3261–3307). Amsterdam: Elsevier Science.

Lumsdaine, R., Stock, J., & Wise, D. (1990). Efficient windows and labor force reduction. *Journal of Public Economics*, *43*, 131–159.

Lundberg, S. (1985). The added worker effect. *Journal of Labor Economics*, *3*, 11–37.

Lundberg, S., Pollak, R., & Wales, T. (1997). Do husbands and wives pool their resources? Evidence from the United Kingdom child benefit. *Journal of Human Resources*, *32*(3), 463–480.

MaCurdy, T. (1983). A simple scheme for estimating an intemporal model of labor supply and consumption in the presence of taxes and uncertainty. *International Economic Review*, *24*(2), 265–289.

MaCurdy, T., Green, D., & Paarsch, H. (1990). Assessing empirical approaches for analyzing taxes and labour supply. *Journal of Human Resources*, *25*, 415–490.

Maddison, A. (1995). *The world economy, 1820–1992*. Paris: OECD Publishing.

Marchand, O., & Thélot, C. (1997). *Le travail en france (1800–2000)*. Paris: Nathan.

Mas-Colell, A., Whinston, M., & Green, J. (1995). *Microeconomic theory*. Oxford, U.K.: Oxford University Press.

OECD. (1991). *OECD employment outlook*. Paris: OECD Publishing.

OECD. (1995). *OECD employment outlook*. Paris: OECD Publishing.

OECD. (1999). *OECD employment outlook*. Paris: OECD Publishing.

OECD. (2011). *OECD pensions at a glance*. Paris: OECD Publishing.

Oreffice, S., & Quintana-Domeque, C. (2012). Fat spouses and hours of work: Are body and Pareto weights correlated? *IZA Journal of Labor Economics*, *1*(4). doi: 10.1184/2193-8997-1-6.

Pencavel, J. (1986). Labor supply of men: A survey. In O. Ashenfelter & R. Layard (Eds.), *Handbook of labor economics* (vol. 1, pp. 3–102). Amsterdam: North-Holland.

Pratt, J. (1964). Risk aversion in the small and in the large. *Econometrica*, *32*, 122–136.

Rogerson, R. (1988). Indivisible labor, lotteries and equilibrium. *Journal of Monetary Economics*, *21*, 3–16.

Rogerson, R., & Wallenius, J. (2007). Micro and macro elasticities in a life cycle model with taxes. *Journal of Economic Theory*, *144*, 2277–2292.

Shimer, R. (2010). *Labor markets and business cycles*. Princeton, NJ: Princeton University Press.

Stock, J., & Wise, D. (1990). Pension, the option value of work and retirement. *Econometrica*, *58*(5), 1151–1180.

Varian, H. (1992). *Microeconomic analysis* (3rd ed.). New York, NY: W. W. Norton.

Wooldridge, J. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). Cambridge, MA: MIT Press.

Wooldridge, J. (2013). *Introductory econometrics: A modern approach* (5th ed.). Boston: MA: Cengage/South-Western.

CHAPTER 2

Labor Demand

In this chapter we will:

- See how firms choose their factors of production
- Analyze substitution between capital and labor
- Analyze substitution between different types of labor
- Study the trade-off between workers and hours
- Review estimates of the elasticities of labor demand with respect to the cost of inputs
- Study the effects of the adjustment costs of labor

INTRODUCTION

The previous chapter was devoted to the supply side of the labor market. But the level of employment does not depend only on decisions made by workers. The desire to perform a certain amount of work at a given wage must also meet the employers' plans. Decisions that firms make about employment depend on key factors altering the labor demand that must be analyzed.

The theory of labor demand is part of a wider context, that of the demand for the factors of production; the basic assumption is that firms utilize the services of labor by combining them with other inputs, such as capital, in order to maximize the profits they derive from the sale of their products. Labor demand theory thus sets out to explain the demand for manpower, as well as the amount of time worked by each employee. An entrepreneur has an interest in hiring a worker whenever the income that worker generates is greater than his cost. The demand for labor must therefore depend not only on the cost of labor but also on the cost of the other factors and on elements that determine what the firm can earn, such as how efficiently its labor force performs and the price at which it can sell its goods. The cost of labor is composed of wages and the social security contributions (also known as payroll taxes) borne by the employer. The efficiency of labor depends on the technology available and the quantities of the other factors of production, such as capital or energy, used by firms. It also depends on the qualities of

each worker, which depend in turn on individual characteristics like motivation, dexterity, and alertness, and on objective factors such as educational level and professional experience. The price of the good produced depends on the quality of the product, the preferences of purchasers, and the characteristics of competitors.

To study labor demand, it is helpful to make a distinction between short-run and long-run decisions. We assume that in the short run the firm adjusts its quantity of labor; its stock of capital we take as given. In the long run, however, it is possible for firms to substitute capital for certain categories of employees. Most works in the field also distinguish the "static" theory of labor demand from the "dynamic" theory. The static theory sets aside the *adjustment costs* of labor, that is, the costs connected solely to *changes* in the volume of this factor. If such costs do not exist, there are really no dynamics, since nothing prevents labor demand from reaching its desired level immediately.

By leaving adjustment delays out of consideration, static theory throws the basic properties of labor demand—the laws, as they are sometimes called—into relief in a simplified manner. Static theory comes to precise qualitative conclusions about the directions in which the quantity of labor demanded varies as a function of the costs of all the factors, and at a deeper level, it also succeeds in characterizing the elements that determine the *extent* of the elasticities of labor demand. Knowing the orders of magnitude of these elasticities is essential when it comes to assessing the effects of economic policy because they make it possible to quantify the response of firms when a change of policy comes into effect. For example, knowledge of the elasticity of unskilled labor with respect to its cost allows us to set out in approximate figures the changes in the demand for this category of wage earners in the wake of a reduction in social security contributions or a rise in the minimum wage.

Dynamic labor demand theory puts flesh on the bones of this knowledge by adding the effects of adjustment costs. Among other things, it furnishes indications concerning the form and speed of labor adjustments (which have also been the object of numerous empirical studies). Taking adjustment costs into account proves especially valuable for random environments in which firms face shocks, sometimes negative and sometimes positive, because it throws light on hiring and separation strategies. The dynamic analysis of labor demand also makes it possible to take into account the turnover of manpower, because a change in the level of employment in a firm is often one facet of a reorganization that requires replacing certain employees with others who have skills better adapted to the firm's plans. Consequently, net variations in employment within a firm are for the most part much more limited than its numbers of hires and separations, which may rise and fall quite steeply, as we see in figure 2.1 for a country like the United States. Even when overall employment is shrinking on average over a trimester, even over a single month, firms are still hiring. Likewise, the number of separations continues to be substantial in firms experiencing growth. It is interesting to note that firms whose workforce is diminishing have a number of hires per quarter equal to around 10% of their total workforce. Conversely, firms where the total workforce is expanding separate from around 10% of their workers every quarter. This phenomenon is not specific to the United States, as shown in figure 2.2,[1] where we observe the same phenomenon,

[1] In these figures, in order to control for the characteristics of establishments each point is estimated by regressing hires or separations rates, respectively, on dummies for each level of employment growth and on establishment fixed effects. A similar analysis can be found for France in Abowd, Corbel, and Kramarz (1999, figure 1, p. 36).

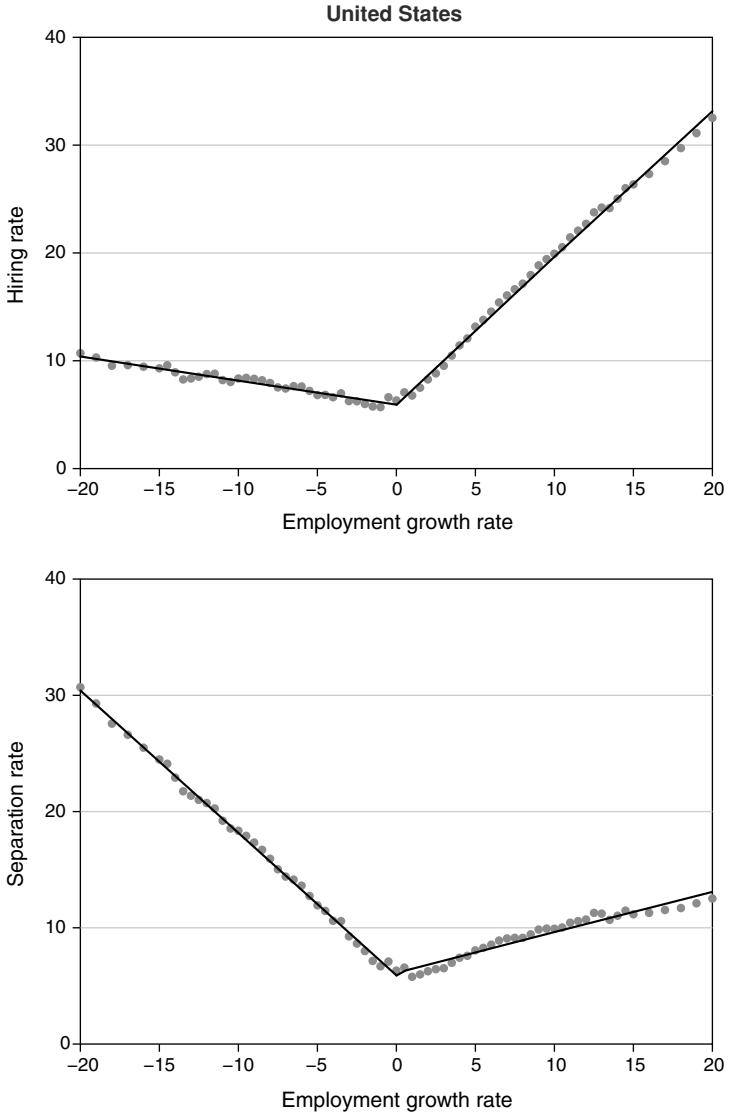


Figure 2.1
Hires, separations, and employment quarterly growth rates (in percentage) in the United States, based on 277,000 establishment quarterly observations from 2001 to 2010.

Source: Davis et al. (2012, figure 6, p. 10), based on JOLTS (Job Openings and Labor Turnover Survey, Bureau of Labor Statistics).

with the same order of magnitude, in French firms. Actually, labor turnover is important in all developed economies.

In this chapter, section 1 sets out the static theory of labor demand. The separation of substitution effects from scale effects supplies an operational grid within which to interpret the long-run determinants of this demand. Section 1 also looks at the case

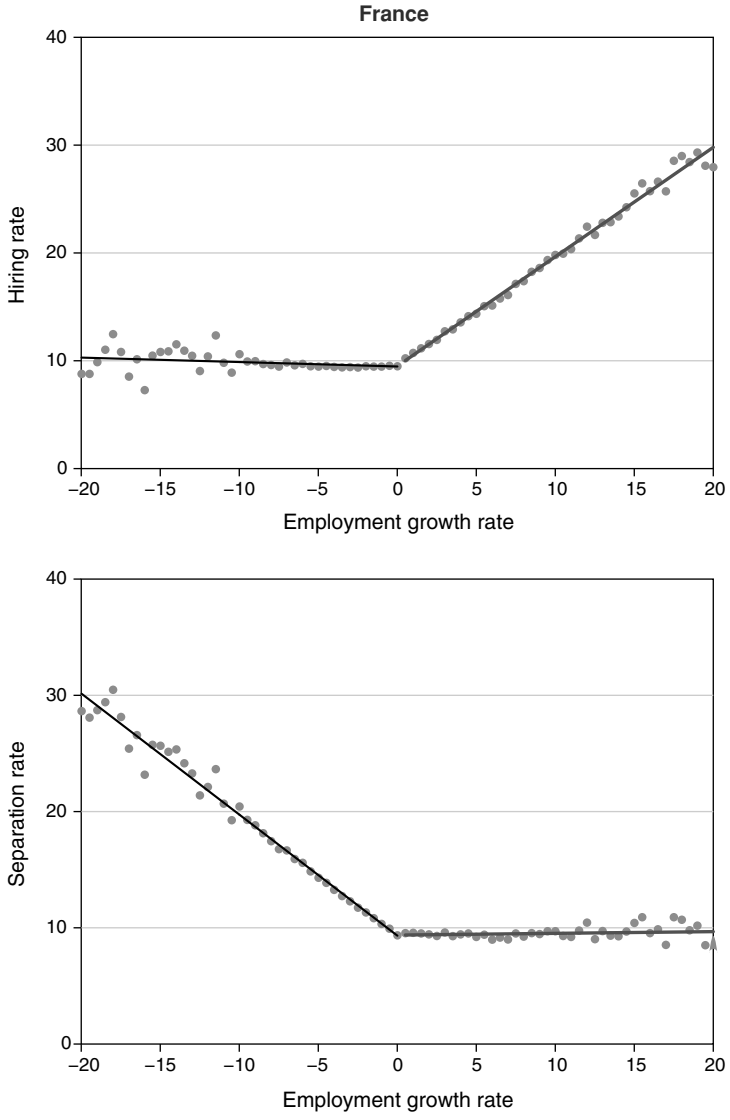


FIGURE 2.2
Hires, separations, and employment quarterly growth rates (in percentage) in France, based on 1,027,564 quarterly observations from 2007 to 2010.

Source: DMMO administrative data (Déclaration de Mouvements de Main-d'Oeuvre, Ministère du Travail).

of multiple (more than two) factors of production and analyzes the trade-off between manpower and hours in this context. Section 2 shows how, by specifying the production function or the cost function explicitly, we can more easily make the transition from theoretical models to estimates. It concludes with a review of the main empirical results. Section 3 integrates adjustment costs into labor demand theory in order to bring out the dynamics of employment more clearly. It shows that these dynamics, and

the properties of the stationary state, depend heavily on the functional form chosen to describe the costs linked to changes in employment. It also highlights the role of forecasts in adjustments of employment. Like section 2, it concludes with a summary of the main results arrived at by empirical studies.

1 THE STATIC THEORY OF LABOR DEMAND

In the short run, we can make the assumption that only the volume of labor services is variable. But in the long term, there exist possibilities of substituting capital for labor that substantially change the determinants of labor demand. When we do set the time horizon farther out, we can no longer study labor demand by focusing narrowly on just two aggregate factors—capital and labor—because the firm can also, for example, change the composition of its workforce by changing the structure of skills it uses. Hence we are led to study the behavior of firms when there are more than two factors of production. The heterogeneity of labor shows up as well in the imperfect substitutability between manpower and number of hours worked. Hence every firm has to make tradeoffs between the number of its employees and the length of time each employee works as a function of the costs incurred when each of these two dimensions of labor demand is utilized with greater or less intensity.

1.1 Labor Demand in the Short Run

In the short run, the volume of work within a firm is more easily adaptable than the stock of capital, so labor demand depends on the real wage and the market power of the firm.

1.1.1 Market Power

The demand $Y(P)$ for a particular good depends, among other things, on the price P at which a firm sells its product. To make the explanation easier, it is preferable to work with the inverse relationship $P = P(Y)$, called *the inverse demand function*. It is assumed to be decreasing and we will denote its elasticity by $\eta_Y^P \equiv YP'(Y)/P(Y)$. A further hypothesis will be made, though it is not necessary to establish most of the results in this chapter: we assume for simplicity's sake that function $P(Y)$ is isoelastic, meaning that the elasticity η_Y^P is a constant independent of Y.

When $\eta_Y^P = 0$, the price of the good does not depend on the quantity produced by the firm. This situation characterizes perfect competition and the firm is then described as a "price taker." On the contrary, if $\eta_Y^P < 0$, the firm finds itself in a situation of imperfect competition and we then say that it is a "price maker." In a general way, the absolute value $\left|\eta_Y^P\right|$ of this elasticity constitutes an indicator of the firm's market power, inasmuch as the larger $\left|\eta_Y^P\right|$ is, the greater the effects on the market price of a change in its level of production. We may also point out that the notation $P(Y)$ does not mean that the price P depends only on the quantity Y produced by the firm. For example, P may vary with decisions taken by competing firms. It is also influenced by the tastes and the incomes of consumers. At partial equilibrium, which is the situation assumed throughout this chapter, it is not useful to bring in all the parameters that have an influence on P explicitly, since only the decisions of a particular firm interest us.

1.1.2 Fixed and Flexible Factors

The factors of production comprise different types of manpower (for example, skilled and unskilled personnel) and different types of plant (machinery and factories). For simplicity, the latter will be represented by a single factor bearing the generic name *capital*. For reasons having to do principally with the time necessary to put them in place and their cost of installation or replacement, certain factors of production cannot be adjusted in the short run. Factors of this kind are called *fixed*, or *rigid*, factors, and we will assume that capital belongs to that category. Conversely, factors whose level can be altered in the short run are called *flexible*, or *variable*, factors. By definition, the levels of all the factors of production can be altered in the long run; hence, all factors of production are flexible in the long run. As regards manpower, certain categories of personnel have to be placed among the fixed factors (choices regarding highly skilled personnel have much in common with decisions about investment), while others (temporary workers, for example) are similar to flexible factors. At the most aggregate level possible, that is, when the ensemble of the services performed by the workforce is represented by a single variable, measured in hours, for example, it is natural to take the view that labor is more flexible than capital.

1.1.3 Cost of Labor and Marginal Productivity

We begin our study of labor demand by assuming that all the services performed by this factor can be represented by a single aggregate L, which is flexible in the short run, the other inputs being considered rigid at that horizon. Their levels can therefore be considered given, and we may, without risk of confusion, represent the production process by a function with a single variable, or $Y = F(L)$. We assume that this function is strictly increasing and strictly concave, that is, that marginal productivity is positive ($F' > 0$) and decreasing with the level of employment ($F'' < 0$).

If we designate the price of a unit of labor by W and set aside the costs tied to the utilization of fixed factors, the firm's profit is written this way:

$$\Pi(L) = P(Y)Y - WL \quad \text{with} \quad Y = F(L)$$

The entrepreneur's only decision is to choose her level of employment so as to maximize her profit. The first-order condition is obtained simply, by setting the derivative of the profit to zero with respect to L, so that:

$$\Pi'(L) = F'(L)[P(Y) + P'(Y)Y] - W = F'(L)P(Y)(1 + \eta_Y^P) - W = 0$$

When $(1 + \eta_Y^P) > 0$, the labor demand is defined by:[2]

$$F'(L) = \nu \frac{W}{P} \quad \text{with} \quad \nu \equiv \frac{1}{1 + \eta_Y^P} \tag{2.1}$$

This relation signifies that the profit of the firm attains its maximum when the marginal productivity of labor is equal to the real wage W/P multiplied by a *markup* $\nu \geq 1$.

[2]The second derivative of the profit is written $\Pi''(L) = (1 + \eta_Y^P)(F'^2 P' + F'' P)$. Since $P' < 0$ and $F'' < 0$, the second-order condition $\Pi''(L) < 0$ dictates that we have $(1 + \eta_Y^P) > 0$.

The latter is an increasing function of the absolute value $\left|\eta_Y^P\right|$ of price elasticity with respect to production. The markup constitutes a measure of the firm's market power. In a situation of perfect competition, the firm has no market power ($\eta_Y^P = 0$) and marginal productivity is equal to the real wage.

The concept of cost function allows us to interpret the optimality condition (2.1) differently. In this model, with just one factor of production, the cost function simply corresponds to the cost of labor linked to the production of quantity Y of a good, or $C(Y) = WL = WF^{-1}(Y)$, where F^{-1} designates the inverse function of F. Since the derivative of F^{-1} is equal to $1/F'$, the marginal cost is defined by $C'(Y) = W/F'(L)$, and relation (2.1) is written:

$$P = \nu \frac{W}{F'(L)} = \nu C'(L) \tag{2.2}$$

In other words, the firm sets its price by applying the markup ν to its marginal cost $C'(Y)$. In the situation of perfect competition ($\nu = 1$), the price of the good exactly equals the marginal cost.

The expression of labor demand allows us to study the impact of a variation in the cost of labor on the volume of labor. Differentiating relation (2.1) with respect to W, we find again that:

$$\frac{\partial L}{\partial W} = \nu/(F'^2 P' + PF'') < 0$$

Hence short-run labor demand and thus the level of supply of the good are *decreasing* functions of labor cost. On the other hand, the selling price of the good produced by the firm rises with W. It could be shown in the same manner that labor demand and the level of production diminish, while the price rises, when the markup ν grows larger.

Thus, in the short run, the cost of labor, the determinants of demand for the good produced by the firm, the firm's technology, and the structure of the market for goods—represented by the markup ν or the elasticity η_Y^P—all influence labor demand. In the longer run, the firm may contemplate replacing part of its workforce with machines, or conversely increasing the numbers of its personnel and reducing its stock of capital. Labor demand will then depend on the technical feasibility of these operations and the price of the other inputs.

1.2 The Substitution of Capital for Labor

We now shift to a long-run perspective, in which capital K also becomes a flexible factor. To appreciate better the different elements that bear on demands for the factors of production, it will be helpful to conduct the analysis in two stages. In the first stage, the level of production is taken as given, and we look for the optimal combinations of capital and labor through which that level can be reached. In the second stage, we look for the volume of output that will maximize the firm's profit. This approach makes it possible to distinguish *substitution effects*, which occur in the first stage, where the volume of production is fixed, from *scale effects*, which are confined to the second stage, in which the optimal level of production is set. More precisely, substitution effects relate to the choice of one factor over another in order to attain a given level of production.

Scale effects (also called quantity effects, or supply effects) have to do with the capacity to alter the level of production while retaining the same proportions among the various inputs. We begin by analyzing the first stage of the producer's problem; scale effects will be studied in section 1.3. The first stage makes it possible to define and characterize the firm's cost function. We can then deduce the properties of the *conditional factor demands*.

1.2.1 Minimization of Total Cost

Assuming a technology with just two inputs, capital and labor, the conditional demands for these inputs depend only on the relative price of each. The properties of these conditional demands can be deduced if we know the cost function of the firm.

A Technology with Two Inputs

Assuming once more that labor can be represented by a single aggregate L, the production function of the firm will now be written $F(K,L)$. If production of level Y requires that capital and labor always be combined in the same proportion—that is, that the ratio K/L remains a constant independent of Y—capital and labor are complementary inputs. In this case, it is enough to know the level of production in order to obtain the quantity of each factor utilized. Formally, we have reverted to the preceding analytical framework, where the production function had only one argument. But we assume from now on that to attain a given level of production, capital and labor can always combine in different proportions. Factors possessing this property are said to be *substitutable*.

More precisely, we posit that the production function is strictly increasing with each of its arguments, so that its partial derivatives will be strictly positive, or, with the obvious notations, $F_K > 0$ and $F_L > 0$. We also assume that this function is strictly concave, which signifies in particular that the marginal productivities of each factor diminish with the quantity of the corresponding factor. We will thus have $F_{KK} < 0$ and $F_{LL} < 0$. To make certain results clearer, it will sometimes be useful to assume that the production function is homogeneous. We may note that if $\theta > 0$ designates the degree of homogeneity, this property is characterized by the following equality:

$$F(\mu K, \mu L) = \mu^{\theta} F(K,L) \quad \forall \mu > 0, \quad \forall (K,L) \tag{2.3}$$

Parameter θ represents the level of *returns to scale*. The homogeneity of the production function implies that this parameter is independent of the level of production. We say that returns to scale are decreasing if $0 < \theta < 1$, constant if $\theta = 1$, and increasing if $\theta > 1$.

Cost Function and Factor Demand

The optimal combination of inputs is obtained by minimizing the cost linked to the production level Y. Let us designate by R and W respectively the price of a unit of capital and a unit of labor; the quantities of inputs corresponding to this choice are given by the solution of the following problem:

$$\min_{\{K,L\}} (WL + RK) \quad \text{subject to constraint } F(K,L) \geq Y \tag{2.4}$$

The solutions, denoted $\bar{L}$ and $\bar{K}$, are called, respectively, the *conditional demand for labor* and the *conditional demand for capital*. The minimal value of the total cost, or $(W\bar{L} + R\bar{K})$, is then a function of the unit cost of each factor and the level of production. This minimal value is called the cost function of the firm, and we will denote it $C(W, R, Y)$.

A figure will help us to understand the solution of problem (2.4). In figure 2.3, we show, in the plane (K, L), an *isoquant* labeled (Y). By definition, this curve designates the set of values of K and L allowing a given level of production to be attained, in other words satisfying $F(K, L) = Y$. In the plane (K, L), an isoquant is thus a curve of equation $K(L)$ such that $F[K(L), L] = Y$. Its slope is negative, and the absolute value of its derivative is, by definition, equal to the *technical rate of substitution* between capital and labor, or $|K'(L)| = F_L/F_K$. The technical rate of substitution defines the quantity of capital that can be saved when the quantity of labor is augmented by one unit. In appendix 7.1 to this chapter it is shown as well that the isoquants are strictly convex $(K'' > 0)$ when the production function is strictly concave. This means that the technical rate of substitution, equal to the absolute value of $K'(L)$, is decreasing: the larger the volume of labor, the less capital can be saved by augmenting the quantity of labor by one unit. In figure 2.3 we have also represented an *isocost* curve (C_0). This corresponds to the values of K and L such that $WL + RK = C_0$, where C_0 is a positive given constant. An isocost curve is thus a straight line with a slope $-(W/R)$ moving out towards the northeast when C_0 increases. It is evident, then, that if the isocost line is not tangent to the isoquant—at point E', for example—it is always possible to find a combination of factors K and L satisfying the constraint $F(K, L) \geq Y$ and leading to a cost inferior to that of the combination represented by point E'. For that, we need only cause line (C_0) to move in towards the origin (for example, at point E'' the total cost of production

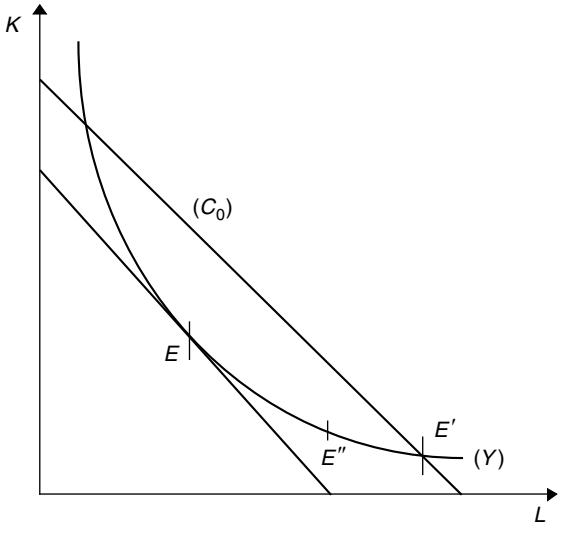


FIGURE 2.3
The minimization of total cost.

is inferior to its value at point E'). To sum up, the producer's optimum lies at point E where the isocost line is tangent to the isoquant. The reader will see that the property of strict convexity of the isoquant guarantees that point E represents a unique minimum for the cost of production. At this point, the technical rate of substitution is equal to the ratio of the costs of inputs. The conditional demands for capital and labor are thus defined by the following equations:

$$\frac{F_L(\bar{K},\bar{L})}{F_K(\bar{K},\bar{L})} = \frac{W}{R} \quad \text{and} \quad F(\bar{K},\bar{L}) = Y \tag{2.5}$$

The Properties of the Cost Function

Relation (2.5) shows that $\bar{K}$ and $\bar{L}$ depend only on the level of production Y and the relative price W/R of labor. Evidently we could deduce the properties of the conditional demands using the two equations of relation (2.5). In fact, though, it proves simpler to proceed indirectly by relying on the cost function $C(W,R,Y)$. Thus in appendix 7.2 of this chapter it is shown that the latter possesses the following properties:

(i) It is *increasing* with respect to each of its arguments and *homogeneous of degree 1* in (W,R).

(ii) It is *concave* in (W,R), which signifies in particular that the second derivatives C_{WW} and C_{RR} are negative.

(iii) It satisfies *Shephard's lemma*, or:

$$\bar{L} = C_W(W,R,Y) \quad \text{and} \quad \bar{K} = C_R(W,R,Y) \tag{2.6}$$

where C_W and C_R designate respectively the partial derivatives of the cost function with respect to W and R.

(iv) It is homogeneous of degree $1/\theta$ with respect to Y when the production function is homogeneous of degree θ. Under this hypothesis, the conditional demands for factors are also homogeneous of degree $1/\theta$ in relation to Y. Formally, we thus have:

$$C(W,R,Y) = C(W,R,1)Y^{1/\theta}, \quad \bar{L}\left(\frac{W}{R},Y\right) = \bar{L}\left(\frac{W}{R},1\right)Y^{1/\theta} \quad \text{and}$$
$$\bar{K}\left(\frac{W}{R},Y\right) = \bar{K}\left(\frac{W}{R},1\right)Y^{1/\theta} \tag{2.7}$$

These properties of the cost function allow us to derive the properties of the conditional factor demands very easily.

1.2.2 The Properties of the Conditional Factor Demands

The most important properties of the conditional demands for labor and capital have to do with the way they vary in the wake of a rise or a fall in the prices of these factors.

The extent of these variations depends on the elasticity of substitution between capital and labor on one hand and the share of each factor in the total cost on the other hand.

Variations in Factor Prices

The differentiation of the first relation of Shephard's lemma (2.6) with respect to W entails:

$$\frac{\partial \bar{L}}{\partial W} = C_{WW} \leq 0$$

The conditional labor demand is thus *decreasing* with the price of this factor. Since the first-order conditions (2.5) show that conditional demand in reality depends only on the relative price of labor, that is, on W/R, we can state that it increases with the price of capital. Symmetrically, we could show that the conditional demand for capital diminishes with R and increases with W.

Shephard's lemma allows us to characterize more precisely the cross effects of a change in the price of one factor on the demand for the other factor. Thus relation (2.6) immediately entails:

$$\frac{\partial \bar{L}}{\partial R} = \frac{\partial \bar{K}}{\partial W} = C_{WR} \tag{2.8}$$

Since it was shown above that the conditional demand for a factor is increasing with the price of the other factor, we can deduce that the cross derivative C_{WR} is necessarily positive.[3] The equality (2.8) portrays the symmetry condition of cross-price effects. It means that at the producer's optimum, the effect of a rise of one dollar in the price of labor on the volume of capital is equal to the effect of a rise of one dollar in the price of capital on the volume of labor. This (astonishing) equality is no longer verified in terms of elasticities.

Cross Elasticities and the Elasticity of Substitution Between Capital and Labor

Let us recall first that the *cross* elasticities $\bar{\eta}_R^L$ and $\bar{\eta}_W^K$ of the conditional demand for a factor with respect to the price of the other factor are defined by:

$$\bar{\eta}_R^L = \frac{R}{\bar{L}} \frac{\partial \bar{L}}{\partial R} \quad \text{and} \quad \bar{\eta}_W^K = \frac{W}{\bar{K}} \frac{\partial \bar{K}}{\partial W} \tag{2.9}$$

At the producer's optimum, relation (2.8) then entails $\bar{\eta}_R^L = (R\bar{K}/W\bar{L})\bar{\eta}_W^K$. Consequently, leaving aside the exceptional case where the cost $W\bar{L}$ of manpower would equal the cost $R\bar{K}$ of capital, the cross elasticities will always be different. They do not, therefore, constitute a significant indicator of the possibilities of substitution between these two factors. To get around this problem, it is preferable to resort to the notion of *elasticity of substitution*, which is the elasticity of the variable $\bar{K}/\bar{L}$ with respect to relative price W/R. The elasticity of substitution between capital and labor, denoted σ, is defined by:

$$\sigma = \frac{W/R}{\bar{K}/\bar{L}} \frac{\partial(\bar{K}/\bar{L})}{\partial(W/R)} \tag{2.10}$$

[3] See as well relation (2.76) in appendix 7.2 to this chapter.

This formula indicates that the capital–labor ratio increases by $\sigma\%$ when the ratio between the price of labor and the price of capital increases by 1%. Figure 2.3 shows that a rise (or a fall) of the relative price W/R increases (or diminishes) the slope of the straight lines of isocost and therefore shifts point E towards the left (or the right) along the isoquant. In other words, the ratio $\bar{K}/\bar{L}$ varies in the same direction as the relative price W/R. The elasticity of substitution between capital and labor is thus always *positive* (though it should be noted that this result is no longer automatically verified when the production function has more than two factors of production; see section 1.4.1).

It is possible to obtain a simple expression of this elasticity of substitution by exploiting the homogeneity of the cost function. In appendix 7.2, it is established that the elasticity of substitution is written in the following manner:[4]

$$\sigma = \frac{CC_{WR}}{C_W C_R} \tag{2.11}$$

The reader can verify that σ is *symmetric* in W and R; therefore this variable also represents the elasticity of the ratio $\bar{L}/\bar{K}$ in relation to the relative cost R/W. It should be noted further that it does not depend on the level Y of production when the hypothesis (2.3) of the homogeneity of the production function is satisfied. Consequently, property (iv) of the cost function set out in the preceding paragraph stipulates that the conditional demands $\bar{K}$ and $\bar{L}$ are homogeneous of degree $1/\theta$ in Y when the production function is homogeneous of degree θ. In this case, the ratio $\bar{K}/\bar{L}$ does not depend on Y and consequently the elasticity of substitution σ depends only on the relative price W/R.

Conditional Demands and the Factor Shares in the Total Cost

It is instructive to express the cross elasticities defined by (2.9) as a function of σ. With the help of relation (2.8), we note that $\bar{\eta}_R^L$ is equal to $(R/\bar{L})C_{WR}$. The expression (2.11) of the elasticity of substitution then leads to $\bar{\eta}_R^L = (RC_W C_R/\bar{L}C)\sigma$. Let us designate by $s \equiv W\bar{L}/C$ the labor share in the total cost. Since, following Shephard's lemma—see (2.6)—we have $\bar{L} = C_W$ and $\bar{K} = C_R$, we immediately arrive at $\bar{\eta}_R^L = (1 - s)\sigma$. Thus, the elasticity of the conditional labor demand with respect to the cost of capital is equal to the elasticity of substitution multiplied by the share of capital in the total cost. It could be shown in the same way that the elasticity of the conditional capital demand with respect to the cost of labor is equal to the elasticity of substitution multiplied by the share of labor in the total cost. There exists as well a link between the *direct* elasticity $\bar{\eta}_W^L$ and the elasticity of substitution σ. The conditional demand for labor depending only on Y and on the ratio W/R, we have $\partial\bar{L}/\partial W = -(R/W)(\partial\bar{L}/\partial R)$, and consequently:

$$\bar{\eta}_W^L = -\bar{\eta}_R^L = -(1 - s)\sigma \tag{2.12}$$

[4] It is possible to obtain an expression of the elasticity of substitution depending only on the partial derivatives of the production function using optimality condition (2.5). We then find:

$$\sigma = \frac{F_K F_L (KF_K + LF_L)}{KL(2F_{KL}F_K F_L - F_{KK}F_L^2 - F_{LL}F_K^2)}$$

When the production function is homogeneous of degree 1, the elasticity of substitution takes a particularly simple form:

$$\sigma = \frac{F_K F_L}{YF_{KL}}$$

Relation (2.12) proves particularly interesting from an empirical point of view for it supplies a simple link between estimates of the elasticity of substitution σ and those of $\bar{\eta}_W^L$ or $\bar{\eta}_R^L$ (see section 2.2.1). What is more, it offers very useful indications of the effect of a variation in the price of the factors on conditional labor demand. In the first place, it is apparent that the greater the possibilities of substitution between capital and labor, the larger this effect is in absolute value. When the value of the elasticity of substitution is high, that means that to obtain a given level of production, the entrepreneur has the possibility of diminishing "greatly" the utilization of one factor and "greatly" increasing that of the other, in the wake of a change in the relative price of the factors. Thus, when W rises or R falls, the firm's interest in diminishing the utilization of labor so as to minimize the total cost is all the greater, the higher the value of σ is. That explains why the elasticities of conditional labor demand are increasing, in absolute value, with the elasticity of substitution σ.

Symmetrically, the influence of the relative share of the cost of a factor can easily be grasped by assuming that σ remains constant. For a given value of the relative price W/R, the fact that the share $(1 - s)$ of capital is "small" reveals that the firm utilizes relatively little of this factor and a great deal of labor. Now, the larger the quantity of labor is, the smaller the variations in the quantity of labor expressed in percentage terms are. The logic goes the other way, of course, if the share of capital is large. Accordingly, the direct and cross elasticities of the conditional labor demand increase in absolute value with the share of capital in the total cost. In an equivalent fashion, these elasticities diminish in absolute value with the share of labor in the total cost.

Variation in the Level of Output

The effects of an exogenous change in the level of output Y on the total cost are easily characterized if total cost is defined by $C = W\bar{L} + R\bar{K}$ with $F(\bar{K}, \bar{L}) = Y$. It suffices to differentiate these last two equalities with respect to Y and to take account of the optimality condition (2.5) to get the following expression of the *marginal cost* (equal by definition to the partial derivative C_Y of the cost function with respect to the output level Y):

$$C_Y(W, R, Y) = \frac{W}{F_L} = \frac{R}{F_K} \tag{2.13}$$

First, it is apparent that the marginal cost is always positive. That signifies that the total cost rises with the level of output. Conversely, it is not possible to know the direction of variations in factor demands without supplementary hypotheses. Clearly, factor demands do not diminish simultaneously when production increases. Thus a rise in production simply requires that the volume of one of the factors increase, but the volume of the other factor is not obliged to do so; it can even decrease. However, when the production function satisfies the homogeneity hypothesis (2.3), a more precise conclusion emerges. The factor demands are then homogeneous of degree $1/\theta$ with respect to Y—see property (iv) of the cost function set out in section 1.2.1—and relation (2.7) clearly shows that the conditional demands for labor and capital then rise simultaneously with the level of output.

Minimization of cost for a given level of output constitutes the first stage of the problem of the firm; we must now examine how the optimal volume of output is determined.

1.3 SCALE EFFECTS

The entrepreneur is generally in a position to choose her level of production. The desired quantities of the factors are then distinguishable from their conditional demands. The analysis of substitution and scale effects yields highly general properties for labor demand; among other things, it brings into play the elasticity of substitution between capital and labor, the share of each factor in the total cost, and the market power of the firm.

1.3.1 UNCONDITIONAL FACTOR DEMANDS

The entrepreneur chooses a level of output that maximizes her profit. Let us again designate by $P(Y)$ the inverse demand function. Then, profit $\Pi(W,R,Y)$ linked to a level of production Y when the unit costs of labor and capital are respectively W and R, takes the following form:

$$\Pi(W,R,Y) = P(Y)Y - C(W,R,Y) \tag{2.14}$$

The first-order condition is obtained by setting the derivative of this expression to zero with respect to Y. Rearranging terms, we find that the optimal level of production is characterized by the equality:

$$P(Y) = \nu C_Y(W,R,Y) \quad \text{with} \quad \nu \equiv 1/(1+\eta_Y^P) \tag{2.15}$$

In the case of a production function homogeneous of degree θ, we can verify[5] that it is indeed a maximum if and only if $\nu > \theta$. We rediscover the result we obtained when we studied short-run labor demand (see equation (2.2)): the firm sets its price by applying the markup ν to its marginal cost C_Y. Taking into account expression (2.13) of marginal cost, the optimality condition (2.15) takes the following form:

$$F_L(K,L) = \nu\frac{W}{P} \quad \text{and} \quad F_K(K,L) = \nu\frac{R}{P} \tag{2.16}$$

In other words, at the firm's optimum the marginal productivity of each factor is equal to its real cost multiplied by the markup. When the competition in the market for the good produced by the firm is perfect ($\nu = 1$), we rediscover the usual equalities between the marginal productivity of a factor and its real cost. The values of K and of L, defined by equations (2.15) and (2.16), are called the *long-run*, or *unconditional*, demands for capital and for labor.

[5] Deriving profit (2.14) with respect to Y gives:

$$\Pi_Y(W,R,Y) = P(Y)(1+\eta_Y^P) - C_Y(W,R,Y)$$

The first equation of (2.7) implies that the marginal cost C_Y is linked to the average cost C/Y by the identity $C_Y \equiv (C/Y)/\theta$. To find the value of the second derivative of the profit at a point satisfying the first-order condition (2.15), we replace C_Y by $C/\theta Y$ in the expression of Π_Y and we differentiate with respect to Y. Taking into account (2.15), the result, after several calculations, is:

$$\Pi_{YY}(W,R,Y) = (1+\eta_Y^P)\frac{P(Y)}{\theta Y}(\theta\eta_Y^P - 1 + \theta) = \frac{P(Y)}{\theta Y}\frac{\theta-\nu}{\nu^2}$$

The second-order condition is thus satisfied, since $\nu > \theta$.

1.3.2 The Laws of Demand

The laws of demand refer to the manner in which *unconditional* demands for the factors of production vary with the unit costs of these factors. They combine substitution and scale effects.

The Decreasing Relation Between the Demand for a Factor and Its Cost

We first demonstrate that the unconditional demand for a factor is decreasing with the cost of this factor. This property possesses a very general character: in particular, it does not depend on the production function of the firm being homogeneous. To establish this result, let us first consider the *profit function*, denoted $\Pi(W,R)$, equal to the maximal value of profit for given values of the costs of the inputs. It is defined by:

$$\Pi(W,R) \equiv \max_{Y} \Pi(W,R,Y)$$

The cost function $C(W,R,Y)$ being concave in (W,R) for all Y, relation (2.14) signifies that function $\Pi(W,R,Y)$ is convex in (W,R), whatever the value of Y may be. Let us denote by Y^* the optimal level of production given by (2.15); by definition, we have $\Pi(W,R) = \Pi(W,R,Y^*)$. It can be shown that the profit function $\Pi(W,R)$ is equally convex in (W,R).[6]

Differentiating relation (2.14) with respect to W we have:

$$\Pi_W(W,R) = \left[P(Y^*)(1+\eta_Y^P) - C_Y(W,R,Y^*)\right]\frac{\partial Y^*}{\partial W} - C_W(W,R,Y^*)$$

According to optimality condition (2.15), the term in brackets is null. Moreover, Shephard's lemma (2.6) states that the partial derivative $C_W(W,R,Y^*)$ is equal to unconditional labor demand L^*. An analogous rationale evidently applies to the unconditional capital demand K^*. We thus arrive at the following relations, known as *Hotelling's lemma*:

$$\Pi_W(W,R) = -L^* \quad \text{and} \quad \Pi_R(W,R) = -K^* \tag{2.17}$$

The profit function $\Pi(W,R)$ being convex, we then have $\Pi_{WW} \geq 0$ and $\Pi_{RR} \geq 0$, and relation (2.17) immediately entails:

$$\frac{\partial L^*}{\partial W} = -\Pi_{WW} \leq 0 \quad \text{and} \quad \frac{\partial K^*}{\partial R} = -\Pi_{RR} \leq 0 \tag{2.18}$$

[6] For every quintuplet (W_1, W_2, R_1, R_2, Y) the convexity in (W,R) of function $\Pi(W,R,Y)$ entails:

$$\Pi[\lambda W_1 + (1-\lambda)W_2, \lambda R_1 + (1-\lambda)R_2, Y] \leq \lambda\Pi(W_1,R_1,Y) + (1-\lambda)\Pi(W_2,R_2,Y)$$

Taking the maximum in Y on the right- and left-hand sides of this inequality, we get:

$$\Pi[\lambda W_1 + (1-\lambda)W_2, \lambda R_1 + (1-\lambda)R_2] \leq \max_{Y}[\lambda\Pi(W_1,R_1,Y) + (1-\lambda)\Pi(W_2,R_2,Y)]$$

But we also have:

$$\max_{Y}[\lambda\Pi(W_1,R_1,Y) + (1-\lambda)\Pi(W_2,R_2,Y)] \leq \max_{Y}\lambda\Pi(W_1,R_1,Y) + \max_{Y}(1-\lambda)\Pi(W_2,R_2,Y)$$

Since the right-hand member of this last inequality is by definition equal to $\lambda\Pi(W_1,R_1) + (1-\lambda)\Pi(W_2,R_2)$, it results that:

$$\Pi[\lambda W_1 + (1-\lambda)W_2, \lambda R_1 + (1-\lambda)R_2] \leq \lambda\Pi(W_1,R_1) + (1-\lambda)\Pi(W_2,R_2)$$

The profit function $\Pi(W,R)$ is thus convex in (W,R).

Thus, under very general conditions, unconditional demand for a factor is a decreasing function of the cost of this factor. It must also be noted that the direction in which this demand varies with the cost of the other factor is not determined a priori—a consequence of the fact that the scale effect may now be opposed to the substitution effect. More generally, it is important to know the determinants of the relative extent of these two effects.

Labor Demand Elasticities

It is possible to be more exact about unconditional labor demand L^* by noting that it always satisfies Shephard's lemma (2.6). Thus we have $L^* = C_W(W, R, Y^*)$. Differentiating this equality with respect to W, we get:

$$\frac{\partial L^*}{\partial W} = C_{WW} + C_{WY}\frac{\partial Y^*}{\partial W}$$

When we multiply the two members of this relation by W/L^*, we bring to light the elasticities η_W^L and η_W^Y of unconditional labor demand and of the level of output with respect to the wage. The result is:

$$\eta_W^L = \frac{W}{L^*}C_{WW} + \left(\frac{Y^* C_{WY}}{L^*}\right)\eta_W^Y$$

Since $L^* = C_W(W, R, Y^*)$, the terms $(W/L^*)C_{WW}$ and $(Y^*/L^*)C_{WY}$ designate respectively the elasticty $\bar{\eta}_W^L$ of the *conditional* labor demand and the elasticity of this demand with respect to the level of output taken at point $Y = Y^*$. This last elasticity can be denoted $\bar{\eta}_Y^L$. We thus finally obtain:

$$\eta_W^L = \bar{\eta}_W^L + \bar{\eta}_Y^L \eta_W^Y \qquad \textbf{(2.19)}$$

This relation clearly reveals the different effects of a rise in wage on the demand for labor. We may start by isolating a *substitution effect* represented by the elasticity $\bar{\eta}_W^L$ of conditional labor demand. We saw in section 1.2.2 that this term is always negative, since for a given level of production, a rise in the cost of labor always leads to reduced utilization of this factor (and increased utilization of capital). Relation (2.19) likewise brings out a *scale effect*, represented by the product $\bar{\eta}_Y^L \eta_W^Y$. The direction of this scale effect is obtained by first noting that the second-order conditions of profit maximization for the firm dictate that η_W^Y should be of the opposite sign to C_{WY}.[7] Since, following Shephard's lemma (2.6), $\bar{\eta}_Y^L$ is of the same sign as C_{WY}, it results that the scale effect is always *negative* and therefore accentuates the substitution effect.

It should be emphasized that formula (2.19) measures the wage elasticity of employment of a given firm, the wages of other firms remaining constant. If the wage rises simultaneously in several competing firms producing substitutable goods, we should expect that the elasticity of employment will be weaker than that defined by relation (2.19) because the prices of competitors must also rise, for the same reason that those of the firm we are considering do. The demand for the goods of this firm

[7] With the help of expression (2.14) of the firm's profit, we can verify that the second-order condition implies $P'(Y) - \nu C_{YY} < 0$. Differentiating equation (2.15) with respect to W, we find that $\partial Y/\partial W$ is of contrary sign to C_{WY}.

thus diminishes less than in the case where competitors' wages remain constant. Consequently, the scale effect is weaker. Formally, if $\eta_{\overline{W}}^{Y}$ and $\eta_{\overline{W}}^{L}$ denote respectively the sum of the elasticities of production and employment of the firm we are considering with respect to its own wage, and with respect to all its competitors' wages (and bearing in mind that conditional elasticity $\bar{\eta}_{W}^{L}$ depends only on the wage of the firm we are considering), we get:

$$\eta_{\overline{W}}^{L} = \bar{\eta}_{W}^{L} + \bar{\eta}_{Y}^{L}\eta_{\overline{W}}^{Y} \tag{2.20}$$

Since $\left|\eta_{W}^{Y}\right| > \left|\eta_{\overline{W}}^{Y}\right|$ when firms produce substitutable goods, $\eta_{\overline{W}}^{L}$ is inferior in absolute value to η_{W}^{L} in the most probable case, where $\bar{\eta}_{Y}^{L}$ is positive. It is important to keep this result in mind when we come to interpret empirical studies, inasmuch as the latter frequently evaluate the impact of variations in the cost of labor that affect several firms, or even several sectors, simultaneously.

Gross Substitutes and Gross Complements

Using the same procedure, it is possible to calculate the cross elasticity η_{R}^{L} of the unconditional labor demand with respect to the cost of capital. This comes to:

$$\eta_{R}^{L} = \bar{\eta}_{R}^{L} + \bar{\eta}_{Y}^{L}\eta_{R}^{Y}$$

In the case of two inputs, we showed in section 1.2.2 that the conditional demand for a factor rises when the price of the other factor rises. The substitution effect, marked by the term $\bar{\eta}_{R}^{L}$, is thus positive. Conversely, the scale effect, represented by the term $\bar{\eta}_{Y}^{L}\eta_{R,}^{Y}$, is a priori ambiguous, except in the case of a homogeneous production function, where it is necessarily positive.[8] The sign of cross elasticity η_{R}^{L} is thus undetermined.

By definition, if $\eta_{R}^{L} > 0$, labor and capital are qualified as *gross substitutes*. When labor and capital are gross substitutes, a rise in the price of capital causes demand for this factor to fall and that of labor to rise: the substitution effect dominates the scale effect. If $\eta_{R}^{L} < 0$, labor and capital are qualified as *gross complements*; a hike in the price of one of these factors signifies that demand for both of them falls off, with the scale effect now dominating the substitution effect.

The Laws of Demand with a Homogeneous Production Function

When the production function is homogeneous, it is possible to express scale effects as a function of the labor share s in the total cost of the markup ν, and of the degree of homogeneity θ. To achieve this, we must first note that relation (2.7) immediately implies that the output elasticity of conditional labor demand $\bar{\eta}_{Y}^{L}$ is equal to $1/\theta$. Then, replacing C_Y by $C/\theta Y$ in the optimality condition (2.15) and taking the logarithmic derivatives with respect to W of this relation, we arrive at:

$$\frac{1}{Y}\left[\frac{YP'(Y)}{P(Y)} - \frac{YC_Y}{C}\right]\frac{\partial Y}{\partial W} = \frac{C_W}{C}$$

[8] A line of reasoning analogous to the one that allowed us to establish the direction of the scale effects in relation (2.19) would show that $\bar{\eta}_{Y}^{L}\eta_{R}^{Y}$ has a sign opposed to that of $C_{WY}C_{RY}$. Now, following Shephard's lemma (2.6), the latter quantity is equal to the product $(\partial\bar{L}/\partial Y)/(\partial\bar{K}/\partial Y)$. We have seen in section 1.2.2 that the conditional demands for factors rise with the level of output when the production function is homogeneous. In all other cases, the sign is ambiguous.

Since $L = C_W$, and following (2.7), $YC_Y/C = 1/\theta$, we find after several calculations:

$$\eta_W^Y = \frac{\theta s}{\theta(\eta_Y^P + 1) - 1} = \frac{\theta \nu}{\theta - \nu} s \tag{2.21}$$

The second-order conditions imposing $\nu > \theta$, we do indeed verify that $\eta_W^Y < 0$. Symmetrically, the value of η_R^Y is obtained by replacing s by $(1 - s)$ in relation (2.21). The scale effect of a rise in price of a factor is proportional to the share of the remuneration of this factor in the total cost. Taking account of relations (2.12) that give the values of the *conditional* demand elasticities, it becomes possible to express, with the help of (2.21), the direct and cross elasticities of *unconditional* labor demand as a function of the share s of this factor in the total cost, as a function of the elasticity of substitution σ between capital and labor, as a function of the margin rate ν, and as a function of the scale θ of overall returns. This is expressed as:

$$\eta_W^L = -(1 - s)\sigma - \frac{\nu}{\nu - \theta} s \quad \text{and} \quad \eta_R^L = (1 - s)\left(\sigma - \frac{\nu}{\nu - \theta}\right) \tag{2.22}$$

Knowledge of the order of magnitude of these elasticities becomes very important when the impact of economic policies must be assessed. That is why we need to understand clearly how they evolve when certain parameters change. Relations (2.22) yield relatively precise predictions concerning labor demand, which in large measure confirm the laws of demand put forward by Marshall (1920) and Hicks (1932) in their time. They are best understood by combining the substitution effect, the absolute value of which is measured by the term $(1 - s)\sigma$, with the scale effect measured by the other terms of these relations.

Market Power

The elasticity of the inverse demand function, η_Y^P, and so market power ν, do not play a role in the substitution effect. Conversely, it is evident that the scale effect *diminishes*, in absolute terms, when ν rises. Faced with a rise in the cost of labor, a firm with weak market power (ν approaching unity) cannot change its selling price very much—it cannot change it at all when competition in the market is perfect ($\nu = 1$)—and so the repercussion of the cost increase will essentially be felt in the output. If, on the other hand, the firm is highly monopolistic, or in other words if the elasticity of the inverse demand function, η_Y^P, is high, the firm can alter its price to a considerable degree without losing too much market share, that is, without changing its output level very much. In sum, the elasticity of output and so that of labor demand with respect to factor costs will diminish in absolute value, the higher the degree of monopoly.

Substitution of Capital for Labor

We see that the elasticity of substitution σ appears only in the substitution effect, with no influence on the scale effect, and since we have already looked at the consequences of a rise in σ for a given level of production, readers may refer back to the comments following relation (2.12). The general conclusion to which we came was that the easier it is to substitute capital for labor, the greater the direct and cross elasticities of labor demand are in absolute value.

The Share of Labor in the Cost of Production

In section 1.2.2 we studied the reasons why the substitution effect, equal in absolute value to $(1 - s)\sigma$, decreases as the share s of labor in the total cost decreases. Formulas (2.22) make it evident that the scale effect is indeed negative, but also that it increases (or diminishes) in absolute value with s if the rise in the cost of production is caused by an increase in W (or R). These movements are to be explained in the following manner: if s is large, then the firm utilizes "a lot" of labor and "little" capital, and in consequence production and employment will be very sensitive to a variation in labor cost but much less influenced by a change in the cost of capital. Hence the share s of labor in the total cost acts in opposite ways on the substitution effect and the scale effect. It is therefore the relative importance of an effect with respect to the other that will determine variations in the elasticities of labor demand. To be more precise, formulas (2.22) show that if capital and labor are gross substitutes—$\sigma > \nu/(\nu - \theta)$—then $\left|\eta_W^L\right|$ and η_R^L are decreasing functions of s. Under this hypothesis, the substitution effect dominates the scale effect and so it is normal that the behavior of unconditional demand should follow that of conditional demand. This result will obviously be inverted when the two factors of production are gross complements.

Adopting a production function limited to two factors thus allows us to assess the determinants of the level of capital and that of aggregate employment. But in many circumstances—for example, if we want to know the impact of an economic policy measure on the employment of unskilled persons—the labor factor can no longer be viewed as a single aggregate and it becomes necessary to work with a production function comprising more than two inputs.

1.4 Beyond Two Inputs

Here again it will be best to proceed in two stages. In the first, we seek to identify the optimal combinations of factors that enable a given level of production to be reached, and in the second, we determine the value of this level that maximizes the firm's profit. The first stage yields conditional demands, which are no longer necessarily characterized by a negative substitution effect. The second allows us to obtain unconditional demands.

1.4.1 Conditional Demands

Conditional factor demands result from the minimization of the total cost for a given level of production. But unlike the case in which there were only two inputs, the cross elasticities and thus the elasticities of substitution are no longer necessarily positive.

The Minimization of Total Cost

The production function of the firm is now written $Y = F(X^1, ..., X^n)$, where X^i is the quantity of factor i utilized in the production of a quantity Y of output. This function is assumed to be strictly increasing with each of its arguments and also strictly concave. If we designate by $W^i > 0$ the price of factor i, the *conditional* demands are obtained by minimizing the total cost linked to the production of a given quantity Y of output. They are thus solutions to the following problem:

$$\min_{(X^1,...,X^n)} \sum_{i=1}^{n} W^i X^i \quad \text{subject to} \quad F(X^1, ..., X^n) \geq Y$$

When there are more than two inputs, this problem cannot be solved graphically and it is therefore necessary to turn to conventional methods of optimization. Let $\lambda \leq 0$ be the multiplier linked to the production constraint; the Lagrangian of this problem is written:

$$\mathcal{L} = \sum_{i=1}^{n} W^i X^i + \lambda \left[F(X^1, ..., X^n) - Y \right]$$

Let F_i designate the partial derivative of function F with respect to its i^{th} argument; differentiating this Lagrangian with respect to X^i gives the first-order conditions. We thus have $(\partial \mathcal{L}/\partial X^i) = W^i + \lambda F_i = 0$, for all $i = 1, ..., n$. Since W^i and F_i are strictly positive, the multiplier λ is strictly negative and the production constraint is always binding. In sum, the conditional factor demands, denoted $\bar{X}^i$ for $i = 1, ..., n$, are defined by the following equations:

$$F(\bar{X}^1, ..., \bar{X}^n) = Y \quad \text{and} \quad \frac{F_i(\bar{X}^1, ..., \bar{X}^n)}{F_j(\bar{X}^1, ..., \bar{X}^n)} = \frac{W^i}{W^j} \quad \forall i, j = 1, ..., n \tag{2.23}$$

The strict concavity of function F guarantees that the necessary conditions for the minimization of total cost are also sufficient conditions. We note that the result described by relation (2.23) generalizes that obtained with two factors of production, that is, the technical rate of substitution (F_i/F_j) between factors i and j is equal to the relative cost (W^i/W^j) of factor i with respect to factor j.

The Cost Function

The minimum value of the total cost, or $\sum W^i \bar{X}^i$, is also called the *cost function* of the firm. It depends on the price of inputs and the output level Y, so it can be denoted $C(W^1, ..., W^n, Y)$. As in the case with two inputs, this function proves very useful for the study of the factor demands. In appendix 7.2 of this chapter, we show that it satisfies the following properties:

(i) It is *increasing* with each of its arguments and it is *homogeneous of degree 1* with respect to $(W^1, ..., W^n)$.

(ii) It is *concave* in $(W^1, ..., W^n)$, which signifies in particular that the partial second derivative C_{ii} is *negative* for all $i = 1, ..., n$.

(iii) It satisfies *Shephard's lemma*:

$$\bar{X}^i = C_i(W^1, ..., W^n, Y) \tag{2.24}$$

where C_i designates the partial derivative of function C with respect to its i^{th} argument.

(iv) It is homogeneous of degree $1/\theta$ in Y when the production function is homogeneous of degree θ. Under this hypothesis, the conditional factor demands are also homogeneous of degree $1/\theta$ in Y.

P-Substitute and P-Complement

Shephard's lemma allows us to obtain a very important property of the demand function of a production factor. Differentiating (2.24) with respect to W^i, we get:

$$\frac{\partial \bar{X}^i}{\partial W^i} = C_{ii} \leq 0 \quad \forall i = 1, \ldots, n \tag{2.25}$$

In other words, the conditional demand for input is always *decreasing* with the price of this input. This is a property of a very general kind, and so does not depend on the number of inputs. However, contrary to the results obtained with a production function having only two arguments, the variation in the conditional demand for a factor resulting from an increase in the cost of another factor does not always have a positive sign. In consequence, when W^i rises, the entrepreneur reduces his demand for factor i—this is the meaning of relation (2.25)—and he must perforce increase that of at least one other factor so as to achieve output level Y. But in the absence of further details about the firm's technology, it is not possible to know either which factor or factors will be utilized more or which ones will be utilized at the same or a lower level. Nonetheless, the symmetry condition of cross-price effects remains satisfied with any number n of inputs since relation (2.24) entails:

$$\frac{\partial \bar{X}^i}{\partial W^j} = \frac{\partial \bar{X}^j}{\partial W^i} = C_{ij} \quad \forall i, j = 1, \ldots, n \tag{2.26}$$

The symmetry condition of cross-price effects is a very general result. It indicates that the effect of a variation in the price of factor j on conditional demand for factor i is the same as that of a variation in the price of factor i on the conditional demand for factor j. As the direction of this effect turns out to be undetermined a priori, however, it will be convenient to make use of the following definitions: when $\partial \bar{X}^i / \partial W^j > 0$—or, in equivalent fashion, $\partial \bar{X}^j / \partial W^i > 0$—goods i and j are called *substitutes in the Hicks-Allen sense* or *p-substitutes* for short. In the opposite case, goods i and j are called *complements in the Hicks-Allen sense*, or simply *p-complements*. To put it another way, factors i and j are p-substitutes (or p-complements) if, to attain a given level of production, the demand for one of the factors increases (or diminishes) when the price of the other factor rises. It should be noted that if there are only two inputs, both are necessarily p-substitutes (see section 1.3.2).

Elasticity of Substitution

Taking into account relation (2.26), the *cross* elasticity of the conditional demand for factor i with respect to the price of factor j, or $\bar{\eta}^i_j$, takes the following form:

$$\bar{\eta}^i_j = \frac{W^j}{\bar{X}^i} \frac{\partial \bar{X}^i}{\partial W^j} = \frac{W^j}{\bar{X}^i} C_{ij} \tag{2.27}$$

As in the case with two inputs, it is apparent that cross elasticity is not a symmetrical notion—as a general rule $\bar{\eta}^i_j \neq \bar{\eta}^j_i$—and that is why we resort to the notion of *elasticity of substitution* when it comes to assessing the extent to which utilization of one factor may replace utilization of another. But here a difficulty arises, having to do

with the number of factors. If we define the elasticity of substitution by a formula analogous to the one employed in the case of two inputs—see (2.10)—we would then, for a given level of production, have to assume that the prices of the other factors do not vary and posit that the elasticity d_j^i of substitution between factors i and j represents the elasticity of ratio $\bar{X}^i/\bar{X}^j$ with respect to the relative cost W^j/W, or:

$$d_j^i = \frac{W^j/W^i}{\bar{X}^i/\bar{X}^j} \frac{\partial(\bar{X}^i/\bar{X}^j)}{\partial(W^j/W^i)}$$

The problem with this definition is that a variation in relative price W^j/W^i will not simply alter the ratio $\bar{X}^i/\bar{X}^j$ of the demands for inputs i and j but can set off a domino effect of substitutions in all the other inputs. In this case, the interpretation of d_j^i in terms of substitution between factors i and j alone becomes obscure, to say the least. A simple alternative, the one most frequently adopted, is to bring in the notion of *partial* elasticity of substitution in Allen's sense (as opposed to *direct* elasticity of substitution d_j^i). It is obtained by weighting the cross elasticity $\bar{\eta}_j^i$ by the inverse of the share of factor j in the total cost. By definition, we will thus have $\sigma_j^i = \bar{\eta}_j^i(C/W^j\bar{X}^j)$. With the help of relation (2.27) characterizing $\bar{\eta}_j^i$ and Shephard's lemma (2.24), we find that the partial elasticity of substitution is expressed by a formula analogous to equation (2.10) obtained with two inputs:

$$\sigma_j^i = \frac{CC_{ij}}{C_iC_j} \tag{2.28}$$

The elasticity of substitution thus defined is quite symmetrical, since $\sigma_j^i = \sigma_i^j$, but is not necessarily positive when there are more than two inputs.

Conditional Demands and Factor Shares

Let $s^j \equiv W^j\bar{X}^j/C$ be the share of factor j in the total cost; since according to Shephard's lemma (2.24), $\bar{X}^i = C_i$ and $\bar{X}^j = C_j$, relations (2.27) and (2.28) lead us to:

$$\bar{\eta}_j^i = s^j\sigma_j^i \quad \forall(i,j) \tag{2.29}$$

This relation is analogous to equality (2.12) from section 1.2.2. It is formally true for every couple (i,j), even when $i = j$, and is illuminating when it comes to interpreting the effect of variation in the price of a factor on conditional demand for the other factor. When the possibilities of substitution between two factors i and j are substantial—that is, when σ_j^i is a fairly large positive number—it is possible to attain an identical level of production by reducing the utilization of one of the factors "a lot." Thus, when W^j rises (or W^i falls) the firm has all the more incentive to replace factor j by factor i, the greater σ_j^i is. As in the case of a production function with two factors, this logic allows us to understand why the elasticity of conditional demand for factor i with respect to the price of factor j rises with the elasticity of substitution σ_j^i when these two factors are p-substitutes. But here factors i and j can also be p-complements ($\sigma_j^i < 0$). Let us suppose that this is in fact the case and that σ_j^i is a relatively large number in absolute value. Faced with a hike in the price of factor j, the producer reduces jointly the quantity of factor j and factor i for reasons having to do with the firm's technology.

The influence of the share s^j of factor j in the total cost is analyzed in the same way as in the case of a production function with two inputs: the elasticity of conditional demand for factor i with respect to the price of factor j rises in absolute value with the share s^j of factor j in the total cost.

1.4.2 UNCONDITIONAL DEMANDS

When overall cost has been minimized, the next stage is to maximize profit. Profit maximization allows us to characterize the unconditional factor demands. As in the case of two inputs, we are able to specify the sign of the cross elasticities by using the concepts of gross complementarity and gross substitutability.

Profit Maximization

Formally, the problem of the firm is analogous to the one dealt with in section 1.3.1, with a production technology comprising just two inputs, on condition that we replace the cost function $C(W, R, Y)$ by function $C(W^1, ..., W^n, Y)$. In particular, equation (2.15) giving the optimal level of output is now written:

$$P(Y) = \nu C_Y(W^1, ..., W^n, Y) \tag{2.30}$$

Consequently, the rule that the firm sets its price by applying the markup ν to the marginal cost C_Y continues to hold with any number of inputs. Moreover, calculations identical to those laid out in section 1.3.1 would show that if the production function is homogeneous of degree θ, the second-order condition always requires that we have $\nu > \theta$.

The procedure adopted to define the profit function in the case of two inputs also applies here. This function, denoted $\Pi(W^1, ..., W^n)$, corresponds to the maximal value of the firm's profit for given factor costs $(W^1, ..., W^n)$. The logic developed in section 1.3.1 will show that the profit function is *convex* and that it always satisfies *Hotelling's lemma*. Using Π_i to designate the partial derivative of the profit function with respect to W^i, and X_i to designate the unconditional demand for factor i, this lemma now takes the following form:

$$X_i = -\Pi_i(W^1, ..., W^n) \quad \forall i = 1, ..., n$$

The profit function being convex, we then have $\Pi_{ii} \geq 0$, and Hotelling's lemma immediately leads to:

$$\frac{\partial X_i}{\partial W^i} = -\Pi_{ii} \leq 0 \quad \forall i = 1, ..., n$$

The property that the (unconditional) demand for a factor diminishes with the price of this factor thus has a very general character, since it is satisfied whatever the number of inputs is. In fact, it is sometimes referred to as the law of demand.

Gross Substitutes and Gross Complements

The respective importance of substitution and scale effects emerges naturally when we note that unconditional factor demands satisfy Shephard's lemma (2.24) *and* that the optimal level of production satisfies relation (2.30). If we simply use X^i and Y to denote

the optimal values of demand for factor i and production, and differentiate (2.24) with respect to W^j, we get:

$$\frac{\partial X^i}{\partial W^j} = C_{ij} + C_{iY}\frac{\partial Y}{\partial W^j} \quad \forall i,j = 1,\ldots,n$$

Multiplying the two members of this equality by W^j/X^i, we find the expression of the elasticity η_j^i of the demand for factor i with respect to the price W^j of factor j. It is:

$$\eta_j^i = \frac{W^j}{X^i}C_{ij} + \left(\frac{YC_{iY}}{X^i}\right)\frac{W^j}{Y}\frac{\partial Y}{\partial W^j}$$

According to (2.27), the term $(W^j/X^i)C_{ij}$ represents the elasticity $\bar{\eta}_j^i$ of conditional demand taken at the profit optimum. Since, following Shephard's lemma, $X^i = C_i$, the term (YC_{iY}/X^i) designates the output elasticity of the conditional demand for input i, we denote it by $\bar{\eta}_Y^i$. Let η_j^Y again designate the elasticity of production with respect to W^j; the end result is:

$$\eta_j^i = \bar{\eta}_j^i + \bar{\eta}_Y^i\eta_j^Y \quad \forall i,j = 1,\ldots,n \tag{2.31}$$

This relation reveals the effects of a rise in price W^j of factor j on the demand for factor i. When $i = j$, relation (2.31) supplies the expression of the direct elasticity of factor i with respect to its price. The substitution effect represented by the direct conditional elasticity $\bar{\eta}_i^i$ is negative. Reasoning analogous to that followed in the case of inputs will show that the scale effect $\bar{\eta}_Y^i\eta_i^Y$ is also negative. Conversely, when $i \neq j$, the term $\bar{\eta}_j^i$ no longer has a determinate sign. We have seen that it is positive (or negative) if the factors i and j, $i \neq j$, are p-substitutes (or p-complements). The second term of the right-hand side of relation (2.31), or $\bar{\eta}_Y^i\eta_j^Y$, reveals a scale effect which, as in the case of two inputs, has an indeterminate sign, except when the production function is homogeneous (in which case it is negative). In sum, it is not possible to state truly general rules regarding the sign of cross elasticity η_j^i for $i \neq j$. That is why it is best to continue with the definitions already given in the case of two inputs. Thus, factors i and j form *gross substitutes* if $\eta_j^i > 0$. They are described as *gross complements* when $\eta_j^i < 0$.

The Case of a Homogeneous Production Function

Proceeding in the same fashion as in section 1.3.2, it is easy to show that if the production function is homogeneous of degree θ, the output elasticity $\bar{\eta}_j^i$ of the conditional demand for factor j is equal to $1/\theta$, for all $j = 1,\ldots,n$. Since, in this case, $C_Y = C/\theta Y$, taking the logarithmic derivatives of the optimality condition (2.30) with respect to W^j, we find, after several calculations:

$$\eta_j^Y = \frac{\theta\nu}{\theta - \nu}s^j$$

Since the second-order conditions dictate $\nu > \theta$, we verify that η_j^Y is negative. A rise in W^j thus entails a negative scale effect measured by the ratio $\nu s^j/(\theta - \nu)$. Finally, if we bring the value of $\bar{\eta}_j^i$ elicited from (2.29) into relation (2.31), we arrive at a formula giving the expression of the cross elasticity η_j^i of the unconditional demand for factor

i with respect to the cost of factor j when the production function is homogeneous of degree θ. It is written:

$$\eta_j^i = s^j \left(\sigma_j^i - \frac{\nu}{\nu - \theta} \right) \quad \forall (i, j) \tag{2.32}$$

This formula generalizes relations (2.22), which applied to the case with two inputs. The observations made there on the respective importance of market power ν, the possibilities of substitution between two factors (represented now by the variable σ_j^i) and the share s^j of the cost of a factor in the total cost, still hold true here and need not be repeated. But the formula (2.32) now allows us to take into account the heterogeneity of the labor factor. If we consider two categories of manpower—skilled and unskilled workers for example—we see that the elasticity of demand for skilled (or unskilled) workers with respect to the cost of unskilled (or skilled) workers is proportional to the share of unskilled (or skilled) manpower in the total cost. These conclusions have to be kept in mind when we want to analyze the effects of a change in minimum wage or a reduction in social security contributions as they apply to unskilled labor. Relation (2.32) shows that if skilled and unskilled workers are gross substitutes—which is the case when $\sigma_j^i > \nu/(\nu - \theta)$—a rise in the cost of unskilled labor provokes a reduction in the demand for unskilled workers and an increase in the demand for skilled ones. Conversely, if $\sigma_j^i < \nu/(\nu - \theta)$, the two categories of workers are gross complements and the rise in the cost of unskilled labor has the effect of reducing the utilization of the two categories of manpower at the same time.

Assuming that workers and hours worked can be considered as different inputs, relations (2.29) and (2.32) then allow us to study the determinants of substitution between workers and hours. To that end, we have to define the relative costs of each of these factors and the technical possibilities of substituting one for the other.

1.5 The Trade-off Between Workers and Hours

It becomes necessary to distinguish the number of workers from the number of hours worked whenever, on one hand, workers and hours are not perfectly substitutable and, on the other, the costs attached to using these two dimensions of the workforce are not identical. The solution to the problem of the firm makes it clear that demands for workers and hours depend on the relative importance of these two costs.

1.5.1 The Distinction Between Workers and Hours

In order to grasp the determinants of the trade-off between workers and hours, it is necessary to distinguish the contributions of these two elements to the production process and to differentiate between the costs arising from an increase in the number of employees and those that arise from a change in the number of hours worked by each employee.

Heterogeneity in the Number of Hours Worked

It is all the more important to make the distinction between workers and hours, to the extent that labor markets show strong heterogeneity when it comes to the number of hours in the working week. Figure 2.4 shows the distribution of time worked by employees in the United States over 2012. We see that the distribution of hours has a spike at

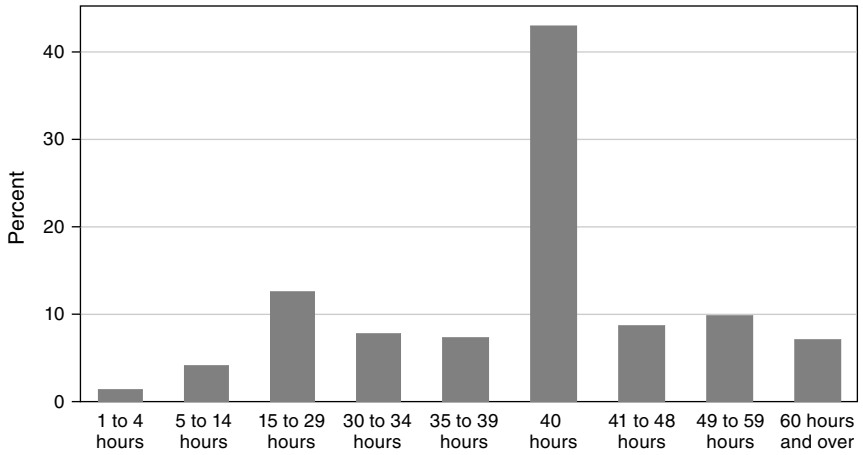


FIGURE 2.4
Distribution of the length of the working week in 2012 in the United States.

Source: Current population survey.

40 hours, which equals standard hours. This figure clearly shows that firms use a large range of options in scheduling work, which makes it important to understand the causes and consequences of these choices.

The Imperfect Substitutability of Workers and Hours

To this point we have not made a distinction between the number of employees in a firm and the overall amount of time that they devote to production. In the case of a production function $F(K,L)$ having only two factors, we thus implicitly assumed that labor services L were simply equal to the product NH, where N designates the number of persons employed and H represents the average individual length of time worked, expressed in hours. But that is a very special perspective because it assumes that workers and hours are perfectly substitutable: the firm would then choose its amount of hours without any thought for the manner in which that amount was divided up among its workforce. This kind of choice is only conceivable if the productivity of an hour of work and the rate of utilization of capital do not depend on the average individual length H of time worked, in other words, if the production of two individuals each working four hours a day is identical to that of a single one working eight hours a day. There are many reasons to think that this is not the case. Set-up costs entail that the relationship between the productivity of an hour of work and the length of working time exhibits increasing returns for small values of the latter. Beyond a certain threshold, fatigue will set in and this relation will exhibit decreasing returns. Moreover, when the duration of individual work changes, the duration of capital utilization, and thus its cost, likewise change if the firm undertakes no reorganization.

Accordingly, the production process should be represented by a function $F(K,N,H)$ having three arguments, which does effectively allow us to distinguish the marginal productivity of workers from that of hours. However, the properties of

demand functions when there are more than two inputs, set out in section 1.4, do not directly apply here, since there is no simple way to separate the cost of labor into a cost assignable to workers and a cost assignable to hours. For that reason we choose a less general representation of technology, but one with the advantage of allowing us to characterize the main elements in the workers/hours trade-off. We will often use the notion of *efficiency* in connection with number of hours worked. It is represented by an increasing function denoted $e(H)$. This function can reveal effects that run counter to each other. Set-up costs should cause the marginal efficiency $e'(H)$ of the number of hours worked to increase for small values of H, the effects of fatigue as the hours pass should cause marginal efficiency to decrease for larger values of H, and in consequence the function ought to be concave past a certain threshold. In sum, if N designates the number of persons employed in the firm, then labor services are expressed by the product $Ne(H)$, assuming, for the sake of simplicity, that all employees work the same amount of hours.

Likewise, the duration of capital utilization depends on H. Denoting this duration by $d(H)$, capital services are expressed by the product $Kd(H)$ where K designates the stock of capital. The production function is then written $Y = F[Kd(H), Ne(H)]$. In what follows, we assume, for the sake of simplicity, that the duration of capital utilization is a constant normalized to 1. In other words, the duration of capital utilization is independent of the individual duration H of work. In that case, any change in the latter necessitates a reorganization of the production process, since the employees are working different hours, but the duration of capital utilization has not changed. This might lead to new work schedules and eventually to a complete rearrangement of shifts in the plant.

The Cost of Labor

The distinction between workers and hours assumes greater importance in light of the fact that the cost of labor is not a linear function of its duration for at least two reasons (Rosen, 1968; Hart, 1987). In the first place, certain costs do not depend on duration, principally the costs of hiring and firing, training costs, and certain social security contributions. We will assume that they can be represented by a single positive scalar, equal to Z for each person employed. These costs can be defined on the basis of different periods, like the day, the week, the month, or the year. For the sake of clarity, we henceforth take the week as the period of reference. In the second place, in many countries there exists a *legal*, or *standard*, work duration and every *overtime* hour worked past that limit is remunerated at a higher rate than regular, or *standard*, hours. For example, in the United States the Fair Labor Standards Act of 1938 defines the standard work week as 40 hours and lays down an overtime rate 50% higher for hours worked past that limit. We will use T to designate the standard work week, Ω to designate the wage for a normal hour, and x to designate the overtime premium. There is generally an absolute limit, legal or physical, on how long anyone can work, but for simplicity we do not consider that here. If R continues to represent the utilization cost of a unit of capital, then the total cost of production is written:

$$C = \begin{cases} [\Omega T + (1+x)\Omega(H-T) + Z]N + RK & \text{if } H > T \\ (\Omega H + Z)N + RK & \text{if } H \leq T \end{cases} \tag{2.33}$$

This expression of the total cost shows that labor demand, here the number of persons employed and hours worked, depends on the comparison between the value of the *variable costs*, represented by Ω and x, and that of the *fixed costs*, represented by Z. Intuition suggests that a reduction in fixed costs gives firms an incentive to substitute workers for hours and thus ought to favor employment. Conversely, a reduction in variable costs ought to increase the number of hours worked, to the detriment of employment. The demand for workers and the demand for hours may thus vary in inverse directions. This logic does not, though, take into account the fact that the firm can also substitute labor services as a whole for those of capital. In order to assess the importance of these different effects, it is therefore necessary to know more precisely the expressions of the demand for workers and for hours.

1.5.2 The Optimal Number of Hours

Drawing upon the notion of efficient labor, the demand functions result from an optimization problem with just two inputs. On this basis, it is easy to show that the optimal number of hours worked depends mainly on how high fixed costs are with respect to variable costs.

Efficient Labor and Minimization of Total Cost

Taking into account cost C defined by equation (2.33), for a given output level Y, the conditional factor demands correspond to the solutions of the following problem:

$$\min_{\{H,N,K\}} C \quad \text{subject to } F[K, Ne(H)] \geq Y$$

If we proceed directly to consider the quantity L of *effective* labor defined by $L \equiv Ne(H)$, this problem takes the form:

$$\min_{\{H,L,K\}} (WL + RK) \quad \text{subject to } F(K, L) \geq Y$$

where the unit cost W of efficient labor is given by:

$$W = \begin{cases} [\Omega T + (1+x)\Omega(H-T) + Z]/e(H) & \text{if } H \geq T \\ (\Omega H + Z)/e(H) & \text{if } H \leq T \end{cases} \tag{2.34}$$

Thus we see that the minimization of the cost of production can be carried out in two stages. In the first stage, we look for the *optimal* number of hours corresponding to the value of H that minimizes the unit cost W. In the second stage, we calculate the values of L and K that minimize the total cost of production, given this optimal value of W. This last problem involves only two inputs (K and L) with costs (W and R) that are given. The properties of the solutions then flow directly from the results we already reached in section 1.2.2.

Relation (2.34) shows that the unit cost W is a function of H, which is not differentiable at point $H = T$. To find the value of H minimizing this function, we thus have to compare its local minima over intervals $H > T$ and $H \leq T$. These calculations

are presented in appendix 7.3, assuming for the sake of simplicity that the elasticity of function $e(H)$ is a positive constant η_H^e belonging to the interval $[0,1]$. This hypothesis may seem restrictive, but it is corroborated by empirical studies (see section 2.2.2). The optimal value H^* of the number of hours is defined by:

$$H^* = \begin{cases} \eta_H^e Z/\left(1-\eta_H^e\right)\Omega \le T & \text{if } Z/\Omega T \le \left(1-\eta_H^e\right)/\eta_H^e \\ T & \text{if } \left(1-\eta_H^e\right)/\eta_H^e \le Z/\Omega T \le \left(1+x-\eta_H^e\right)/\eta_H^e \\ \eta_H^e(Z-x\Omega T)/(1+x)\left(1-\eta_H^e\right)\Omega \ge T & \text{if } Z/\Omega T \ge \left(1+x-\eta_H^e\right)/\eta_H^e \end{cases} \tag{2.35}$$

We should first note that the optimal number of hours depends neither on quantity K of capital nor on level Y of output; this is a consequence of the particular form of the production function and would no longer hold if the technology were described by any function $F(K,N,H)$. It does however fit well with observation, for there is little difference among the actual numbers of hours worked in firms that are large and small, capital intensive and not.

We also see that the optimal value of H depends on the elasticity η_H^e of the function $e(H)$ measuring the efficiency of the number of hours worked by individuals. In this respect, it is illuminating to consider first the case where $\eta_H^e < 1$, and then the case where $\eta_H^e = 1$. When the elasticity of the efficiency of an employee with respect to hours is small (η_H^e is close to 0), the firm does not utilize overtime hours, for to do so would increase efficiency by only a small amount. On the other hand, the more the efficiency of labor depends on its duration—the more η_H^e approaches 1— the more the firm will tend to resort to overtime hours. When $\eta_H^e = 1$, workers and hours are perfectly substitutable. The interior solutions, described by equation (2.35), are no longer defined.

It should also be noted that the number of hours is low ($H \le T$) when the fixed cost Z is small in comparison with variable cost ΩT corresponding to standard hours. Conversely, the firm uses overtime hours ($H \ge T$) when the ratio $Z/\Omega T$ grows larger, that is, when the level of variable cost ΩT sinks relative to that of fixed cost Z. The optimal number of hours coincides with standard hours ($H = T$) for the intermediate values of ratio $Z/\Omega T$. In this situation the firm desires to set its number of hours beyond that of the legal limit T, but the rate x of extra pay for overtime hours proves too high for it to do so. The optimal solution is then $H = T$.

Fixed Costs, Variable Costs, and the Number of Hours Actually Worked

Relations (2.35) show precisely how the optimal number of individual hours of work varies when the exogenous parameters change. We have:

$$\begin{aligned} &\frac{\partial H^*}{\partial Z} \ge 0, \quad \frac{\partial H^*}{\partial \Omega} \le 0 \quad \text{and} \quad \frac{\partial H^*}{\partial x} \le 0, \\ &\frac{\partial H^*}{\partial T} = 0 \text{ if } H^* < T, \quad \frac{\partial H^*}{\partial T} = 1 \text{ if } H^* = T, \quad \text{and} \quad \frac{\partial H^*}{\partial T} < 0 \text{ if } H^* > T \end{aligned} \tag{2.36}$$

As intuition would suggest, a rise in fixed costs Z tends to increase the number of hours, while an increase in variable costs Ω or x tends to reduce it.

A change in standard hours has contrasting effects according to whether or not the firm makes use of overtime. In particular, when the optimal number of hours exceeds the legal limit ($H > T$), a reduction in the latter raises the number of hours worked by

all employees. In other words, a *reduction* in standard hours has the effect of *increasing* the actual work week by causing the number of overtime hours to rise. This result seems counterintuitive at first sight, and it runs counter to the overt purpose of a reduction in standard hours, which is precisely to bring down the actual number of hours worked by every individual so as to increase the number of jobs. But if we look closely at relation (2.35), which defines the optimal number of hours, we find that it arises because the propensity to make use of overtime, that is, the ratio H/T, does not depend on Z but on the *ratio* $Z/\Omega T$. A lowering of standard hours is thus like a relative *rise* in fixed costs (see Calmfors and Hoel, 1988). The variable costs have sunk in relative importance, and thus we can see why the firm would increase the number of hours actually worked (in the following paragraph, we will demonstrate that this increase ought, as a general rule, to occur at the expense of the number of jobs). On the other hand, a reduction in standard hours evidently leads to a reduction in the actual work week when these two variables are equal ($H = T$). It should be noted, however, that in this situation a drastic reduction in standard hours might cause firms to start making use of overtime, and we would no longer have the equality $H = T$.

1.5.3 Cost of Labor and Demand for Workers

The demand for workers is deducible from the optimal amount of efficient labor. When there are opportunities to trade off between workers and hours, analysis of the impact on employment of variations in the elements that influence the cost of labor requires very lengthy calculations. We begin by presenting these calculations, before summarizing them and giving quantitative results in tables 2.1 and 2.2. Readers pressed for time may refer directly to these tables in order to get an idea of the underlying economic mechanisms and the relevant orders of magnitude.

A Synthetic Formula

Given the optimal values of H and W, we have seen that total cost minimization took the form studied in section 1.2.1. The solutions of this minimization correspond to the

Table 2.1
The signs of the elasticities of hours worked and the conditional demand for workers.

	η_Z^H	η_Ω^H	η_x^H	η_T^H	$\overline{\eta}_Z^N$	$\overline{\eta}_\Omega^N$	$\overline{\eta}_x^N$	$\overline{\eta}_T^N$
$H^* < T$	+	−	0	0	−	+(a)	0	0
$H^* = T$	0	0	0	+	−	−	0	−(a)
$H^* > T$	+	−	−	−	−	+(a)	+	+

Note: (a) if $\overline{\eta}_w^L$ is less than 1 in absolute value.

Table 2.2
Values of elasticities of hours and conditional demand for workers.

	η_Ω^H	η_x^H	η_T^H	$\overline{\eta}_\Omega^N$	$\overline{\eta}_x^N$	$\overline{\eta}_T^N$
$H^* = 0.9 \times T$	−1	0	0	0.63	0	0
$H^* = T$	0	0	1	−0.21	0	−0.96
$H^* = 1.04 \times T$	−3	−2.23	−2	2.49	2.00	1.86

conditional demands for capital and efficient labor, and we continue to denote them by $\bar{K}$ and $\bar{L}$. They are functions of W^*, R, and Y, where W^* designates the optimal value of the unit cost of efficient labor. It is given by relation (2.34) when we replace H by its optimal value H^* defined by (2.35). We thus arrive at:

$$W^* = \begin{cases} (\Omega H^* + Z)/e(H^*) & \text{if } Z/\Omega T \leq \left(1 + x - \eta_H^e\right)/\eta_H^e \\ (1+x)\Omega H^*/\eta_H^e e(H^*) & \text{if } Z/\Omega T \geq \left(1 + x - \eta_H^e\right)/\eta_H^e \end{cases} \tag{2.37}$$

Remember also that in reality $\bar{L}$ is only an auxiliary variable linked to the conditional demand for workers $\bar{N}$ by relation $\bar{L} \equiv \bar{N}e(H)$. If v represents one of the parameters Ω, x, Z, or T, differentiating this identity then implies:

$$\bar{\eta}_v^N = \bar{\eta}_v^L - \eta_H^e \eta_v^H \quad \forall v = (\Omega, x, Z, T)$$

In this expression, $\bar{\eta}_v^N$ and $\bar{\eta}_v^L$ designate respectively the elasticities of $\bar{N}$ and of $\bar{L}$ with respect to v, and η_v^H represents the elasticity of the optimal number of hours with respect to this parameter. Since $\bar{L}$ depends only on W^*, R, and Y, $\bar{\eta}_v^L$ will always equal $\bar{\eta}_W^L \eta_v^W$ where $\bar{\eta}_W^L$ and η_v^W are respectively the elasticity of $\bar{L}$ with respect to cost W of efficient labor taken at W^*, and the elasticity of W^* with respect to parameter v. We thus finally get:

$$\bar{\eta}_v^N = \bar{\eta}_W^L \eta_v^W - \eta_H^e \eta_v^H \quad \forall v = (\Omega, x, Z, T) \tag{2.38}$$

This relation allows us to deduce the properties of conditional demand $\bar{N}$ from those of $\bar{L}$, W, and H. It shows that in order to attain a given output level, it is possible for the firm to substitute employees for hours, in which case $\bar{N}$ and H vary in opposite directions. This eventuality is represented by the term $-\eta_H^e \eta_v^H$. But the firm can also substitute labor services as a whole (employees and/or hours) with capital services. The term $\bar{\eta}_W^L \eta_v^W$ conveys this possibility. Thus $\bar{N}$ and H do not necessarily vary in opposite directions and the comparative statics of the demand for workers is not directly deducible from that of hours worked. We must also take into account capital/labor substitution, encapsulated in the extent of elasticity $\bar{\eta}_W^L$. According to the laws of demand, we know only that $\bar{\eta}_W^L < 0$, but all the estimates carried out in this domain indicate that the latter is, in absolute value, clearly inferior to unity (see section 2.2.1 below).[9] We may thus assume, without gravely compromising what follows, that the absolute value of $\bar{\eta}_W^L$ belongs to the interval $[0, 1]$. We assume that this spread of variation also applies, for that matter, to unconditional elasticity η_W^L. We begin by discussing the general results, insisting on their economic interpretation, then go on to give orders of magnitude for a particular form of the production function and probable values for the parameters.

[9] Rigorously speaking, the term $\bar{\eta}_W^L$ means something measurably different compared to what it represented before: the elasticity of the labor demand, expressed in terms of hours or number of employees, with respect to its cost. Now, L refers to a number of units of efficient labor. But the function linking the demand for labor to its cost is, by construction, identical to that linking the demand for efficient labor to the cost of efficient labor. The elasticity $\bar{\eta}_W^L$ is thus the same in the two configurations. Relation (2.12) indicates that $\bar{\eta}_W^L = -(1-s)\sigma$, where s designates the share of labor cost in total cost and σ the elasticity of substitution between capital and labor. We will see further on that the majority of empirical studies suggest that σ is smaller than 1, and even close to 1 on the basis of macroeconomic data. The absolute value of $\bar{\eta}_W^L$ is thus likely smaller than 1.

Variations in Fixed Costs

The first-order condition of the minimization of the unit cost of efficient labor dictates that the optimal value of the number of hours worked should be such that $\partial W^*/\partial H = 0$ for $H^* \neq T$. At the optimum, we thus have:

$$\frac{\mathrm{d}W}{\mathrm{d}Z} = \frac{\partial W^*}{\partial H}\frac{\partial H^*}{\partial Z} + \frac{\partial W^*}{\partial Z} = \frac{\partial W^*}{\partial Z}$$

Definition (2.37) of W^* shows that $\partial W^*/\partial Z$ is always positive for all H^*. Consequently η_Z^W is positive, and as we know that $\bar{\eta}_W^L \leq 0$ and $\eta_Z^H \geq 0$, relation (2.38) entails $\bar{\eta}_Z^N \leq 0$. As intuition suggests, a rise in the fixed costs of labor tends to increase utilization of hours to the detriment of the number of workers and to favor the utilization of capital over labor. These two effects combine to reduce the number of workers.

The study of variations in the demand for workers as a function of other parameters proves a more delicate business. It is best to pursue it by distinguishing situations in which the firm utilizes overtime from those in which it does not.

Variations in the Hourly Wage

- When $H^* < T$, relation (2.35) shows that, setting fixed costs aside, the number of hours worked depends only on the hourly wage Ω. More precisely, we see that the elasticity of an individual's hours of work with respect to the hourly wage, or η_Ω^H, is equal to -1. It is possible to save several calculations by noting, with the help of equations (2.37) and (2.35), that the optimal values of W and H satisfy $W^* = \Omega H^*/\eta_H^e e(H^*)$. Differentiating this equality with respect to Ω, we get:

  $$\eta_\Omega^W = 1 + (1 - \eta_H^e)\eta_\Omega^H = \eta_H^e$$

  Bringing this value of η_Ω^W into (2.38), we finally arrive at:

  $$\bar{\eta}_\Omega^N = \eta_H^e(1 + \bar{\eta}_W^L) = \eta_H^e[1 - (1 - s)\sigma]$$

  As we adopt the hypothesis that $(1 - s)\sigma$ is inferior to unity, we have $\bar{\eta}_\Omega^N \geq 0$, which signifies that an increase in the hourly wage entails an increase in employment at the expense of hours. In other words, it would be necessary for the elasticity of capital/labor substitution to be very great, which is unlikely, for a rise in the hourly wage to be accompanied both by a reduction in the number of hours of work and by a reduction in employment. To attain a given output level, firms prefer to substitute workers for hours rather than to substitute capital for workers.
- If $H^* = T$, the optimal value of W is given by:

  $$W^* = \frac{\Omega T + Z}{e(T)}$$

  It is evident immediately that W^* rises with Ω (thus $\eta_\Omega^W \geq 0$) and, as η_Ω^H is null, (2.38) then implies $\bar{\eta}_\Omega^N \leq 0$. Differently to the previous case, the level of employment falls when the level of the hourly wage rises. This result is not

hard to understand: a rise in the cost of labor means that the firm uses less of this factor and more capital to attain the same output level. Since hours worked do not vary ($H = T$), the adjustment necessarily takes place through a reduction in employment.

- If $H^* > T$, equation (2.37) defining W^* gives $\eta_\Omega^W = 1 + (1 - \eta_H^e)\eta_\Omega^H$. Bringing this value of elasticity η_Ω^W into (2.38) with $v = \Omega$, we get:

$$\bar{\eta}_\Omega^N = \bar{\eta}_W^L \eta_\Omega^W - \eta_H^e \eta_\Omega^H = \bar{\eta}_W^L + \eta_\Omega^H \left[(1 - \eta_H^e)\bar{\eta}_W^L - \eta_H^e \right] \tag{2.39}$$

The expression (2.35) of the optimal number of hours of work implies, after several calculations, $\eta_\Omega^H = -Z/(Z - x\Omega T) < -1$. Taking this inequality into account, relation (2.39) entails $\bar{\eta}_\Omega^N > \eta_H^e(1 + \bar{\eta}_W^L)$. As we may consider that elasticity $\bar{\eta}_W^L$ is smaller in absolute value than unity, a rise in the hourly wage will lead to an increase in the number of workers. Consequently, when the hourly wage rises, firms reduce individual hours of work, and in order to attain a given output level, the elasticity of substitution between capital and labor would have to reach unimaginable values for firms to reduce their demand for workers as well.

Variations in the Overtime Premium

A variation in the overtime premium x influences the optimal level of hours worked only when the latter exceeds standard hours T. Differentiating equation (2.37) with respect to x, for $(Z/\Omega T) > (1 + x - \eta_H^e)/\eta_H^e$, after several rearrangements we find $\eta_x^W = -x\Omega T/(Z - x\Omega T) - \eta_H^e \eta_x^H$. The sign of η_x^W is thus ambiguous since η_x^H is a negative quantity. In bringing this value of η_x^W into (2.38) with $v = x$, however, we arrive at:

$$\bar{\eta}_x^N = \bar{\eta}_W^L \eta_x^W - \eta_H^e \eta_x^H = -\eta_H^e \eta_x^H (1 + \bar{\eta}_W^L) - \bar{\eta}_W^L \frac{x\Omega T}{Z - x\Omega T}$$

It is evident that an increase in x *increases* the conditional demand for workers once we assume that $\bar{\eta}_W^L$ is, in absolute value, smaller than 1. The explanation is the same as that for a rise in hourly wage: any increase in the variable cost leads to a reduction in individual hours worked, and the possibilities of capital/labor substitution would have to extend farther than any empirical study warrants in order for firms to have an interest in reducing their level of employment as well.

The Reduction in Standard Hours

A change in standard hours T acts on the actual work week H whenever H is not inferior to T. It is evident that the impact of such a change is not the same when $H > T$ and when $H = T$.

- If $H > T$, the derivative with respect to T of equation (2.37) defining W^* yields the equality $\eta_T^W = (1 - \eta_H^e)\eta_T^H$. Since, following (2.36), η_T^H is negative, it is certain that η_T^W is also negative. In these conditions, relation (2.38) with $v = T$ indicates that the effect of substituting hours for workers $(-\eta_H^e \eta_T^H)$ is positive and that the effect of substituting labor for capital $(\bar{\eta}_W^L \eta_T^W)$ is equally positive. In consequence, we may conclude unambiguously that $\bar{\eta}_T^N > 0$. In other words, a

reduction in standard hours has the effect of *diminishing* the demand for workers, which probably runs directly counter to the objective aimed at with such a measure. This result, which may cause surprise, springs from the fact that a reduction in standard hours is the exact equivalent of a reduction in variable costs compared to fixed costs, which, as we have seen, will provoke an *increase* in the actual number of hours worked (and a more intensive use of capital), to the detriment of the number of persons employed.

- If $H = T$, the impact of a rise in T is a priori ambiguous. On one hand, this rise amounts to a reduction in fixed costs, which tends to reduce employment, but on the other, it also signifies that the efficiency of labor, $e(T)$, is raised, which may give the firm an incentive to raise its employment level. To escape this ambiguity, we have to be able to assign an order of magnitude to the different elasticities that occur in formula (2.38). Noting that $\eta_T^W = [\Omega T/(\Omega T + Z)]$ and $\eta_T^H = 1$, relation (2.38) gives:

$$\bar{\eta}_T^N = \bar{\eta}_W^L \eta_T^W - \eta_H^e \eta_T^H = \bar{\eta}_W^L \left[\frac{\Omega T}{\Omega T + Z} - \eta_H^e \right] - \eta_H^e$$

Using the existence conditions (2.35) for the solution $H^* = T$, it is evident that $\bar{\eta}_T^N$ is negative given that the absolute value of $\bar{\eta}_W^L$ is inferior to $(1 + x)/x$. Since the hypothesis of an absolute value of $\bar{\eta}_W^L$ inferior to unity is the most probable one, we can conclude that $\bar{\eta}_T^N \leq 0$. Thus, a reduction in standard hours leads to a *rise* in employment when the actual work week coincides with the standard one. In this case, a reduction in standard hours is equivalent to a reduction in fixed costs, which has the effect of increasing employment. It is evident that this last effect outweighs the countervailing effect on productivity (a reduction in hours worked reduces average production per employee, which may give the firm an incentive to restrain its demand for workers).

Synthesis of Results

The signs of the elasticities of the conditional demands for workers and hours with respect to the various parameters are summarized in table 2.1. The reader will see that the behavior of firms is very different, according to whether they utilize overtime hours or not. When the optimal number of hours H^* differs from standard hours T, a rise in the hourly wage induces an extension of working time, and in general, an increase in employment. Conversely, when the work week chosen by the firm is equivalent to standard hours, a rise in the hourly wage reduces employment.

Reducing standard hours probably leads to increased employment in firms where the optimal work week is equivalent to the standard one. Actually, for a given level of production, the reduction of hours has two opposing effects on employment. It gives the firm incentive to hire more workers in order to meet its orders. But it also produces a rise in the fixed costs of labor, which pushes firms to substitute capital for labor. The first effect dominates for reasonable values of the elasticity of substitution between capital and labor. Reducing standard hours has a different impact on employment for firms that resort to overtime hours. A reduction in standard hours pushes these firms to *increase* hours worked by using more overtime hours. This increase in the actual work week,

combined with the rise in the cost of labor flowing from the remuneration of overtime hours, leads to a reduction in employment.

Finally, table 2.1 shows that an increase in the overtime premium pushes firms to reduce hours worked. The impact on employment is positive for probable values of the elasticity of substitution between capital and labor. The empirical study conducted by Hamermesh and Trejo (2000) on data from California confirms the result, according to which an increase in the overtime premium reduces the hours worked. Hamermesh and Trejo find an elasticity η_x^H of -0.5.

Some Quantitative Results

Table 2.2 gives the values for the elasticities of optimal hours and employment, assuming that the share s of the cost of labor in the total cost is equal to 0.7 and that the elasticity of substitution between capital and labor is equal to one. As we shall see, empirical studies suggest that such values are relevant for an aggregate production function that represents the technology of the economy as a whole. This implies that $\bar{\eta}_W^L = -(1-s)\sigma = -0.3$. We assume further that the elasticity of labor efficiency η_H^e is equal to 0.9. Relation (2.35) shows that firms in which the ratio of the fixed cost of labor to the variable cost corresponding to standard hours, or $(Z/\Omega T)$, is less than 0.11 choose a work week shorter than the standard one. The optimal number of hours is equal to standard hours if $(Z/\Omega T)$ lies somewhere between 0.11 and 0.44. The firm resorts to overtime when $(Z/\Omega T)$ is greater than 0.44. Thus we distinguish three types of firm according to the level of their fixed cost: (1) those with a share of fixed cost $(Z/\Omega T)$ equal to 10% and whose work week is equal to 90% of standard hours, following relation (2.35); (2) firms for which $(Z/\Omega T) = 0.3$ and whose work week is the same as the standard one; and (3) firms for which $(Z/\Omega T) = 0.45$ and whose optimal work week is 4% longer than the standard one, assuming that the overtime premium x is equal to 30%.

Table 2.2 shows that variations in hourly wage have very different effects on employment, since elasticity $\bar{\eta}_\Omega^N$ runs from -0.21 to 2.49, when the only source of heterogeneity in firms is the extent of the fixed costs of labor. The same observation can be made about a reduction in the number of hours worked, which allows employment to be significantly increased (at a given hourly wage) when the actual number of hours is the same as the standard one, but has a very strong negative effect on employment in firms that make use of overtime. From this point of view, it is interesting to note that a reduction in standard hours has often been proposed in the United States (see Hamermesh, 2006) and adopted in certain European countries like Germany in the 1980s and France in the 1980s and the 2000s in order to increase employment in periods of recession. Models of labor demand suggest, however, that the effects of this measure on employment are ambiguous, which the empirical studies made in the case of Germany by Hunt (1999) and in the case of France by Crépon and Kramarz (2008) confirm.

Taking Scale Effects into Account

To this point we have assumed that output level Y was given. But when the firm maximizes its profit, this level becomes a choice variable, and so-called scale effects (see sections 1.3.2 and 1.4.2) have to be added to the results obtained when Y was fixed. Since we are interested in the impact of changes in standard hours, the hourly wage,

and overtime premium on labor demand at the macroeconomic level, scale effects have to be gauged by taking into account variation in the cost of labor across the whole economy and not in one firm alone. These scale effects must therefore be calculated using relation (2.20). Formally, it suffices to replace conditional elasticity $\bar{\eta}_W^L$ by unconditional elasticity η_W^L in all the equations in this section. Empirical studies show, as we will see, that the term η_W^L is certainly negative and that in absolute value it is superior to $\bar{\eta}_W^L$, since it is derived from it by adding scale effects, which are negative—see relations (2.19), (2.20), and (2.22) in section 1.3.2. Empirical studies suggest that -0.5 is a probable order of magnitude for η_W^L at the macroeconomic level, whereas the value of conditional elasticity $\bar{\eta}_W^L$ that we have used for an individual firm is -0.3. The difference between these two elasticities is thus slight, which implies that taking scale effects into consideration does not modify the conclusions reached for a given output level very much when we are at the macroeconomic level.

To be more precise: taking scale effects into consideration does not modify our results concerning the actual duration H^*, which is independent of the output level. Nor does it modify our results relative to a rise in fixed costs on employment: when Z rises, we have seen that conditional demand $\bar{N}$ for workers diminishes, and since scale effects do not affect the length H^* of the work week, the rise in Z must indeed diminish the demand for workers. Conversely, scale effects might affect results concerning the signs of the impact of variations in variable costs Ω and x on employment. But for that, unconditional elasticity η_W^L would have to take values higher than unity. At the aggregate level, this eventuality is not in the least realistic.

Taking scale effects into consideration leads to a greater absolute value for elasticity of employment with respect to standard hours when $H^* > T$. This is because scale effects have a tendency to accentuate the impact of the rise in the cost of labor on employment. When the actual work week is the same length as the standard one, $H^* = T$, the reduction in hours, which has a positive impact on employment for a given level of production, has a smaller impact, which can even become negative if the scale effect is large. Nevertheless, with a value of η_W^L equal to -0.5, the elasticity of employment with respect to standard hours amounts to -0.79. So it remains negative, which means that a reduction in standard hours always creates jobs. Overall though, reducing standard hours tends to be more unfavorable for employment than it would be in a setting where production was given.

It is important to emphasize, in closing this discussion of the trade-off between workers and hours, that what we have done is to look at the impact of variations in standard hours and the overtime premium, while taking the hourly wage as given. Now there are good reasons to think that the hourly wage is influenced by these two variables, for with a constant real wage a reduction in time worked entails a reduction in monthly earnings. We can well imagine that wage earners would resist such a reduction in income by demanding higher wages (see Cahuc and Zylberberg [2008] for an analysis of the impact of reduction in standard hours when wages react to changes in hours). Conversely, a rise in the overtime premium brings them extra income, and that might lead to a reduction in the hourly wage. The empirical study of Trejo (1991), carried out using North American data, finds this type of effect. These problems will be tackled in chapter 7, where we will study the setting of wages in the framework of collective bargaining models.

2 FROM THEORY TO ESTIMATE

Empirical studies based on the static theory of labor demand aim principally to estimate the different elasticities set out above. First we show how it is possible, on the basis of explicit functional forms, to utilize theoretical results in empirical investigations. We then sum up the main conclusions to be drawn from all the empirical work dedicated to labor demand.

2.1 Specific Functional Forms for Factor Demands

There are two methods for estimating the parameters of the factor demand functions. The first consists of postulating a particular production function on the basis of which it becomes possible to state explicitly the cost and profit functions; they in turn make it possible to arrive at the factor demands. The second is based directly on a cost function defined a priori, without specifying the associated production function.

2.1.1 The Choice of a Production Function

The solution of the problem of cost minimization allows us, with the help of a particular form of the production function, to obtain conditional demand functions in explicit form. The most commonly utilized production functions are the Cobb-Douglas type, and CES (for *Constant Elasticity of Substitution*).

The Cobb-Douglas Function with Two Factors

When we take into consideration no more than two different factors of production, for example, capital K and labor L, a Cobb-Douglas production function (Cobb and Douglas, 1928) possesses the following form:

$$Y = AK^{\theta(1-\alpha)}L^{\theta\alpha}, \qquad 0 < \alpha < 1, \qquad A > 0 \tag{2.40}$$

In this expression the parameter $\theta > 0$ designates the degree of homogeneity of the production function. It is easy to verify that the technical rate of substitution F_L/F_K is equal to $\alpha K/(1-\alpha)L$. Now, according to relation (2.5) from section 1.2.1, minimization of the total cost of production requires that this rate should coincide with the ratio of the cost of the factors. If W and R again designate respectively the unit costs of labor and capital, we get:

$$\frac{F_L}{F_K} = \frac{\alpha K}{(1-\alpha)L} = \frac{W}{R} \tag{2.41}$$

These equalities show that the capital/labor ratio K/L is proportional to the ratio W/R. Since by definition—see (2.10)—the elasticity of substitution σ between capital and labor measures precisely the elasticity of the ratio K/L with respect to relative cost W/R, we will have $\sigma = 1$ here. Moreover, relation (2.41) implies that the share s of labor in the total cost is simply equal to parameter α. Equation (2.12), which gives the value of the elasticities of conditional labor demand as functions of s and σ, then takes the following form:

$$\bar{\eta}_W^L = -\bar{\eta}_R^L = -(1-\alpha)$$

Using a Cobb-Douglas production function thus imposes very restrictive conditions with regard to the possibilities of substitution between the inputs—since σ is always equal to 1—but it does allow a very simple estimation of the elasticities of labor demand.

The expressions of the conditional factor demands are deduced from relations (2.40) and (2.41). After several calculations, we get:

$$\bar{L} = \left[\frac{\alpha}{(1-\alpha)}\frac{R}{W}\right]^{1-\alpha}\left(\frac{Y}{A}\right)^{1/\theta} \quad \text{and} \quad \bar{K} = \left[\frac{(1-\alpha)}{\alpha}\frac{W}{R}\right]^{\alpha}\left(\frac{Y}{A}\right)^{1/\theta}$$

The cost function $C(W,R,Y)$, equal by definition to $W\bar{L} + R\bar{K}$, is then written:

$$C(W,R,Y) = \left(\frac{W}{\alpha}\right)^{\alpha}\left(\frac{R}{1-\alpha}\right)^{1-\alpha}\left(\frac{Y}{A}\right)^{1/\theta}$$

The CES Function with Two Inputs

Let $\theta > 0$ continue to designate the degree of homogeneity; if we consider only two inputs K and L, the CES (*Constant Elasticity of Substitution*) function proposed by Arrow et al. (1961) is expressed this way:

$$Y = \left[(a_L L)^{\frac{\sigma-1}{\sigma}} + (a_K K)^{\frac{\sigma-1}{\sigma}}\right]^{\frac{\theta\sigma}{\sigma-1}}, \qquad \sigma > 0, \qquad a_K > 0, \qquad a_L > 0 \tag{2.42}$$

If we equalize the technical rate of substitution with the ratio of the costs of inputs, we get:

$$\frac{K}{L} = \left(\frac{R}{W}\right)^{-\sigma}\left(\frac{a_K}{a_L}\right)^{\sigma-1} \tag{2.43}$$

We thus observe that parameter σ represents the elasticity of substitution between the two inputs. It must also be noted that equation (2.43), when put in logarithmic form, makes it possible to estimate this elasticity of substitution in linear form. Relations (2.42) and (2.43) supply the conditional demands of the two inputs. After several calculations, we find the following expressions:

$$a_L\bar{L} = \left(\frac{W}{a_L}\right)^{-\sigma}\left[\left(\frac{W}{a_L}\right)^{1-\sigma} + \left(\frac{R}{a_K}\right)^{1-\sigma}\right]^{-\frac{\sigma}{\sigma-1}} Y^{1/\theta}$$

$$a_K\bar{K} = \left(\frac{R}{a_K}\right)^{-\sigma}\left[\left(\frac{W}{a_L}\right)^{1-\sigma} + \left(\frac{R}{a_K}\right)^{1-\sigma}\right]^{-\frac{\sigma}{\sigma-1}} Y^{1/\theta}$$

With the help of these two equations, we deduce the cost function, which comes to:

$$C(W,R,Y) = \left[\left(\frac{W}{a_L}\right)^{1-\sigma} + \left(\frac{R}{a_K}\right)^{1-\sigma}\right]^{\frac{1}{1-\sigma}} Y^{1/\theta}$$

2.1.2 The Choice of a Cost Function

Empirical studies aiming to estimate a cost function directly postulate an analytic form satisfying the theoretical properties of such a function, that is, concavity, homogeneity of degree 1 with respect to the costs of the factors, as well as being increasing with respect to the output level and the input quantities. Thanks to Shephard's lemma, the partial derivatives of the cost function give the conditional factor demands, which it thus becomes possible to estimate.

The Generalized Leontief Function (Diewert, 1971)

If we consider a production function homogeneous of degree $\theta > 0$ with n inputs, the generalized Leontief cost function is written:

$$C(W^1, ..., W^n, Y) = Y^{1/\theta} \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \left(W^i\right)^{1/2} \left(W^j\right)^{1/2}, \qquad a_{ij} = a_{ji}$$

Following Shephard's lemma (2.24), the conditional demand $\bar{X}^i$ for factor i is given by the partial derivative C_i of the cost function with respect to W^i. We thus get:

$$\bar{X}^i = Y^{1/\theta} \sum_{j=1}^{n} a_{ij} \left(\frac{W^i}{W^j}\right)^{1/2}$$

This expression allows us to estimate the coefficients a_{ij} and then from that to deduce the elasticities of substitution σ_j^i between two factors i and j by the formula:

$$\sigma_j^i = \frac{a_{ij} \left(W^i W^j\right)^{1/2}}{2 s^i s^j} \quad \forall (i, j), j \neq i$$

We see that the elasticity of substitution between two inputs is no longer a constant, for it depends on the costs of the factors as well as on the share of each input in the total cost. In this sense, it is less restrictive to utilize a generalized Leontief cost function than a CES production function to define and estimate the demand functions.

The Translog Cost Function (Christensen, Jorgenson, and Lau, 1973)

Assuming once more a production function homogeneous of degree $\theta > 0$ with n inputs, the translog (transcendental logarithmic) cost function is defined by:

$$\ln C = a_0 + \sum_{i=1}^{n} a_i \ln W^i + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \ln W^i \ln W^j + \frac{1}{\theta} \ln Y$$

In this expression, parameters a_i and a_{ij} must be such that $\sum_{i=1}^{n} a_i = 1$, $a_{ij} = a_{ji}$, $\sum_{j=1}^{n} a_{ij} = 0, \forall i = 1, ..., n$. For $a_i > 0$, and $a_{ij} = 0, i, j = 1, ..., n$, this function is of the Cobb-Douglas type. But in the general case ($a_{ij} \neq 0$), the conditional demand functions are not linear with respect to the parameters. With the help of Shephard's lemma, though, we

can show that the shares s^i of each factor are linear functions of the coefficients of the cost function. Thus we have:

$$s^i = a_i + \sum_{j=1}^{n} a_{ij} \ln W^j$$

It then becomes possible to estimate the parameters of this equation and from that to deduce the elasticities of substitution. The resulting expression is:

$$\sigma_j^i = \frac{a_{ij} + s^i s^j}{s^i s^j}, \qquad \forall (i,j), i \neq j, \qquad \sigma_i^i = \frac{a_{ii} - s^i + \left(s^j\right)^2}{\left(s^i\right)^2}$$

Here again, the elasticities of substitution are not constant and can vary among the factors, taken two at a time. The cross and direct elasticities of the conditional demand functions are subsequently obtained using relation (2.29).

2.2 Main Issues and Main Results

Much research has attempted to estimate the elasticities of labor demand and the possibilities of substitution. As is often the case in labor economics, the estimation of parameters faces the difficulty of isolating genuinely exogenous variables from pertinent explanatory variables like factor costs in the case of labor demand. The identification problem previously raised recurs. Numerous studies carried out down to the end of the 1990s estimated the elasticity of labor demand on the basis of cross-section or time-series data and treated this problem without clearly identifying exogenous variations in the factor costs paid by firms. These studies yielded results that are hard to interpret. Thus, in cross-section data the correlation between wages and employment might be due to the fact that the most productive firms hire more, and also pay higher wages. The researcher would then observe an increasing relation between wages and employment that might cause her to underestimate the elasticity of labor demand with respect to the cost of labor. Such identification problems do not disappear with time series. For example, a positive shock to the demand for a firm's goods might induce a simultaneous increase in employment and wages that would cause her to identify a positive relation between these two variables.

A strategy frequently utilized to isolate exogenous variations in the cost of labor is to exploit changes in the amount of the minimum legal wage (Neumark and Wascher, 2008). The effects of minimum wage are studied in chapter 12. We will see that they affect not only the demand for but also the supply of labor. Hence it is important to take these two dimensions into account in assessing the impact of minimum wage on employment. This impact cannot be systematically attributed solely to a reaction of labor demand to the cost of labor. It is thus a high-risk strategy to assess the elasticity of labor demand on the basis of variations in the level of employment consequent upon changes in the minimum wage. Another strategy is to estimate an inverse labor demand, in other words, wages as a function of the level of employment, when there occur exogenous variations in labor supply, such as a large influx of immigrants (Angrist, 1996) or a massive entry of women onto the labor market (Acemoglu et al., 2004). This strategy also relies on a model of labor market equilibrium in which supply and demand interact. This model will be studied in chapter 3.

With these preliminary points stipulated, it remains the case that empirical studies of labor demand do arrive at some convergent results. It emerges, among other things, that the elasticity (conditional and unconditional) of labor demand with respect to the cost of this factor is negative. It has also been found that unskilled labor is more easily substitutable for capital than skilled labor and that skilled labor and capital may even be p-complements.

2.2.1 Aggregate Labor Demand

The estimate most frequently made is that of the conditional elasticity $\bar{\eta}_W^L$ of aggregate labor demand. It is effected by positing that the labor factor L is a *homogeneous* quantity equal to the sum of hours worked, or the level of employment. The cost of labor W is most often assimilated to the total amount of wages divided by the number of workers or by their hours. In reality, the definition of W raises numerous problems, for variations in the total amount of wages may correspond to deformations in the structure of employment arising, for example, from different levels of seniority or skill. We have also seen in the preceding section that the distinction between fixed costs and variable costs plays an important role when firms have to choose between the number of workers and the number of hours worked.

These difficulties notwithstanding, studies devoted to estimating $\bar{\eta}_W^L$ yield converging results, whatever the level (firm, sector, or nation) at which the data are collected. They show that the elasticity of conditional demand for labor with respect to the cost of this factor is *negative* and, in absolute value, *inferior to 1*. Hamermesh (1993), building on more than 70 studies, takes the view that the most probable interval for $\left|\bar{\eta}_W^L\right|$ is $[0.15 - 0.75]$. If a single figure were to be chosen, 0.30 would surely be the best estimate. Knowledge of $\bar{\eta}_W^L$ allows us to deduce the value of the elasticity of substitution σ between capital and labor, since, according to (2.12), we know that these two quantities are linked by relation $\bar{\eta}_W^L = -(1 - s)\sigma$, where s represents the share of labor in total cost. Overall, s is close to 0.7. With $\bar{\eta}_W^L = -0.3$, we arrive at $\sigma = 1$. In other words, the use of a global Cobb-Douglas production function, or $Y = K^{\theta(1-\alpha)}L^{\theta\alpha}$ with $\alpha = 0.7$, is not without empirical relevance when we are considering only two inputs.

Taking scale effects into account increases the absolute value of the elasticity of employment to its cost, which conforms to theoretical results. Works dedicated to estimating the elasticity η_W^L of the unconditional labor demand are less numerous, and show wider divergence, than those dedicated to estimating $\bar{\eta}_W^L$. Still, on the basis of macroeconomic data, η_W^L is negative and Hamermesh estimates that its absolute value lies[10] on average at around 1. If we assign a value of 0.3 to $\bar{\eta}_W^L$, it becomes evident that the extent of the scale effect is far from negligible.

2.2.2 Complementarity and Substitution Between Inputs

The degree to which one input is capable of replacing another in the production process has an important place in the assessment of the effects of economic policy. Several major results stand out.

If we take labor services into account with the help of a *sole* aggregate, the latter is, as a general rule, a p-substitute with any other aggregate input. Hence labor is

[10] More precisely, in this case we estimate $\eta_{\frac{L}{W}}$ defined by relation (2.20).

p-substitute with capital, energy, and raw materials, which, as readers will recall, means that the *conditional* labor demand rises with the cost of these three inputs. This result is somewhat surprising if one thinks back, for example, to the effects the hikes in the cost of oil in the 1970s had on the level of employment. But it should be remembered that such hikes are accompanied by a scale effect, in other words, by reduced production, which can lead in the end to reduced employment. In other words, labor and energy are p-substitutes but are probably not gross substitutes.

Unskilled labor is easier to substitute for capital than skilled labor. There are even good reasons to think that at the overall level, or even at the level of one of the large sectors of the economy, skilled labor and capital are p-complements (for a far-reaching review of the literature, see Hamermesh, 1993, chapter 3). These results are confirmed by the fact that direct elasticity of the conditional labor demand, for a given category of manpower, diminishes in absolute value with the level of education in this category. Likewise, this elasticity diminishes, still in absolute value, with the level of skill. The results are evidently sensitive to the manner in which the breakdown between skilled and unskilled labor is carried out. In the United States, one breakdown views unskilled workers as those with a high school diploma at most and skilled workers as all those with a higher qualification than that; the authors come to an estimate of the elasticity of substitution between skilled and unskilled labor lying between 1 and 2 (see Johnson, 1997, and Autor et al., 1998, for whom this elasticity of substitution lies rather between 1.4 and 1.5). In his study of the Israeli labor market, Angrist (1996) finds that the elasticity of unskilled labor is equal to 3.

Results concerning the substitution between workers and hours do not yet display a real consensus. In large measure, the lack of precision comes from the difficulty of attributing different costs to workers and hours, that is, of assessing what share to assign to variable costs, and what share to fixed costs. According to Hamermesh (1993), the only property firmly established is that workers and hours are both p-substitutes for capital. With a reasonable degree of confidence, we may likewise assume that workers and hours are p-substitutes. For example, most studies show that the employment level rises unambiguously when the cost of overtime hours rises, a conclusion which also conforms to the theoretical analysis presented in section 1.5.3. We note further that Leslie and Wise (1980) and Hart and MacGregor (1988) give estimates of the elasticity of production with respect to hours equal to 0.64 and 0.87 respectively. Using French data Gianella and Lagarde (1999) arrive at a figure equal to 0.9, whatever the size of the firm examined. As Hart and MacGregor (1988) emphasize, however, these results are fragile, and their summary of the empirical literature leads them to conclude that the elasticity of production with respect to hours is close to (in fact not significantly different from) unity.

3 DYNAMIC LABOR DEMAND

The static theory of labor demand furnishes valuable indications about what determines elasticities, and about the possibilities of substitution over the long run between the different inputs. About the manner in which the inputs reach their long-run values, however, and the length of time that these adjustments take, it gives no firm detail.

Moreover, it does not take into account the fact that firms are faced with an ongoing process of reorganization arising from technological constraints, market fluctuations, and manpower mobility.

In order to be able to assess these phenomena, we have to resort to the notion of adjustment cost. Firms may incur adjustment costs when they decide to change their level of employment. But the fact that firms must deal with quits by workers entails that they may incur adjustment costs simply in order to maintain a constant level of employment, since they may have to hire to adjust employment to the desired level.

The functional form chosen to describe adjustment costs conditions the dynamics of labor demand and the properties of stationary solutions. That is why we look at different functional forms in this section—so as to take into account different types of adjustment cost. Additionally, we examine the dynamics of labor demand in a setting without uncertainty, then introduce stochastic elements into the models. The setting without uncertainty serves as a baseline and allows us to grasp the principal mechanisms at work. The models set in stochastic environments bring out the role of expectations in labor adjustment.

3.1 The Costs of Labor Adjustment

Adjusting the size of the workforce entails costs. Numerous studies show that the size of these costs is far from insignificant, and for that reason they play a large role in decisions to hire and fire. No real consensus has yet been reached as regards the analytical representation of these costs, but the quadratic symmetric form, historically the one used most frequently, is gradually being abandoned.

3.1.1 Definition and Size of Adjustment Costs

Adjustment costs are evaluated on the basis of several sources. Some studies give estimates of the difference between optimal employment, that is, what the firm would choose if adjustment costs were absent and the level of employment actually observed. Others supply indications of what the costs of hiring and firing workers amount to. Yet others attempt to assess the effects of employment protection, which plays an important role in many OECD countries.

A Typology

Labor adjustment costs arise from variations in the volume of employment and from the replacement of former employees by new ones. When the work process is reorganized, causing temporary loss of efficiency, we say that the firm is undergoing internal adjustment costs. Examples might be the adaptation of the workforce to new machinery, or the settling-in period for new workers. Costs like this are difficult to evaluate, because they do not show up as distinct items in the firm's accounts. But when the reorganization is accompanied by costs that can be distinguished from variations in production, for example, if a change in the work routine requires the advice of experts who charge a fee for their services, or severance pay for workers who are fired, we say that the firm is dealing with external adjustment costs.

It is important to distinguish gross costs, which are caused by gross changes in employment (the sum of all who join or leave the workforce), from net costs, which flow from net changes in employment (the difference between the joiners and the leavers).

The existence of gross costs highlights the possibility of there being turnover costs, even when the size of the firm's workforce remains constant. These costs are due to the operations of hiring and firing and to voluntary departures. The rate of rotation of a workforce is defined as the average of the number of entries into and exits out of employment divided by the stock of jobs at the outset of the period. Even when a firm's level of employment remains constant, it generally has a positive rate of workforce rotation. This positive rate is the consequence of voluntary quits by workers and reorganizations that may require the firm to renew its mix of workers. This rotation is not negligible: in each quarter, firms with unchanging levels of employment separate from around 10% of their manpower and hire an equivalent number, for a quarterly rate of rotation of 10% (see figures 2.2 and 2.1, which concern France and the United States).

Evaluating Hiring and Separation Costs

Studies carried out by recruiting agencies and human resources departments indicate that in the United States the replacement cost of a worker who quits a firm oscillate between 25% of the annual wage for less qualified wage earners and more than 100% for those more qualified (Nase, 2009). As a general rule, studies carried out on American data come to the conclusion that hiring costs are much greater than separation costs (Hamermesh, 1993).

Studies based on French data also show that the adjustment costs of employment are substantial. Abowd and Kramarz (2003), utilizing a representative sample of French firms and their employees, show that in France the costs of hiring are due solely to the hiring of skilled workers on long-term contracts and are clearly less than the costs of separation. The average cost of a separation represents 56% of the annual cost of labor, whereas a hire (not including training costs) represents only 3.3% of the same cost. The cost of a separation itself depends heavily on the context. Rigorous employment protection in France means that to let an employee go for economic reasons brings a cost equivalent to 126% of the annual cost of his labor. Goux et al. (2001) come to conclusions of the same order using longitudinal data on 1,000 French firms followed from 1988 to 1992. They estimate that for long-term contracts, the cost of hiring represents no more than 2.5% of the cost of separation.

Although measurements of hiring costs are not always homogeneous (for example, studies on French data leave training costs out of account), what emerges from all these studies is that in the United States, the costs of hiring are high and outstrip the costs of separation, while in countries where strong legal measures are in place to enhance job security, the costs of separation far outstrip recruitment costs.

Employment Protection Measures

The usual view is that the higher the cost of a firing, the stronger employment protection is. International comparisons try to rank job security norms by how strict they are (see Venn, 2009, and chapter 13). The wide range of criteria utilized shows us at a glance how complex this exercise is and how difficult it is to evaluate precisely the effective cost of job protection. These criteria concern things like the possibility of using contracts of limited duration and the services of agencies supplying temporary labor, how long a period of trial employment can last, the administrative procedure to follow when terminating employment (notification, summons, authorization from a public agency), the amount of advance notice and severance pay applicable to different types

of termination (firing for cause, firing for economic reasons, etc.), and the definition of wrongful termination and the possibility that a person wrongfully terminated can get his or her job back. These criteria identify the strictness imposed by written rules but leave aside their application and most of the case law.

Most assessments conclude that employment protection is less strict in the United States, Canada, and the United Kingdom than in France, Germany, and the countries of southern Europe. Japan occupies an intermediate position (for more details, see chapter 13). In Europe a large part of the cost of termination is regulatory in nature (period of advance notice, administrative procedure, etc.). The result, since the beginning of the 1980s, has been a massive recourse to short-term contracts precisely to avoid these administrative costs.

3.1.2 The Specification of Adjustment Costs

For ease of analysis, adjustment costs have most often been represented using a convex symmetric function (in general quadratic) of net employment changes. But this way of specifying them does not allow us to explain asymmetric and discontinuous adjustments in employment and the consequence of gross employment changes. For this reason, it is now gradually being replaced by a representation including fixed costs, linear costs, quadratic costs, and gross employment changes.

Quadratic Costs

The first analyses of the decisions made by a firm facing adjustment costs adopted a quadratic relation between the variations (gross or net) in employment and adjustment costs. This representation was introduced by Holt et al. (1960), who viewed net adjustment costs as equal to $b(\Delta L_t - a)^2$, $a, b > 0$, with $\Delta L_t = L_t - L_{t-1}$ or $\Delta L_t = \dot{L}_t$ according to whether time was represented discretely or continuously. This specification has the advantage of introducing an asymmetry between the costs of positive and negative variations in employment ($a > 0$). But this asymmetry has a drawback: cost is strictly positive in the absence of any variation in employment. Eisner and Strotz (1963) got around this problem by assuming quadratic and symmetric adjustment costs ($a = 0$). A hypothesis of this kind allows us to obtain simple analytic results, which is why it was adopted in numerous studies. It proves vulnerable to criticism, however, on two points. First, it does not allow us to distinguish costs arising from recruitment from those arising from departure; but the numerous studies referred to above show that these costs differ in amount and effect. Second, it implies that there is a *gradual* adjustment of employment since the marginal cost of adjustment rises with a change in the level of employment. This property gives firms an incentive not to vary their labor demand too much at each period so as to minimize adjustment costs. So the quadratic form does not allow us to explain the sudden adjustments in employment often observed in real life.

Asymmetric Convex Costs

For the reasons just mentioned, more recent studies postulate asymmetric adjustment costs. Pfann and Palm (1993) assume a relation of this form:

$$C(\Delta L) = -1 + \exp(a\Delta L) - a\Delta L + \frac{b}{2}(\Delta L)^2, \qquad a > 0,\ b > 0$$

This specification implies an asymmetry between positive and negative variations in employment. We return to the symmetrical formulation with $a = 0$. Conversely, when $a > 0$ (or $a < 0$), the marginal cost of an increase in employment is greater (or less) than that of a reduction. The asymmetry may also originate in a function that is not continuously differentiable. For example, Chang and Stefanou (1988) and Jaramillo et al. (1993) adopt the following specification:

$$C(\Delta L) = c_h(\Delta L)^2 \quad \text{if } \Delta L \geq 0 \quad \text{and} \quad C(\Delta L) = c_f(\Delta L)^2 \quad \text{if } \Delta L \leq 0, \quad c_h > 0, c_f > 0$$

Linear Costs

The specification of adjustment costs in the form of a piecewise linear function offers the advantage of achieving a more realistic representation of labor demand, in which firms hire in some circumstances, let employees go in others, and sometimes leave their workforce unchanged (see section 3.2.2). The use of piecewise linear costs expanded greatly in the 1990s, with the works of Bentolila and Bertola (1990), Bertola (1990), Bentolila and Saint-Paul (1994), and Bertola and Rogerson (1997), who examine linear adjustment costs of the form:

$$C(\Delta L) = c_h \Delta L \quad \text{if } \Delta L \geq 0 \quad \text{and} \quad C(\Delta L) = -c_f \Delta L \quad \text{if } \Delta L \leq 0, \quad c_h > 0, c_f > 0$$

The coefficients c_h and c_f represent the respective unit costs of a hiring and a termination. The adjustment of employment is asymmetric, since $c_h \neq c_f$.

Lump-Sum Costs

In many circumstances, the adjustment costs of employment include a component that is fixed and therefore not directly linked to the size of the adjustment. For example, the costs of searching for certain categories of personnel, or the administrative costs incurred in a mass termination are in large part independent of the number of individuals involved in these operations. Hamermesh (1989, 1993, 1995) adopts the hypothesis of a discontinuity in adjustment costs when he postulates that firms undergo a strictly positive fixed cost when $\Delta L \neq 0$, but that they are not subject to any cost if $\Delta L = 0$. Abowd and Kramarz (2003) consider different fixed costs for hirings and terminations. The existence of lump-sum costs allows us to explain why firms of a certain size sometimes have an interest in doing their hirings, and their terminations, in groups.

Empirical studies have sought to discover which representation fits best. With that in mind, certain studies posit functions of adjustment costs that comprise fixed costs, linear costs, and quadratic costs (Hamermesh, 1992; Nielsen et al., 2007). These studies are all the more necessary in that the analysis of the determinants of labor demand dynamics proves particularly sensitive to the specification of adjustment costs.

3.2 The Adjustment of Employment in a Deterministic Environment

We here consider a firm situated in a deterministic environment, which must support adjustment costs when it alters its workforce. To make things easier from a technical point of view, a large part of the literature has assumed that these costs were symmetric

and could be represented by a quadratic function. We begin by studying this case, which always serves as a baseline in this domain. But criticisms directed at the hypothesis of quadratic and symmetric costs, and outlined above, have led to the use of asymmetric functional forms, the linear one being chosen most often.

3.2.1 Quadratic and Symmetric Adjustment Costs

The use of quadratic costs presents the advantage of leading to a very simple dynamic representation of the trajectory of employment, in which employment gradually returns to its stationary value.

The Behavior of the Firm

We will work with a dynamic model in continuous time, in which, at each date, $t \geq 0$, the adjustment cost is restricted to labor alone. When the firm utilizes a quantity L_t of this factor, it obtains a level of output $F(L_t)$ that is strictly increasing and concave with respect to L_t. Taking other inputs into account, such as capital, greatly complicates the analysis without changing the import of the results that we want to highlight. We likewise simplify by leaving quits out of the reckoning, on the assumption that net variations in employment are equal to gross variations. Introducing quits modifies the cost of labor in the stationary state, for it must now incorporate not only the wage but also the turnover cost. Any rise in turnover costs induces an increase in the cost of the workforce leading to a diminution of employment, as we will see in further detail in chapter 13. Such a modification aside, introducing quits into the reckoning would not change the main dynamic properties of the model substantially.

Let $\dot{L}_t$ be the derivative with respect to t of the variable L_t; we will assume that variations in the level of employment are accompanied at every date t by an adjustment cost represented by the quadratic function $(b/2)\dot{L}_t^2$, $b \geq 0$.

To simplify the notations and calculations, from now on we will omit the index t and assume that at every date the cost of labor and the interest rate are exogenous constants denoted respectively by W and r. At date $t = 0$, the discounted present value of profit, Π_0, is written:

$$\Pi_0 = \int_0^{+\infty} \left[F(L) - WL - \frac{b}{2}\dot{L}^2 \right] e^{-rt} \mathrm{d}t$$

In this environment, free of random factors, the firm chooses its present and future levels of employment so as to maximize the discounted present value of profits Π_0. This is a classic problem of calculus of variations for which the first-order condition is given by the Euler equation:[11]

$$\frac{\partial J}{\partial L} = \frac{\partial}{\partial t}\left(\frac{\partial J}{\partial \dot{L}}\right) \quad \text{with} \quad J(L, \dot{L}, t) = \left[F(L) - WL - \frac{b}{2}\dot{L}^2 \right] e^{-rt} \tag{2.44}$$

[11] See Takayama (1986, chapter 5) and the mathematical appendix B at the end of this book. The Euler condition is also sufficient if function J is concave in L and $\dot{L}$, which is the case here.

After several simple calculations, we find that the employment path is described by a nonlinear second-order differential equation that takes the form:

$$b\ddot{L} - rb\dot{L} + F'(L) - W = 0 \tag{2.45}$$

The stationary value L^* of employment is obtained by making $\dot{L} = \ddot{L} = 0$ in this equation. It is thus defined by the usual equality between marginal productivity and wage, or $F'(L^*) = W$. In this simple model, the stationary level of employment does not depend on parameter b measuring the extent of adjustment costs, for $\dot{L} = 0$ in the stationary state, and there is no flow of hirings or terminations to give rise to costs of this type. This would no longer be the case if, for example, the stationary state were characterized by a permanent flow of hirings compensating for exogenous departures. On the other hand, the employment path described by differential equation (2.45) always depends on parameter b measuring the size of adjustment costs.

The Dynamics of Employment

It is possible to specify precisely the properties of the trajectory of employment in the neighborhood of the stationary state by taking the first-order approximation of $F'(L)$ around L^*. Replacing $F'(L)$ by $F'(L^*) + (L - L^*)F''(L^*)$ in equation (2.45), we arrive at:

$$b\ddot{L} - rb\dot{L} - aL = -aL^*, \quad \text{with} \quad a = -F''(L^*) > 0$$

Let A_1 and A_2 be two arbitrary constants. The general solution of this linear second-order differential equation is written:[12]

$$L = L^* + A_1 e^{\lambda_1 t} + A_2 e^{\lambda_2 t} \tag{2.46}$$

with

$$\lambda_1 = \frac{1}{2}\left[r + \sqrt{r^2 + \frac{4a}{b}}\right] > 0 \quad \text{and} \quad \lambda_2 = \frac{1}{2}\left[r - \sqrt{r^2 + \frac{4a}{b}}\right] < 0 \tag{2.47}$$

The coefficient λ_1 being positive, it is necessary that A_1 be equal to 0 in order to have a stable path. Let L_0 be the (given) level of employment at date $t = 0$; the value of A_2 is found by making $t = 0$ in (2.46), which gives $A_2 = L_0 - L^*$. The employment trajectory is thus completely defined by:

$$L = L^* + (L_0 - L^*)e^{\lambda_2 t} \tag{2.48}$$

[12] Readers are reminded that the solution of a linear second-order differential equation $af''(t) + bf'(t) + cf(t) = d$, where a, b, c, d are given constants, is found by first calculating the solution of the homogeneous equation $af'' + bf' + cf = 0$. This solution is of the form $f(t) = A_1 e^{\lambda_1 t} + A_2 e^{\lambda_2 t}$, where A_1 and A_2 are arbitrary constants and λ_1 and λ_2 are the roots of the "characteristic" equation $a\lambda^2 + b\lambda + c = 0$. We then calculate the solution of the non-homogeneous equation, which is equal to the sum of the solution of the homogeneous equation and a particular solution of the non-homogeneous equation. Here a particular solution is d/c. So the general solution is of the form $f(t) = A_1 e^{\lambda_1 t} + A_2 e^{\lambda_2 t} + (d/c)$. In the end we get a particular solution on the basis of a known value, generally the initial or terminal value of $f(t)$. The constants A_1 and A_2 are determined by the initial conditions and the stability conditions.

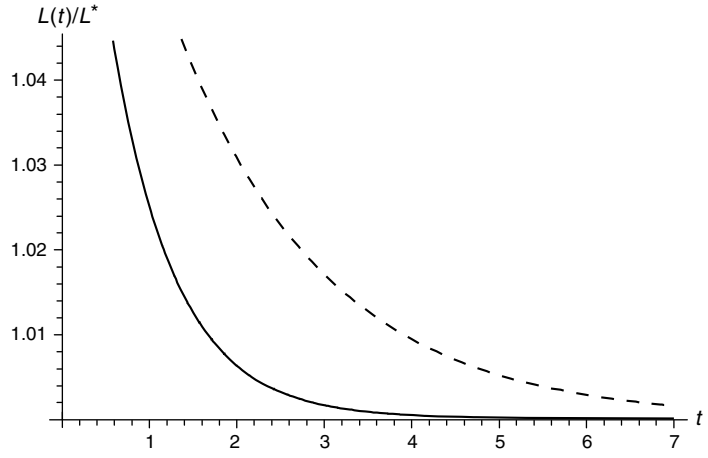


Figure 2.5
Employment adjustment in the model with quadratic adjustment costs. The initial level of employment is 10% greater than its stationary value. The broken line corresponds to an adjustment cost b equal to 80% of the annual labor cost, and the unbroken line to a value of b equal to 15% of the same annual cost.

This equality shows that employment *gradually* moves to its stationary value L^*. This property is the direct consequence of the utilization of a quadratic form to represent adjustment costs. With this specification, the firm has an interest in "smoothing out" the changes it makes to its workforce, for if the adjustment were to be made all at once at the initial date, the instantaneous cost of the hirings and terminations, or $b(L_0 - L^*)^2$, would exceed the total cost of an adjustment spread out over time.

Figure 2.5 gives an illustration of the adjustment trajectories, assuming a homogeneous production function $F(L) = L^{0.7}$, a labor cost W equal to 0.7, and an annual interest rate $r = 0.05$. We thus have $L^* = 1$. We assume that the initial level of employment is 10% greater than its stationary value. As well, we distinguish two kinds of job: skilled jobs (the broken line), for which the cost of adjustment is 80% of the annual labor cost, and unskilled jobs (the unbroken line), for which the cost of adjustment is 15% of the annual cost. We may note that the trajectory of unskilled jobs approaches the stationary value more rapidly than that of skilled jobs, for which the costs of adjustment are greater. In this regard, a graphic representation is particularly useful because it allows us to visualize the amounts of time that adjustments take. But it is also useful to have a measure of the adjustment speed.

Median Lag and the Adjustment Speed

The time required for employment adjustment is conventionally measured by a *median lag* which, by definition, indicates the time required for the level of employment to settle at a point equidistant from its initial value L_0 and its stationary value L^*. Consequently, the median lag, denoted δ, is implicitly defined by the equality $L_\delta = (L_0 + L^*)/2$. Therefore, taking into account equation (2.48), which describes the employment trajectory, the median lag is defined by the formula $\delta = -\ln 2/\lambda_2$. Given the expression of λ_2 that appears in relation (2.47), we see that the median lag increases with b. Hence a rise in adjustment costs prolongs the time that employment adjustment takes.

Staying with a quadratic function means that employment adjustment takes place gradually. For one thing, this does not always correspond to observed facts. For another, the hypothesis of symmetry prevents us from distinguishing between effects arising from the costs of terminating employment and those arising from the costs of hiring. In what follows, we examine the consequences of the asymmetry between these two types of cost with a linear adjustment costs function.

3.2.2 Linear and Asymmetric Adjustment Costs

It is possible to distinguish the costs of hiring and firing by adopting a piecewise linear specification. The hypothesis of linearity also brings out the fact that, contrary to the model with quadratic costs, employment adjustment can take place immediately.

The Demand for Workers

Let c_h and c_f be two positive constants, and let us assume from now on that the adjustment costs are represented by the function:

$$C(\dot{L}) = c\dot{L} \quad \text{with} \quad c = c_h \quad \text{if } \dot{L} > 0 \quad \text{and} \quad c = -c_f \quad \text{if } \dot{L} < 0 \tag{2.49}$$

Parameters c_h and c_f allow us to distinguish hiring costs ($\dot{L} > 0$) from termination costs ($\dot{L} < 0$). As in the previous model with quadratic adjustment costs, it is assumed, for the sake of simplicity, that there are no quits so that the net employment changes are equal to gross employment changes. The firm's problem consists of choosing, at date $t = 0$, levels of employment that maximize the discounted present value of profit Π_0. The latter is expressed thus:

$$\Pi_0 = \int_0^{+\infty} \left[F(L) - WL - C(\dot{L})\right] e^{-rt} \mathrm{d}t$$

Once again, this is a problem of calculus of variations to which the Euler equation (2.44) applies when the quadratic function $-(b/2)\dot{L}^2$ is replaced by the linear function $C(\dot{L}) = c\dot{L}$. After several simple calculations, we find that the employment path is defined by the equation $F'(L) = W + rc$, which entails:

$$F'(L) = W + rc_h \quad \text{if } \dot{L} > 0 \quad \text{and} \quad F'(L) = W - rc_f \quad \text{if } \dot{L} < 0$$

These conditions signify that the firm hires when marginal productivity is sufficiently high to cover the wage W and the hiring cost rc_h. Conversely, the firm fires when productivity is so low that it just equals wage W less the provision rc_f for the termination cost. In all other cases, that is, when productivity lies in the interval $[W - rc_f, W + rc_h]$, the firm has no interest in altering the size of its workforce since the gains due to hiring and firing are less than the costs incurred by adjusting employment.

Labor adjustments take a particularly instructive form when the parameters W, r, c_h, and c_f are constants, which we have assumed. Let us define the employment levels L_h and L_f by the equalities:

$$F'(L_h) = W + rc_h \quad \text{and} \quad F'(L_f) = W - rc_f \tag{2.50}$$

We see that the optimal values L_h and L_f do not depend on date t. That means that labor demand *immediately* (i.e., in $t = 0$) "jumps" to its stationary value. The firm adjusts its workforce to the value L_h (or L_f) if the latter is superior (or inferior) to the initial value L_0 of employment. In the opposite case, that is, if L_0 falls in the interval $[L_h, L_f]$, the optimal solution for the firm consists of making no change to the size of its workforce. In sum, labor demand is defined by:

$$L = \begin{cases} L_h & \text{if } L_0 \leq L_h \\ L_0 & \text{if } L_h \leq L_0 \leq L_f \\ L_f & \text{if } L_f \leq L_0 \end{cases} \tag{2.51}$$

The result—that the level of employment immediately jumps to its stationary value—arises from our choice of a linear form to represent adjustment costs. In this case, it is not necessary to smooth out the trajectory so as to reduce costs. The firm always has an interest in reaching the stationary state as quickly as possible. Consequently, the use of a linear form allows us to account for brutally rapid changes in the employment level. It can be remarked that lumpy adjustment costs would yield a similar employment path.

The Effects of Hiring and Firing Costs

The choice of optimal employment is represented in figure 2.6. The upper part of the graph represents the marginal productivity of the initial level of employment $F'(L_0)$. The boldface curve in the lower part of the graph represents the relationship between the initial level and the optimal level of employment chosen by the firm. We see that if the marginal productivity of initial employment is superior to $W + rc_h$ the firm hires, whereas it fires if the marginal productivity of initial employment is inferior to $W - rc_f$. In all cases lying in between, the firm does not alter its employment level.

Figure 2.6 and relations (2.50) also show that the costs of hiring and firing have *opposing* effects on labor demand. If the size of the workforce is low at the outset ($L_0 \leq L_h$), then optimal employment is equal to L_h and a rise in the hiring cost c_h reduces employment. Conversely, if there is a large number of workers at the outset ($L_f \leq L_0$), the optimal level of employment takes the value L_f and we clearly see that a *rise* in the termination cost c_f has the effect of *increasing* employment. We should not, however, conclude on the basis of this analysis that a rise in the termination cost (or a fall in the hiring cost) "augments" the firm's labor demand. In reality, since this demand immediately jumps to L_h or L_f (unless it simply remains at L_0), the level of employment is always equal to one of the three quantities L_h, L_f, or L_0. Let us suppose that the number of workers is L_f, a rise in the termination cost c_f will augment L_f up to a certain value L_f^+ and will thus have the effect of placing the *outset* level of the workforce (now equal to L_f) somewhere in the interval $[L_h, L_f^+]$. In this case, relation (2.51) describing labor demand shows that the firm then has an interest in remaining at L_f. In other words, a rise in the cost of terminating *hinders* the firm from going ahead with reductions in personnel but gives it no incentive to hire. An analogous line of reasoning would show that a rise in the costs of hiring has the effect of *discouraging* further recruitment but does not lead to a reduction in employment. Conversely, a reduction in hiring costs always has a positive effect on employment to the extent that it increases the value L_h of optimal employment.

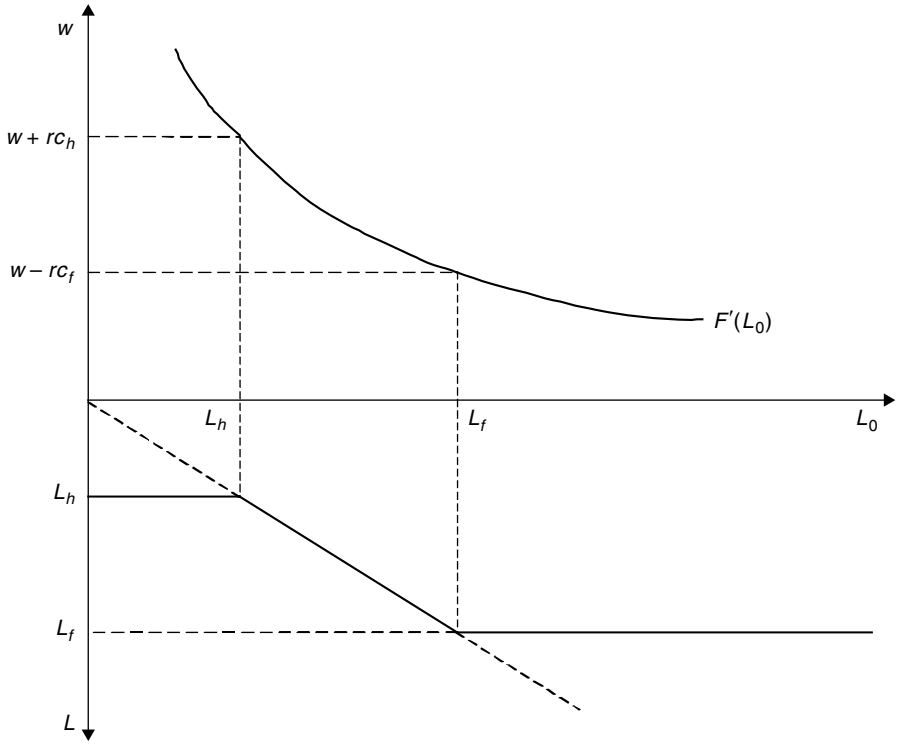


FIGURE 2.6
Labor demand in the model with linear adjustment costs of employment.

It emerges from this analysis that a rise in the termination cost of employment leads to a stabilization of labor demand when the latter is high ($L_t = L_f$) and that a fall in the hiring cost has the effect of increasing labor demand when it is low ($L_t = L_h$). Conclusions of this nature cannot be reached with a quadratic representation of adjustment costs. Moreover, the model just presented suggests that appropriate management of hiring and firing costs may play a stabilizing role vis-à-vis labor demand. This result must however be reexamined in an environment with uncertainty.

3.3 THE ADJUSTMENT OF LABOR DEMAND IN A STOCHASTIC ENVIRONMENT

In order to compare the results to follow with those already obtained in the absence of uncertainty, we will take a stochastic environment and will examine the consequences of representing adjustment costs by a quadratic and symmetric function, and then by a linear asymmetric function.

3.3.1 QUADRATIC AND SYMMETRIC ADJUSTMENT COSTS

The quadratic case serves as our baseline. Under the hypothesis of rational expectations, the dynamic path of employment is described by a linear equation with lag that fits

estimates well. This representation extends to multiple inputs and allows us to define the notion of dynamic complementarity and substitution.

The Firm's Problem and the Euler Equation

Models in continuous time are not very well adapted to understanding the formation of expectation. For this reason we (provisionally) abandon the continuous representation of time in favor of a model in discrete time. The output of the firm is now written $F(A_t, L_t)$ where $A_t > 0$ is a random variable representing, for example, a shock to the selling price or to productivity, falling at the beginning of period t. We assume that A_t has been realized and observed prior to the decisions to hire and fire made in period t. Production is always strictly increasing and concave with respect to employment L_t. In the course of period t, the firm supports adjustment costs arising from manpower turnover represented by a quadratic and symmetric function, which is expressed as $(b/2)(L_t - L_{t-1})^2$, $b > 0$. The firm's behavior is analyzed following the same procedure followed in the dynamic model with certainty. The problem of maximizing the expected profit yields optimality conditions that allow us to know the labor demand at each date. These conditions are generally equations in differences defining current employment, hirings, and firings as a function of past employment and expected future employment. At each date t, the expected discounted present value of profit is written:

$$\Pi_t = \mathbb{E}_t \left\{ \sum_{i=0}^{+\infty} \left(\frac{1}{1+r} \right)^i \left[F(A_{t+i}, L_{t+i}) - W_{t+i} L_{t+i} - \frac{b}{2} (L_{t+i} - L_{t+i-1})^2 \right] \right\}$$

In this expression, $\mathbb{E}_t$ designates the expectation operator conditional upon all the information available to the employer at date t. The strict concavity of the production function and the convexity of the adjustment costs imply that the first-order condition defines a maximum. Differentiating the expression of expected discounted present value of profit with respect to L_t, we obtain the Euler equation, or:

$$F_L(A_t, L_t) = W_t + b(L_t - L_{t-1}) - \frac{b}{1+r} \mathbb{E}_t \left(L_{t+1} - L_t \right), \quad \forall t \geq 1 \tag{2.52}$$

The dynamics of employment is thus described by a second-order difference equation where current employment L_t depends both on past employment L_{t-1} and on expected employment $\mathbb{E}_t L_{t+1}$.

The Formation of Expectations

At this point it is necessary to spell out the process by which expectations are formed. We assume here that the producer is capable of forming *rational* expectations. This signifies that the expectation formed at date t about the value X_{t+i} of a variable X at date $t + i$ is then equal to the mathematical expectation of X_{t+i} conditional upon all the information available at date t. This expectation is denoted $\mathbb{E}_t X_{t+i}$ (for more detail on expectations in general, see chapter 8, section 3). Under the hypothesis of rational expectations, the "true" model of the economy is one of the available pieces of information. In particular, the employer knows that the future level of employment L_{t+1} is given by relation (2.52) applied to date $(t + 1)$. Step by step, we thus see that employment L_t

will be a function of past employment L_{t-1} and of expectations formed at date t about all future shocks A_{t+i}, $i \geq 1$. In order to obtain an explicit solution for labor demand, we assume that the production function can be approximated by a linear quadratic function of the form $F(A_t, L_t) = A_t L_t - (B/2)L_t^2$, with $B > 0$. Equation (2.52) is then written:

$$a_0 \mathbb{E}_t L_{t+1} - L_t + a_1 L_{t-1} + a_t = 0 \tag{2.53}$$

with

$$a_0 = \frac{b}{(B+b)(1+r)+b}, \quad a_1 = (1+r)a_0 \quad \text{and} \quad a_t = \frac{(A_t - W_t)\, a_1}{b}$$

The Dynamics of Employment

The solution of equation (2.53) can be obtained thanks to the "indeterminate coefficients" method (see Blanchard and Fischer, 1989, p. 261; Sargent, 1986, chapter 14). It consists of postulating a particular form for the solution, then identifying the unknown parameters by writing that this particular form satisfies equation (2.53). Since L_t depends on its past value L_{t-1}, on the present realization of the random variable a_t, and on all the future expected values of the latter, we can seek a solution that is a linear form of these quantities. In this case, L_t is written:

$$L_t = \lambda L_{t-1} + \sum_{i=0}^{+\infty} \mu_i \mathbb{E}_t a_{t+i} \tag{2.54}$$

In this expression, λ and μ_i $(i \geq 0)$ are unknown parameters that have to be determined. To do so, we begin by calculating the expectation at date t of L_{t+1} resulting from relation (2.54). We thus find that $\mathbb{E}_t L_{t+1} = \lambda L_t + \sum_{i=0}^{+\infty} \mu_i \mathbb{E}_t a_{t+i+1}$, and in substituting this expression of $E_t L_{t+1}$ in (2.53), we finally get:

$$L_t = \frac{1}{1 - \lambda a_0} \left(a_0 \sum_{i=0}^{+\infty} \mu_i \mathbb{E}_t a_{t+i+1} + a_1 L_{t-1} + a_t \right) \tag{2.55}$$

It suffices now to identify the coefficients of L_{t-1} and of a_{t+i} $(\forall i \geq 0)$ in the expressions of L_t given by (2.54) and (2.55) in order to obtain the values of the coefficients λ and μ_i. These are:

$$\lambda = \frac{a_1}{1 - \lambda a_0}, \quad \mu_0 = \frac{1}{1 - \lambda a_0} \quad \text{and} \quad \mu_i = (a_0 \mu_0)^i \mu_0, \quad \forall i \geq 1$$

Coefficient λ linked to lagged employment L_{t-1} is the root of the polynomial $a_0 \lambda^2 - \lambda + a_1$. We verify that this polynomial admits two real positive roots, one inferior to one and the other superior to one. Only the value of λ inferior to unity gives a stable nonexplosive solution and is thus the only root that can be retained. Substituting the values of the coefficients calculated above in equation (2.54), we arrive at the definitive expression of the solution, or:

$$L_t = \lambda L_{t-1} + \mu_0 \sum_{i=0}^{+\infty} (a_0 \mu_0)^i \mathbb{E}_t a_{t+i} \quad \text{with} \quad \lambda = \frac{1 - \sqrt{1 - 4a_0 a_1}}{2a_0} \tag{2.56}$$

It is easy to verify that the value of λ linked to lagged employment increases with parameter b, which measures the extent of the adjustment costs. The weight of past employment is thus more important, the higher adjustment costs are. In other words, fluctuations in labor demand are less marked when adjustment costs are large. Equations analogous in form to (2.56) have served as the foundation of numerous empirical estimates that attempt to measure the speed of employment adjustment. For that, we have to postulate a particular form for the stochastic process governing the path of the random variables a_t and, if possible, to link the parameters of this process to certain observable variables (see section 3.4.1 for an example).

Dynamic Substitution and Complementarity

Only adjustment costs linked to employment, assumed to be homogeneous, have been taken into consideration. But the firm incurs this type of cost for other inputs, notably capital. Lucas (1967) and Nadiri and Rosen (1973) have studied the case of quadratic adjustment costs with multiple inputs in a stationary environment. Messe (1980) has extended this study in a stochastic framework with rational expectations. The dynamics of employment is then described by an equation the form of which is very close to (2.56). To be precise, if there are n inputs, of which the i^{th} is utilized in quantity X_t^i at date t, the equation of the path of this input is written:

$$X_t^i = \sum_{j=1}^{n} \lambda_{ij} X_{t-1}^j + \sum_{k=0}^{+\infty} \gamma_i^k \mathbb{E}_t x_{t+k}$$

In this expression, λ_{ij} are adjustment parameters, γ_i is a vector of parameters dependent on technology and adjustment costs, and x_t represents a vector relative to the price of the inputs. It is evident that the quantity of input i utilized at date t depends on the past quantities of all the inputs that give rise to adjustment costs. By extension of the definitions we gave in section 1.4.2 when looking at labor demand in the absence of adjustment cost, inputs i and j are called *dynamically substitutable* if $\lambda_{ij} < 0$ and *dynamically complementary* if $\lambda_{ij} > 0$. When two factors are dynamically substitutable (or complementary) the direction of their adjustments is identical (or inverted). We also see that the average time it takes to adjust an input is influenced by the adjustment costs of all the inputs. So the slowness with which employment is adjusted may be a consequence of the adjustment costs of capital if these two inputs are dynamically complementary.

3.3.2 LINEAR AND ASYMMETRIC ADJUSTMENT COSTS

In a model set in a stochastic environment, the costs of hiring and firing *jointly* influence employment; the firm, in making decisions in the present, takes into account possible future upturns and downturns in the health of the economy. Where adjustment costs are sizable, we should expect to observe low rates of employment turnover. However, the influence of adjustment costs on average employment is a priori indeterminate in sign.

A Model with Two States of Nature

We return to the model in continuous time of section 3.2.2 in which a firm faces linear adjustment costs described by the formula (2.49). We assume that instantaneous production is represented by the function $F(A, L)$ where A and L designate respectively a

parameter affecting productivity and the employment level (the indicator t is left out in order to simplify the notation). To bring out the contrast between the firm's behavior in booms and slumps, it is assumed that parameter A is a random variable following a Poisson process[13] with two states denoted A_G and A_B, with $A_G > A_B$ and $F_{AL} > 0$. The realization A_G then represents the "good" state in which marginal productivity is highest for a given level of employment. The instantaneous transition probability from state A_G to state A_B is denoted q_G, while the instantaneous transition probability from state A_B to state A_G is denoted q_B. The ratio $1/q_G$ (or $1/q_B$) represents the average length of time the economy remains in state A_G (or A_B): it is a measure of the *persistence* of state A_G (or A_B).

The complete and rigorous solution of the optimization problem of a firm that finds itself in an environment of this type is possible but encounters substantial technical difficulties (see Bentolila and Saint-Paul, 1994; Dixit, 1997). For the sake of simplicity, we start by considering a stationary policy linking constant levels of employment L_G and L_B when the productivity variable takes the values A_G and A_B, respectively. We assume moreover that the different parameters of the model are such that $L_G > L_B$, which means that the firm hires when the economy passes from state A_B to state A_G and that it fires when the economy passes from state A_G to state A_B (employment remains unaltered when productivity does not change).

The Decisions of the Firm

Let Π_G and Π_B be the stationary present discounted values of expected profit when the productivity variable is equal to A_G and A_B, respectively. Let W_G and W_B be the real wages linked to these states; expected profits are then defined by the following trade-off equations:

$$r\Pi_G = F(A_G, L_G) - W_G L_G + q_G \left[-c_f(L_G - L_B) + \Pi_B - \Pi_G\right] \tag{2.57}$$

$$r\Pi_B = F(A_B, L_B) - W_B L_B + q_B \left[-c_h(L_G - L_B) + \Pi_G - \Pi_B\right] \tag{2.58}$$

We will return to this type of equation in more detail in chapter 5, section 2.1, dedicated to job search theory, as well as in chapter 9, section 3.2, dedicated to the matching model. We interpret them by reasoning as though there were multiple trade-off possibilities in the investment of an asset. In the present case, an asset worth Π_G brings in $r\Pi_G$ at every date if it is invested in the financial market. An asset corresponding to the same amount of money invested in the labor market brings in, at every date, instantaneous profit $F(A_G, L_G) - W_G L_G$, to which must be added the average gain linked to a change in the state of the economy. This eventuality comes about with a probability of q_G, in which case the firm lets $(L_G - L_B)$ individuals go, which costs it $c_f(L_G - L_B)$, and it then gets an expected profit equal to Π_B. Relation (2.58) defining Π_B is interpreted in analogous manner.

When the level of employment is, for example, equal to L_B and state A_G comes about, the firm makes its hiring decisions in such a way as to maximize the value of its expected profit net of the costs of hiring. So it must solve the following problem:

$$\max_{L_G} \left[\Pi_G - c_h(L_G - L_B)\right] \quad \text{with } L_B \text{ given}$$

[13] The properties of a Poisson process are set out in mathematical appendix D at the end of this book.

In symmetric fashion, if the number of workers equals L_G and state A_B comes about, it decides to terminate employment so as to maximize the value of its profit net of the termination costs. So it must solve the following problem:

$$\max_{L_B} \left[\Pi_B - c_f(L_G - L_B)\right] \quad \text{with } L_G \text{ given}$$

The first-order conditions of these two problems come down to two equations $(\partial\Pi_G/\partial L_G) = c_h$ and $(\partial\Pi_B/\partial L_B) = -c_f$. These two conditions are easy to grasp: the firm increases its workforce as long as the marginal profit of a hire surpasses its cost, and it terminates jobs to the point where the marginal loss due to a termination—equal to $-(\partial\Pi_B/\partial L_B)$—just covers the cost c_f of a termination.

Relations (2.57) and (2.58) allow us to find the partial derivatives of profits Π_G and Π_B with respect to employment levels:

$$\frac{\partial\Pi_G}{\partial L_G} = \left(\frac{1}{r+q_G}\right)\left[F_L(A_G, L_G) - W_G - q_G c_f + \frac{\partial\Pi_B}{\partial L_G}\right]$$

$$\frac{\partial\Pi_B}{\partial L_B} = \left(\frac{1}{r+q_B}\right)\left[F_L(A_B, L_B) - W_B + q_B c_h + \frac{\partial\Pi_G}{\partial L_B}\right]$$

Relations (2.57) and (2.58) also give $(\partial\Pi_B/\partial L_G) = q_B\left[-c_h + (\partial\Pi_G/\partial L_G)\right]$, and $(\partial\Pi_G/\partial L_B) = q_G\left[c_f + (\partial\Pi_B/\partial L_B)\right]$, which implies, with optimality conditions $(\partial\Pi_G/\partial L_G) = c_h$ and $(\partial\Pi_B/\partial L_B) = -c_f$, that $(\partial\Pi_B/\partial L_G) = (\partial\Pi_G/\partial L_B) = 0$. Consequently, the optimal levels L_G and L_B satisfy the following equations:

$$F_L(A_G, L_G) = W_G + q_G c_f + (r + q_G)c_h \tag{2.59}$$

$$F_L(A_B, L_B) = W_B - q_B c_h - (r + q_B)c_f \tag{2.60}$$

The values L_G and L_B correspond respectively to the levels of labor demand in states A_G and A_B, if and only if these two equations imply $L_G > L_B$, which we assume. In this case, employment rises when the firm passes from a bad state to a good one, diminishes when it passes from a good state to a bad one, and remains constant in all other circumstances.

Fluctuations in Employment

We see that taking uncertainty into account through a two-state Poisson process considerably alters the results obtained from models in a stationary deterministic environment. Hiring phases (which correspond to the good state of nature A_G) are linked to a level of employment L_G superior to the one L_B existing in firing phases (i.e., when the bad state A_B is realized). Unlike the case with certainty, the level of employment does not settle definitively on value L_G or L_B; rather it alternates from one value to the other according to the states of nature. Moreover, relations (2.59) and (2.60) indicate that labor demand depends, whatever the state of nature, on both turnover costs c_h and c_f. So it is that we see L_G *decrease* with c_h and c_f. The fact that recruitment is weaker when the cost c_h of a hire rises has nothing surprising about it; yet it appears that the same thing happens

when it is the termination cost c_f that increases. This comes about simply because the entrepreneur foresees that in the future she will have to deal with less favorable phases in the cycle, when terminations will have to be made. Hence high costs of termination put a brake on hires in the upward phases of the cycle. Conversely, relation (2.60) shows that L_B *increases* with c_f and c_h. A rise in the termination cost c_f gives the firm an incentive to do less firing in the downward phases of the cycle, and a rise in the hiring cost c_h gives it incentive to act in the same way, since it foresees that it will have to set about recruiting personnel when the economic cycle turns up again. This analysis suggests that adjustment costs ought to have a stabilizing effect, to the extent that a rise in these costs reduces hires when the economy turns up and puts a brake on firings when it turns down. In certain circumstances, it is even possible that adjustment costs may have a beneficial effect on average employment.

The Labor Turnover Rate

Let us suppose that the economy is composed of a continuum of identical firms and let us designate by ρ the proportion of these that, at a given date, find themselves in the good state of nature. The variable ρ then represents the proportion of firms for which $A = A_G$ holds. For the sake of simplicity, the measure of the continuum of firms is normalized to 1. At any date t, there are ρq_G firms that pass from state A_G to state A_B and that each fire $(L_G - L_B)$ workers. The destruction of jobs thus amounts to $\rho q_G(L_G - L_B)$. Conversely, there are $(1 - \rho)q_B$ firms whose state passes from A_B to A_G and which each hire $(L_G - L_B)$ workers. The creation of jobs thus amounts to $(1 - \rho)q_B(L_G - L_B)$. At stationary equilibrium of the economy, the number of jobs created is equal to the number destroyed, and parameter ρ is thus defined by the equality $\rho = q_B/(q_B + q_G)$. One interesting indicator often utilized to measure job flows is the *turnover rate*, equal, by definition, to the sum of all the jobs created and destroyed. In this model, the turnover rate, denoted τ, is given by:

$$\tau = [(1 - \rho)q_B + \rho q_G](L_G - L_B) = 2\frac{q_G q_B}{q_B + q_G}(L_G - L_B)$$

Since, following (2.59) and (2.60), the employment levels L_G and L_B are functions, respectively decreasing and increasing, of adjustment costs, it results that the turnover rate falls when the "rigidity" of the labor market increases, that is, when the costs of hiring c_h and firing c_f increase. Conversely, the turnover rate is a decreasing function of the wage differential $(W_G - W_B)$. All other things being equal, for that matter, an economy with rigid wages that vary little over the cycle will have a higher labor turnover rate than an economy with more flexible wages. This property may contribute to increase labor turnover rates in certain European countries like Spain, France, and Germany (Bertola and Rogerson, 1997; Bertola, 1999). This observation, which has to do with labor market equilibrium and not just labor demand, is more thoroughly documented in chapter 13, where wages are endogenous and so react to the adjustment costs of employment.

Average Employment

A "rigid" labor market will thus create and destroy fewer jobs than a "flexible" one, but we cannot a priori state anything about the *average* level of employment, which comes under pressure from two opposing directions. In certain circumstances, it is possible

that the average employment level may be higher in a rigid economy than in a flexible one. To see why, let us suppose that the production function takes the quadratic form $F(A, L) = AL - (B/2)L^2$; the marginal productivities appearing on the left-hand side of relations (2.59) and (2.60) are then equal to $A_G - BL_G$ and $A_B - BL_B$, respectively. Let us denote average employment by $\bar{L} = \rho L_G + (1 - \rho)L_B$, average productivity by $\bar{A} = \rho A_G + (1 - \rho)A_B$, and the average wage by $\overline{W} = \rho W_G + (1 - \rho)W_B$. Since $\rho = q_B/(q_B + q_G)$, the addition of relations (2.59) and (2.60) defining L_G and L_B comes to:

$$\bar{A} - B\bar{L} = \overline{W} + \frac{r}{q_B + q_G}\left(q_B c_h - q_G c_f\right)$$

Consequently, under the hypothesis of a quadratic production function, average employment is an *increasing* function of termination costs and a *decreasing* function of hiring costs. This result, however, does not bear a general character: it depends on the specification of the production function and the nature of the shocks. With a homogeneous production function, the termination costs have ambiguous effects. Bertola (1999) has shown, with the help of numerical examples, that a rise in these costs likely has a positive impact, but one small in extent, on average employment. Using a discrete time model, Bentolila and Bertola (1990) have studied the case of a homogeneous production function with shocks that follow a random walk of the type $A_t = A_{t-1} + \epsilon_t$, where ϵ_t is a white noise. The shocks have a permanent effect on the level of parameter A_t. These authors likewise conclude that there is a positive relationship between firing costs and average employment. Nonetheless, for realistic values of the parameters, they show that the impact of firing costs on employment is small in extent. Conversely, Bentolila and Saint-Paul (1994) arrive at markedly different results by assuming that the shocks are independent and have a uniform distribution. They bring to light a nonmonotonic relationship between firing costs and average employment. When these costs are low, the relationship is negative, but it becomes positive when they rise sufficiently high. The consequences of firing costs are analyzed further in chapter 13 in a search and matching model that takes into account job creation, job destruction, and the endogeneity of wages.

3.4 Empirical Aspects of Labor Demand in the Presence of Adjustment Costs

To estimate the importance of employment adjustment costs has been the aim of many in-depth studies; until recently, a quadratic and symmetric representation of these costs was always used. Today, however, studies using microeconomic data generally abandon this representation.

3.4.1 On Estimates

For convenience, numerous studies postulate that adjustment costs take quadratic and symmetric form. In a stochastic environment, and under the hypothesis of rational expectations, the level of present employment L_t is given by the difference equation (2.56), which brings in past employment L_{t-1} and expectations regarding shocks a_{t+i} $(i \geq 1)$ affecting the firm's environment. When expectations are rational, the producer is capable, like the econometrician, of estimating the stochastic process of the a_{t+i}.

For that, it is enough to substitute expectations of these variables at date t by the values predicted for them by the stochastic process estimated by the econometrician. For example, if the stochastic process generating the shocks is autoregressive of order one, or $a_\tau = \alpha a_{\tau-1} + \epsilon_\tau$, $0 < \alpha < 1$, where ϵ_τ is a white noise, then $\mathbb{E}_t a_{t+i} = \alpha^i a_t$ and equation (2.56) reads:

$$L_t = \lambda L_{t-1} + \frac{\mu_0}{1 - \alpha(a_0\mu_0)} a_t$$

In this way we can deduce the median lag[14] δ. When the random variable a_t follows a more complex process, the hypothesis of rational expectations allows us to obtain an equation linking present employment to the (observable) values of shocks past and present. Using panel data, this leads to estimating equations of the form:

$$L_{it} = \lambda L_{i,t-1} + \boldsymbol{X}_{it}\boldsymbol{\beta} + \eta_i + \varepsilon_{it}$$

where L_{it} designates the level of employment in firm i at date t, $\boldsymbol{X}_{it}$ designates a vector of the characteristics of the firm, η_i designates a fixed effect independent of time associated to firm i, and ε_{it} designates an error term. In this dynamic setting the ordinary least squares estimator for λ is inconsistent because the regressor $L_{i,t-1}$ is correlated with the error term $\eta_i + \varepsilon_{it}$. It then remains to estimate this equation with adequate methods, which are presented in econometric textbooks (see, e.g., Balgati, 2008; Wooldridge, 2010).

The expression (2.56) of labor demand upon which the preceding method is based is obtained using precise hypotheses concerning the production function (linear quadratic) and adjustment costs (quadratic and symmetric). To get around having to postulate such restrictive hypotheses, another approach consists of estimating the Euler equations directly. These indicate (see, e.g., (2.52)) that employment at date t depends on both past and expected future variables. The hypothesis of rational expectation allows us, in making our estimates, to replace expectation variables by their realizations, using the technique of generalized moments or that of instrumental variables, with instruments belonging to the information set of the firm at date t (Hamilton, 1994, chapter 14).

3.4.2 Main Results

The results obtained from estimating dynamic equations of labor demand are given by Hamermesh (1993, chapters 7 and 8) and Hamermesh and Pfann (1996). From this it emerges, among other things, that the adjustment costs of employment cannot be validly represented by a simple quadratic and symmetric component.

On the Form of Adjustment Costs

Until the end of the 1980s, the great majority of empirical studies used quadratic and symmetric cost functions. Most often they found that adjustment costs were minor, on the order of 20% of the annual labor cost for the United States and United Kingdom.

[14] In a discrete time model, the median lag is equal to $-\ln 2 / \ln \lambda$.

But since then, studies grounded in microeconomic data have developed notably, and all of them reach the same conclusion: the hypothesis that adjustment costs are symmetric and convex (like quadratic functions, for example) must be rejected. A good representation must, in all likelihood, be asymmetric, piecewise linear, and involve fixed costs (Hamermesh, 1989; Hamermesh and Pfann, 1996). Nielsen et al. (2007) find similar results on Norwegian firms. Their econometric evidence supports the existence of purely fixed components, unrelated to plant size. They also estimate that quadratic components of costs are important and that both fixed and convex costs are higher for employment contractions.

The work of Abowd and Kramarz (2003) and Kramarz and Michaud (2010), grounded in French data, confirm this judgment. They find that the costs of terminating employment are almost linear functions of terminations, with a very high lump-sum component, explainable by the existence in France of economically motivated procedures for mass termination. They estimate that separation costs are significantly larger than hiring costs, that collective terminations (dismissal of at least 10 workers during a 30-day period) are more expensive than individual terminations, and that costs are often concave and induce firms to group their hiring and separations.

It is important to note that adjustment costs can have different forms at firm level and aggregate level. Cooper and Willis (2009) specify a dynamic optimization problem at the plant level, allowing for both convex and nonconvex adjustment costs. Contrary to evidence at the micro level in support of nonconvex adjustment costs, their findings indicate that piecewise quadratic adjustment costs are sufficient to match the aggregate dynamics of employment. Caballero et al. (1997) and King and Thomas (2006) have developed models in which fixed adjustment costs imply that individual firms' employment adjustments are discrete, but their asynchronous timing implies a smooth aggregate employment series similar to that implied by the traditional adjustment model with quadratic costs.

On the Speed of Adjustments

Many studies published in the 1980s and at the beginning of the 1990s have tried to estimate the speed of adjustment of labor demand. They have adopted a quadratic and symmetric representation of labor turnover costs and have not taken into account possible adjustment costs for other inputs. It appears that the speed of adjustment is relatively high, since according to Hamermesh (1993, p. 261), a reasonable estimate of the median lag is 1 to 2 quarters (1.4 quarters on the basis of quarterly data, and 1.2 quarters on the basis of monthly data). Estimating simultaneous adjustments of multiple inputs does not seem to change this conclusion. With a moderate degree of confidence, certain studies do show, however, that labor services would be dynamic substitutes with the rate of capital utilization. In other words, firms would adjust the utilization of their equipment all the more quickly, the greater the disequilibrium between desired employment and actual employment. It is worth noting that most of the estimates apply to the United States and Canada.

Firms adjust hours of work more rapidly than numbers of workers. This result points to the conclusion that adjustment costs are greater for workers than for hours, which also explains why workers are kept on during cyclical downturns. There exists no robust result, however, allowing us to assert that workers and hours are dynamic substitutes or complements.

Most international comparisons indicate that employment adjusts more rapidly in the United States than anywhere else. They also suggest that the adjustment takes place more rapidly in Europe than in Japan. The reasons for these divergences are not well established. Contrary to what one might think, the degree of unionization does not appear to be a significant variable. The greater or lesser rigor of legislation regarding the termination of employment might, however, be an explanation for this phenomenon. Abraham and Houseman (1993) compare labor adjustment practices in the United States and in Germany. Lazear (1990) and Dertouzos and Karoly (1990) find that strengthened job security, that is, an increased cost of terminating employment, has a negative impact, but Bertola (1990) estimates that these costs have practically no influence. We return to this problem in detail in chapter 13.

4 SUMMARY AND CONCLUSION

- *Conditional* demands represent the quantities of each input that a firm desires to use to attain a *given* level of output. The cost function is the minimal value of the total cost of the inputs corresponding to this operation. *Unconditional* demands designate the quantities of each input that a firm desires to use to maximize its profit. The conditional and unconditional demands for an input always decrease with the cost of the input. In absolute value, the wage elasticity of unconditional labor demand diminishes, the more market power the firm has. It increases with the elasticity of capital/labor substitution.
- Labor and capital are called *gross substitutes* when a rise in the price of a factor leads the firm to reduce the *unconditional* demand for this factor and increase that for another. When this rise implies a reduction in the unconditional demand for each factor, labor and capital are described as *gross complements*. Two factors are *p-substitutes* (or *p-complements*) if *conditional* demand for one of them increases (or falls off) when the cost of the other factor rises. If the production function includes only two inputs, then they are necessarily p-substitutes.
- Cross elasticity of conditional demand for a factor i with respect to the price of a factor j increases in absolute value with the share of factor j in the total cost and with the elasticity of substitution between these two factors.
- A reduction in standard hours has the same impact on employment as a rise in fixed costs. That is why, when a firm makes use of overtime hours, a reduction in standard hours increases the actual work week by inflating the number of overtime hours used. The rise in fixed costs tends to hold back the level of production and hence that of employment. Therefore, a reduction in standard hours may have deleterious effects on employment if it is not accompanied by a reduction in fixed costs.
- At the aggregate level, we may take it that the absolute value of the elasticity of conditional labor demand with respect to the cost of labor falls in the interval [0.15–0.75], with consensus settling on a figure of 0.30. Unskilled labor is more easily substitutable for capital than skilled labor is. Skilled labor and capital are p-complements. Workers and hours are p-substitutes with capital.

- The adjustment costs of labor are often sizable. In the United States, the hiring costs are higher than the termination costs. In France, the termination costs clearly outrank other adjustment costs.
- When adjustment costs are quadratic, the firm gradually adjusts the size of its workforce. But it alters the size of the workforce instantaneously if adjustment costs are linear. Under this hypothesis, a rise in the costs of terminating employment allows the firm to stabilize labor demand when labor demand is high. A decline in hiring costs has the effect of increasing labor demand when it is low. In a stochastic environment, a rise in hiring costs generally has a negative impact on average employment. But a rise in firing costs may have a positive impact on average employment.
- Studies grounded on microeconomic data reject the hypothesis of quadratic and convex adjustment costs. A good representation of these costs must be asymmetric, be piecewise linear, and include a lump-sum component.
- Firms adjust the volume of their hours more quickly than they do that of their workforce. Adjustment times are shorter for unskilled labor than for skilled labor.
- The fiction of a firm that lasts forever is no doubt inadequate to the task of characterizing fully the behavior of labor demand. We have to take into consideration firms that fail and explain how new ones come into being. Empirically, job creation and destruction due to the closing down and starting up of firms may be as great as, or greater than, that caused by the expansion and contraction of existing firms. These problems will be tackled in chapters 9 and 10 dealing with employment and unemployment in a macroeconomic perspective.
- The functioning of the firm is studied in abstraction from specific problems linked to the management of human resources. In reality, wages, working conditions, the scheduling of hours of work, and employment itself are all objects of formal or informal negotiation. As well, the efficiency of labor may be sensitive to the level and form of remuneration paid and the hierarchical structure prevailing in the firm. These features of the wage relationship may affect labor demand. For example, the linkage between employment and wages may be affected by the bargaining power of the workers and their preferences. Such considerations are absent from the traditional theory of labor demand; they will be dealt with in chapters 6 and 7.

5 RELATED TOPICS IN THE BOOK

- Chapter 3, section 1.2: The question of tax incidence
- Chapter 3, section 1.3: The effect of a shock on labor supply
- Chapter 7, section 3: Models of collective bargaining
- Chapter 9, section 2: The competitive model with adjustment costs
- Chapter 9, section 3: The matching model
- Chapter 10, section 2.2: A model with skills and tasks
- Chapter 12, section 2: The minimum wage
- Chapter 13, section 2: Employment protection
- Chapter 14, section 2.3: Employment subsidies and the creation of public jobs

6 FURTHER READINGS

Bertola, G. (1999). Microeconomic perspectives on aggregate labor markets. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 3C, chap. 45, pp. 2985–3028). Amsterdam: Elsevier Science.

Hamermesh, D. (1993). *Labor demand*. Princeton, NJ: Princeton University Press.

Hamermesh, D., & Pfann, G. (1996). Adjustment costs in factor demand. *Journal of Economic Literature*, *34*, 1264–1292.

7 APPENDICES

7.1 THE CONVEXITY OF ISOQUANTS

In this appendix, we show that the isoquants of a production function with two inputs, denoted $F(K,L)$, are strictly convex when the production function is strictly increasing with respect to each of its arguments and strictly concave in (K,L). Readers will recall that a function $f:\mathbb{R}^n \to \mathbb{R}$ is strictly convex (or strictly concave) if and only if:

$$f[\lambda x + (1-\lambda)y] < (\text{resp. } >)\lambda f(x) + (1-\lambda)f(y), \quad \forall (x,y) \in \mathbb{R}^n \times \mathbb{R}^n, \quad \forall \lambda \in (0,1)$$

By definition, the isoquant corresponding to a given output level Y is a curve $K(L)$ defined by $F[K(L),L] \equiv Y$. This equality implies in particular:

$$F[K(\lambda L_1 + (1-\lambda)L_2),\ \lambda L_1 + (1-\lambda)L_2] = Y \quad \forall (L_1,L_2), \quad \forall \lambda \in (0,1) \tag{2.61}$$

The production function being strictly concave, for each quadruplet (K_1,K_2,L_1,L_2) we always have:

$$F[\lambda K_1 + (1-\lambda)K_2,\ \lambda L_1 + (1-\lambda)L_2] > \lambda F(K_1,L_1) + (1-\lambda)F(K_2,L_2), \quad \forall \lambda \in (0,1) \tag{2.62}$$

Let us posit $K_1 = K(L_1)$ and $K_2 = K(L_2)$, which implies $F[K(L_1),L_1] = F[K(L_2),L_2] = Y$; the right-hand side of (2.62) is then equal to Y. Whatever the values of L_1 and L_2, and for all $\forall \lambda \in (0,1)$, relation (2.61) then gives:

$$F[\lambda K(L_1) + (1-\lambda)K(L_2),\ \lambda L_1 + (1-\lambda)L_2] > F[K(\lambda L_1 + (1-\lambda)L_2),\ \lambda L_1 + (1-\lambda)L_2] \tag{2.63}$$

The production function being taken as strictly increasing with respect to each of its arguments, inequality (2.63) allows us to write:

$$K[\lambda L_1 + (1-\lambda)L_2] < \lambda K(L_1) + (1-\lambda)K(L_2) \quad \forall (L_1,L_2), \quad \forall \lambda \in (0,1)$$

This last relation shows that the isoquant $K(L)$ is represented by a strictly convex curve in the plane (K,L).

7.2 The Properties of Cost Functions

Let us consider a firm producing a unique good, whose technology can be represented by a production function with n arguments, denoted $Y = F(X^1, ..., X^n)$. Let us designate the vector indicating the quantities of the inputs utilized in the production of a quantity Y of the good by $\mathbf{X} = (X^1, ..., X^n)$, and the vector indicating their respective price by $\mathbf{W} = (W^1, ..., W^n)$. Let $\mathcal{Y}$ be the set of the vectors $\mathbf{X}$ such that $F(\mathbf{X}) \geq Y$ for a given output level Y. The cost function of this firm, denoted $C(\mathbf{W}, Y)$, is then defined by the following relation:

$$C(\mathbf{W}, Y) = \min_{\mathbf{X} \in \mathcal{Y}} \sum_{i=1}^{n} W^i X^i \tag{2.64}$$

(i) *$C(\mathbf{W}, Y)$ is increasing and homogeneous of degree 1 in $\mathbf{W}$.*

The cost function evidently increases with the price of each input, since for a given vector $\mathbf{X}$ of inputs, the rise in price W^i of input i increases the total cost of production. To show that the cost function is homogeneous, it is enough to note that for any positive number λ we have:

$$\min_{\mathbf{X} \in \mathcal{Y}} \sum_{i=1}^{n} (\lambda W^i) X^i = \lambda \min_{\mathbf{X} \in \mathcal{Y}} \sum_{i=1}^{n} W^i X^i$$

Put another way:

$$C(\lambda \mathbf{W}, Y) = \lambda C(\mathbf{W}, Y), \quad \forall \lambda \geq 0, \ \forall (\mathbf{W}, Y)$$

Consequently, the cost function is homogeneous of degree 1 with respect to vector $\mathbf{W} = (W^1, ..., W^n)$ of the input costs.

(ii) *$C(\mathbf{W}, Y)$ is concave in $\mathbf{W}$.*

Given two vectors $\mathbf{W} = (W^1, ..., W^n)$ and $\mathbf{V} = (V^1, ..., V^n)$ of the input costs, we always have:

$$C(\mathbf{W}, Y) \leq \sum_{i=1}^{n} W^i X^i \quad \forall \mathbf{X} \in \mathcal{Y} \tag{2.65}$$

$$C(\mathbf{V}, Y) \leq \sum_{i=1}^{n} V^i X^i \quad \forall \mathbf{X} \in \mathcal{Y} \tag{2.66}$$

Let us take a scalar $\lambda \in [0, 1]$ and let us multiply relations (2.65) and (2.66) respectively by λ and $(1 - \lambda)$. If we add the results obtained side by side, we get:

$$\lambda C(\mathbf{W}, Y) + (1 - \lambda) C(\mathbf{V}, Y) \leq \sum_{i} \left[\lambda W^i + (1 - \lambda) V^i \right] X^i, \quad \forall \lambda \in [0, 1], \quad \forall \mathbf{X} \in \mathcal{Y}$$

This inequality being satisfied for any vector of inputs $\mathbf{X}$ of the set $\mathcal{Y}$, it implies in particular:

$$\lambda C(\mathbf{W}, Y) + (1 - \lambda) C(\mathbf{V}, Y) \leq \min_{\mathbf{X} \in \mathcal{Y}} \sum_{i} \left[\lambda W^i + (1 - \lambda) V^i \right] X^i, \quad \forall \lambda \in [0, 1] \tag{2.67}$$

By the definition of the cost function, we also have:

$$C[\lambda \mathbf{W} + (1-\lambda)\mathbf{V}, Y] \equiv \min_{\mathbf{X}\in\mathcal{Y}} \sum_{i=1}^{n} \left[\lambda W^i + (1-\lambda)V^i\right] X^i \tag{2.68}$$

Comparison of relations (2.67) and (2.68) then shows that the cost function satisfies the following inequality:

$$C[\lambda \mathbf{W} + (1-\lambda)\mathbf{V}, Y] \geq \lambda C(\mathbf{W}, Y) + (1-\lambda)C(\mathbf{V}, Y) \quad \forall \lambda \in [0,1], \quad \forall (\mathbf{W}, \mathbf{V}, Y)$$

This proves the concavity of function $C(\mathbf{W}, Y)$ with respect to $\mathbf{W}$.

(iii) *Shephard's lemma*

Let $\bar{\mathbf{X}} = (\bar{X}^1, ..., \bar{X}^n)$ be a vector minimizing the total cost when the unit prices of inputs are given by the vector $\mathbf{W} = (W^1, ..., W^n)$. In other terms, $\bar{\mathbf{X}}$ is a solution of the problem described by relation (2.64). For given $Y, \mathbf{W}$ and so $\mathbf{X}$, let us consider the function with n arguments $\Phi = \Phi(\mathbf{V})$, with $\mathbf{V} = (V^1, ..., V^n)$, defined by:

$$\Phi(\mathbf{V}) \equiv C(\mathbf{V}, Y) - \sum_{i=1}^{n} V^i \bar{X}^i \tag{2.69}$$

Since, by construction, we have:

$$C(\mathbf{V}, Y) = \min_{\mathbf{X}\in\mathcal{Y}} \sum_{i=1}^{n} V^i X^i, \quad \forall \mathbf{V}$$

Relation (2.69) implies $\Phi(\mathbf{V}) \leq 0, \ \forall \mathbf{V}$. Still by definition of the cost function, relation (2.69) likewise entails $\Phi(\mathbf{W}) = 0$. Vector $\mathbf{W}$ thus represents a maximum for function $\Phi(.)$. For all i, the partial derivative of the latter with respect to V^i is thus null at point $\mathbf{W}$. Differentiating the two members of relation (2.69) with respect to V^i, we get:

$$\bar{X}^i = C_i(\mathbf{W}, Y), \quad \forall i = 1, ..., n \tag{2.70}$$

where C_i designates the partial derivative of the cost function with respect to its i^{th} argument. Relation (2.70) constitutes Shephard's lemma.

(iv) *The case of a homogeneous production function*

Let us henceforth assume that the production function is homogeneous of degree $\theta > 0$. By the definition of the cost function, we have:

$$C(\mathbf{W}, \lambda Y) = \min_{\mathbf{X}} \sum_{i=1}^{n} W^i X^i \quad \text{subject to constraint } F(\mathbf{X}) \geq Y \tag{2.71}$$

In this problem, let us make the change of variable $\mathbf{Z} = \lambda^{-1/\theta}\mathbf{X}$, that is, $Z^i = \lambda^{-1/\theta}X^i$ for all $i = 1, \ldots, n$. The problem (2.71) is then written:

$$C(\mathbf{W}, \lambda Y) = \lambda^{1/\theta} \min_{\mathbf{Z}} \sum_{i=1}^{n} W^i Z^i \quad \text{subject to constraint } F(\mathbf{Z}) \geq Y$$

We can immediately deduce:

$$C(\mathbf{W}, \lambda Y) = \lambda^{1/\theta} C(\mathbf{W}, Y) \tag{2.72}$$

This last equation shows that the cost function is indeed homogeneous of degree $1/\theta$ in Y when the production function is homogeneous of degree θ. Making $\lambda = 1/Y$ in (2.72), we arrive at:

$$C(\mathbf{W}, Y) = C(\mathbf{W}, 1) Y^{1/\theta}, \quad \forall (\mathbf{W}, Y)$$

Applying Shephard's lemma (2.70), we find:

$$\bar{X}^i = C_i(\mathbf{W}, Y) = C_i(\mathbf{W}, 1) Y^{1/\theta}$$

Consequently, the conditional demands functions are equally homogeneous of degree $1/\theta$ with respect to Y.

(v) *Production function with two inputs*

When the only arguments of the production function are capital and labor, or $Y = F(K, L)$, all the relations previously established of course remain satisfied. In particular, if W designates the labor cost and R the user cost of capital, Shephard's lemma is written with the obvious notations:

$$\bar{L} = C_W(W, R, Y) \quad \text{and} \quad \bar{K} = C_R(W, R, Y) \tag{2.73}$$

To find a simple expression of the elasticity of substitution between capital and labor, we must first note that the homogeneity to degree 1 of the cost function with respect to (W, R) implies:

$$C(W, R, Y) = RC(W/R, 1, Y), \quad \forall (W, R, Y)$$

Differentiating this relation with respect to W and R entails successively:

$$C_W(W, R, Y) = C_W(W/R, 1, Y) \tag{2.74}$$

$$C_R(W, R, Y) = C(W/R, 1, Y) - (W/R) C_W(W/R, 1, Y) \tag{2.75}$$

If we now, for example, derive (2.74) with respect to R, we get:

$$C_{WR}(W, R, Y) = -\frac{W}{R^2} C_{WW}(W/R, 1, Y) \tag{2.76}$$

The cost function being concave, C_{WW} is negative or null, and in consequence we will necessarily have $C_{WR} \geq 0$. In the case of two factors of production, the elasticity of substitution σ is defined by:

$$\sigma = \frac{W/R}{\bar{K}/\bar{L}} \frac{\partial(\bar{K}/\bar{L})}{\partial(W/R)}$$

With the help of Shephard's lemma (2.73) and relations (2.74) and (2.75), we can write:

$$\frac{\bar{K}}{\bar{L}} = \frac{C_R(W,R,Y)}{C_W(W,R,Y)} = \frac{C(W/R,1,Y) - (W/R)C_W(W/R,1,Y)}{C_W(W/R,1,Y)}$$

Or again:

$$\frac{\bar{K}}{\bar{L}} = \frac{C(W/R,1,Y)}{C_W(W/R,1,Y)} - \frac{W}{R}$$

Differentiating this equation with respect to W/R, we arrive at:

$$\frac{\partial(\bar{K}/\bar{L})}{\partial(W/R)} = -\frac{C(W/R,1,Y)C_{WW}(W/R,1,Y)}{C_W^2(W/R,1,Y)}$$

Using (2.76) and Shephard's lemma (2.73), we find after rearranging terms that the elasticity of substitution between capital and labor satisfies the relation:

$$\sigma = \frac{C(W,R,Y)C_{WR}(W,R,Y)}{C_W(W,R,Y)C_R(W,R,Y)}$$

7.3 THE OPTIMAL VALUE OF HOURS WORKED

If the amount of hours desired is such that $H \leq T$, the following inequality is satisfied:

$$\frac{\Omega T + (1+x)\Omega(H-T) + Z}{e(H)} \geq \frac{\Omega H + Z}{e(H)} \tag{2.77}$$

In this case, if the minimum of function $\varphi(H) \equiv (\Omega H + Z)/e(H)$ lies within interval $[0, T]$, it represents a *global* minimum for function W defined by (2.34). Differentiating $\varphi(H)$ with respect to H, we find after several calculations:

$$\varphi'(H) = \frac{1}{He(H)}\left[(1-\eta_H^e)\Omega H - Z\eta_H^e\right] \tag{2.78}$$

From that we deduce that the optimal number of hours worked is given by:

$$H^* = \frac{\eta_H^e}{1-\eta_H^e}\frac{Z}{\Omega} \tag{2.79}$$

For this value of H to be smaller than T it is necessary and sufficient that the following inequality be satisfied:

$$\frac{Z}{\Omega T} \leq \frac{1-\eta_H^e}{\eta_H^e}$$

Moreover, equations (2.78) and (2.79) show that at the optimum, we have:

$$\varphi''(H^*) = \frac{(1-\eta_H^e)\Omega}{H^* e(H^*)}$$

And the second-order condition for a minimum, or $\varphi'' > 0$, then dictates $\eta_H^e < 1$. The first line of relation (2.35) is thus proved.

If the desired number of hours is such that $H \geq T$, the inequality (2.79) is inverted and the minimum of the function $\psi(H) \equiv [\Omega T + (1+x)\Omega(H-T) + Z]/e(H)$ represents a global minimum for function W. Differentiating $\psi(H)$ with respect to H, we get:

$$\psi'(H) = \frac{1}{He(H)}\left[(1-\eta_H^e)(1+x)\Omega H - (Z - \Omega x T)\eta_H^e\right] \tag{2.80}$$

The optimum number of hours worked is then given by:

$$H = \frac{\eta_H^e}{1-\eta_H^e}\frac{Z - \Omega x T}{(1+x)\Omega} \tag{2.81}$$

This value of H is greater than T when:

$$\frac{Z}{\Omega T} \geq \frac{1+x-\eta_H^e}{\eta_H^e}$$

Equations (2.80) and (2.81) again imply at the optimum:

$$\psi''(H) = \frac{(1-\eta_H^e)(1+x)\Omega}{He(H)}$$

And the second-order condition for a minimum always comes down to $\eta_H^e < 1$. The second line of relation (2.35) is thus established.

Finally, the optimum number of hours worked coincides with standard hours ($H = T$) when the minima of functions $\varphi(H)$ and $\psi(H)$ are not respectively in the intervals $[0, T]$ and $[T, +\infty]$. This configuration appears when the following inequalities are satisfied:

$$\frac{1-\eta_H^e}{\eta_H^e} \leq \frac{Z}{\Omega T} \leq \frac{1+x-\eta_H^e}{\eta_H^e}$$

Thus the third line of relation (2.35) is proved.

REFERENCES

Abowd, J., Corbel, P., & Kramarz, F. (1999). The entry and exit of workers and the growth of employment. *Review of Economics and Statistics*, *81*(2), 170–187.

Abowd, J., & Kramarz, F. (2003). The costs of hiring and separations. *Labour Economics*, *10*(5), 499–530.

Abraham, K., & Houseman, S. (1993). *Job security in America: Lessons from Germany*. Washington, DC: Brookings Institution.

Acemoglu, D., Autor, D., & Lyle, D. (2004). Women, war and wages: The effect of female labor supply on the wage structure at mid-century. *Journal of Political Economy*, *112*(3), 497–551.

Angrist, J. (1996). Short-run demand for Palestinian labor. *Journal of Labor Economics*, *14*, 425–453.

Arrow, K., Chenery, H., Minhas, B., & Solow, R. (1961). Capital-labor substitution and economic efficiency. *Review of Economics and Statistics*, *43*, 225–250.

Autor, D., Katz, L., & Krueger, A. (1998). Computing inequalities: Have computers changed the labor market? *Quarterly Journal of Economics*, *113*, 1169–1213.

Balgati, B. (2008). *Econometric analysis of panel data* (4th ed.). Chichester, U.K.: Wiley.

Bentolila, S., & Bertola, G. (1990). Firing costs and labor demand: How bad is eurosclerosis? *Review of Economic Studies*, *57*, 381–402.

Bentolila, S., & Saint-Paul, G. (1994). A model of labor demand with linear adjustment costs. *Labour Economics*, *1*, 303–326.

Bertola, G. (1990). Job security, employment and wages. *European Economic Review*, *34*, 851–886.

Bertola, G. (1999). Microeconomic perspectives on aggregate labor markets. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 3C, chap. 45, pp. 2985–3028). Amsterdam: Elsevier Science.

Bertola, G., & Rogerson, R. (1997). Institutions and labor reallocation. *European Economic Review*, *41*, 1147–1171.

Blanchard, O., & Fischer, S. (1989). *Lectures on macroeconomics*. Cambridge, MA: MIT Press.

Caballero, R., Engel, E., & Haltiwanger, J. (1997). Aggregate employment dynamics: Building from microeconomic evidence. *American Economic Review*, *87*(1), 115–137.

Cahuc, P., & Zylberberg, A. (2008). Reduction of working time and employment. In T. Boeri, M. Burda, & F. Kramarz (Eds.), *Working hours and job sharing in the EU and USA: Are Europeans lazy? Or Americans crazy?* Oxford, U.K.: Oxford University Press.

Calmfors, L., & Hoel, M. (1988). Work sharing and overtime. *Scandinavian Journal of Economics*, *90*, 45–62.

Chang, C., & Stefanou, S. (1988). Specification and estimation of asymmetric adjustment rates for quasi fixed factors of production. *Journal of Economic Dynamics and Control*, *12*, 145–151.

Christensen, L., Jorgenson, D., & Lau, L. (1973). Transcendental logarithmic production frontiers. *Review of Economics and Statistics*, *55*, 28–45.

Cobb, C., & Douglas, P. (1928). A theory of production. *American Economic Review, Papers and Proceedings*, *18*, 139–165.

Cooper, R., & Willis, J. (2009). The cost of labor adjustment: Inferences from the gap. *Review of Economic Dynamics*, *12*(4), 326–347.

Crépon, B., & Kramarz, F. (2008). The two French work-sharing experiments: Employment and productivity effects. In T. Boeri, M. Burda, & F. Kramarz (Eds.), *Working hours and job sharing in the EU and USA: Are Europeans lazy? Or Americans crazy?* Oxford, U.K.: Oxford University Press.

Davis, S., Faberman, J., & Haltiwanger, J. (2012). Labor market flows in the cross section and over time. *Journal of Monetary Economics*, *59*(1), 1–18.

Dertouzos, J., & Karoly, L. (1990). Labor market responses to employer liability. Mimeo, Rand Corporation.

Diewert, W. (1971). An application of the Shephard duality theorem, a generalized Leontief production function. *Journal of Political Economy*, *79*, 481–507.

Dixit, A. (1997). Investment and employment dynamics in the short run and the long run. *Oxford Economic Papers*, *49*, 1–20.

Eisner, R., & Strotz, R. (1963). Determinants of business investment. Commission on Money and Credit. In *Impacts of monetary policy*, part I (pp. 59–223). Englewood Cliffs, NJ: Prentice Hall.

Gianella, C., & Lagarde, P. (1999). Productivity of hours in the aggregate production function: An evaluation on a panel of French firms from the manufacturing sector (Document de travail G 9918). Direction des Etudes et Synthèses Economiques, INSEE.

Goux, D., Maurin, E., & Pauchet, M. (2001). Fixed-term contracts and the dynamics of labour demand. *European Economic Review*, *45*, 533–552.

Hamermesh, D. (1989). Labor demand and the structure of adjustment costs. *American Economic Review*, *79*(4), 674–689.

Hamermesh, D. (1992). A general model of dynamic labor demand. *Review of Economics and Statistics*, *74*(4), 733–737.

Hamermesh, D. (1993). *Labor demand*. Princeton, NJ: Princeton University Press.

Hamermesh, D. (1995). Labour demand and the source of adjustment costs. *Economic Journal*, *105*, 620–634.

Hamermesh, D. (2006). Overtime laws and the margins of work timing. In P. Askenazy, D. Carton, F. de Coninck, & M. Gollac (Eds.), *Organisation et intensité du travail*. Paris: Octares.

Hamermesh, D., & Pfann, G. (1996). Adjustment costs in factor demand. *Journal of Economic Literature*, *34*, 1264–1292.

Hamermesh, D., & Trejo, S. (2000). The demand for hours of labor: Direct evidence from California. *Review of Economics and Statistics*, *82*, 38–47.

Hamilton, J. (1994). *Time series analysis*. Princeton, NJ: Princeton University Press.

Hart, R. (1987). *Working time and employment*. Boston, MA: Allen and Unwin.

Hart, R., & MacGregor, P. (1988). The returns to labour services in West German manufacturing industry. *European Economic Review*, *32*, 947–963.

Hicks, J. (1932). *The theory of wages*. New York, NY: Macmillan.

Holt, C., Modigliani, F., Muth, J., & Simon, H. (1960). *Planning, production, inventories and work force*. Englewood Cliffs, NJ: Prentice Hall.

Hunt, J. (1999). Has work-sharing worked in Germany? *Quarterly Journal of Economics*, *114*, 117–148.

Jaramillo, F., Schiantarelli, F., & Sembenelli, A. (1993). Are adjustment costs for labor asymmetric? An econometric test on panel data for Italy. *Review of Economics and Statistics*, *75*, 640–648.

Johnson, G. (1997). Changing in earnings inequality: The role of demand shift. *Journal of Economic Perspectives*, *11*(2), 41–54.

King, R., & Thomas, J. (2006). Partial adjustment without apology. *International Economic Review*, *47*(3), 779–809.

Kramarz, F., & Michaud, M.-L. (2010). The shape of hiring and separation costs in France. *Labour Economics*, *17*(1), 27–37.

Lazear, E. (1990). Job security provision and employment. *Quarterly Journal of Economics*, *105*, 699–726.

Leslie, D., & Wise, J. (1980). The productivity of hours in UK manufacturing and production industries. *Economic Journal*, *90*, 74–84.

Lucas, R. (1967). Optimal investment policy and the flexible accelerator. *International Economic Review*, *8*, 78–85.

Marshall, A. (1920). *Principles of economics*. New York, NY: Macmillan.

Messe, R. (1980). Dynamic factor demand schedules for labor and capital under rational expectations. *Journal of Econometrics*, *14*, 141–158.

Nadiri, M., & Rosen, S. (1973). *A disequilibrium model of production*. New York, NY: National Bureau of Economic Research.

Nase, D. (2009). The high cost of turnover, 25 September, articlebase, www.articlesbase.com/human-resources-articles/the-high-cost-of-turnover-1271345.html.

Neumark, D., & Wascher, W. (2008). *Minimum wages*. Cambridge, MA: MIT Press.

Nielsen, O., Salvanes, K., & Schiantarelli, F. (2007). Employment adjustment, the structure of adjustment costs, and plant size. *European Economic Review*, *51*, 577–598.

OECD. (1999). *OECD employment outlook*. Paris: OECD Publishing.

Pfann, G., & Palm, F. (1993). Asymmetric adjustment costs in non-linear labour demand models for the Netherlands and UK manufacturing sectors. *Review of Economic Studies*, *60*, 397–412.

Rosen, S. (1968). Short run employment variation on class-1 railroads in the U.S., 1947–63. *Econometrica*, *36*, 511–529.

Sargent, T. (1986). *Macroeconomic theory*. Boston, MA: Academic Press.

Takayama, A. (1986). *Mathematical economics* (2nd ed.). New York, NY: Cambridge University Press.

Trejo, S. (1991). The effects of overtime pay regulation on worker compensation. *American Economic Review*, *81*, 719–740.

Venn, D. (2009). Legislation, collective bargaining and enforcement: Updating the OECD employment protection indicators. OECD Social, Employment and Migration Working Papers. www.oecd.org/els/workingpapers.

Wooldridge, J. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). Cambridge, MA: MIT Press.

CHAPTER 3

Competitive Equilibrium and Compensating Wage Differentials

In this chapter we will:

- Describe the basic model of the labor market in competitive equilibrium
- See how the analysis of interactions between supply and demand offers insight into the problem of fiscal incidence
- See how the interactions between supply and demand allow us to estimate the elasticity of labor demand. Apply this strategy using data and programs allowing us to replicate the main results of the paper of Acemoglu, Autor, and Lyle (2004) that uses the instrumental variable method to estimate the elasticity of labor demand in a consistent way
- Understand why, in a situation of perfect competition, the hedonic theory predicts that wage differentials compensate for the laboriousness or danger of tasks
- Provide evaluations of the value of statistical life
- Understand how the assortative matching model shows that very small differences in talent can lead to huge remuneration differentials
- Use the assortative matching model to explain the soaring remuneration of superstars and chief executive officers (CEOs)

INTRODUCTION

Why does John earn a lower wage than Jane? A number of possible reasons come to mind. Jane stayed in school longer or obtained a more prestigious diploma. Jane's work is more demanding, with heavy responsibilities. Jane is older or has been with her company longer. She is more highly motivated and efficient. John works in a region where the average wage is lower or Jane works in a firm with higher productivity or in a region where the demand for labor is stronger, and so on.

One of the purposes of labor economics is to assess how relevant, and how significant, each of these explanatory factors is. On the theoretical level, we must specify which hypotheses are being used to justify every answer proposed to the question of

why some people earn higher wages than others. The answers to this question are not trivial, and without elaborating a simple yet rigorous conceptual framework to represent the different elements that influence wages, they cannot be given. The basic frame of reference adopted by economic analysis is the model of perfect competition. When applied to labor economics, it explains the formation of wages by assuming that they match all labor supply with all labor demand; the attendant hypotheses are that agents have no market power because there is free entry into the market and information is perfect. This frame of reference leads to positive conclusions about the setting of compensation for labor, which empirical studies allow us to confirm or reject.

In the first section of this chapter, we will describe the basic model of the labor market in competitive equilibrium. As we shall see, the interface between supply and demand in a market where agents are price takers leads to an efficient allocation of resources. We shall see as well that the model of perfect competition is very useful for evaluating the consequences of taxation. We have broached this theme already in chapter 1, where we studied the effect of taxes on labor supply without analyzing their impact on wages. But in modifying labor supply, taxes affect the equilibrium of the labor market and thus wages. So in order to evaluate tax incidence correctly, it is necessary to take wage adjustment into account. We shall see how the impact of taxes on employment and wages depends on the interplay between labor supply and demand. More generally, the model of perfect competition is helpful in analyzing the effects on employment and wages of shocks that affect this interplay, such as the massive entry of women into the labor market after World War II. In this connection, we shall see that the model furnishes an adequate framework for estimating functions of labor supply and demand.

In section 2, we will see that the hypothesis of perfect competition yields a very rich theory of wage setting, with a number of implications when the ensemble of the characteristics of jobs, especially working conditions, is taken into account. Differences that arise from hard working conditions are explained by the *hedonic theory of wages*, the premises of which were sketched by Adam Smith at the end of the eighteenth century, and have more recently been formalized by Rosen (1974). We will see in chapter 4 that wage differences linked to individual competence are explained by the *theory of human capital*, which rests on the idea that education leads individuals to become competent in ways that have value on the labor market. The foundations of this theory were laid by Becker (1964). According to the hedonic theory of wage, in exchange for a wage, a worker must carry out a set of tasks which may be more or less burdensome according to the speed at which he has to perform them, the work environment, the risk of accidents, and even the social prestige attached to that job. Adam Smith noted at the outset that workers with the same level of competence should be paid different wages if their working conditions are different. The hedonic theory of wages proposed by Rosen (1974) accounts for wage heterogeneity arising from these "compensating differentials." It shows that the mechanism of perfect competition provides reimbursement for the workers who hold the hardest jobs. In section 2 we present a simple setup derived from the model of Roy (1951), which shows how wage earners can choose among jobs with different degrees of arduousness offered in different competitive markets to which firms have free access by creating jobs adapted to the preferences of workers. This mechanism also allows workers, whose preferences are by nature heterogeneous, to choose how hard a job they are willing to take in view of the wage differentials created by

competition. These mechanisms also ensure that the allocation of workers over a range of jobs is socially efficient.

The third and last section of this chapter describes the competitive functioning of the labor market in a context where agents and jobs are heterogeneous. The fact is that for certain occupations the heterogeneity of the services traded is persistent and plays an important role. This holds particularly true of the markets for "superstars," whether they be athletes, artists, journalists, lawyers, doctors, scientists, or managers of large firms who dispose of specific talents that are hard or impossible to replicate. In such a setting, at the limit, each agent is unique in the sense that she possesses a characteristic that the others do not. So we must seek an explanation of how athletes of varying ability are allocated among different teams, how journalists of varying talent are allocated among different periodicals, and for that matter what kind of CEO arrives at the helm of what kind of company. We will see that the competitive functioning of this type of market may lead to steeply unequal compensation packages, which are nevertheless socially efficient inasmuch as they ensure an optimal allocation of talent.

1 THE COMPETITIVE EQUILIBRIUM

A market works according to the principles of perfect competition if agents are perfectly informed about the quality and the price of all the goods and services exchanged on that particular market. Another requirement for perfect competition is that all agents must be price takers. Under these hypotheses, perfectly competitive equilibrium is characterized by prices (including wages) that match supply and demand. We will start by presenting a simple model of perfect competition that allows us to shed light on the consequences of taxation and analyze the impact of shocks that affect labor supply and demand.

1.1 Perfect Competition with Identical Workers and Jobs of Equal Difficulty

The model of perfect competition assumes that a job and the wage it pays are determined when the demand put forth by firms is matched or met by the supply put forth by workers, both sides being price takers. Here we will illustrate the functioning of a market on which a perfectly homogeneous service is traded: every worker offers a service of the same quality, and the working conditions are the same everywhere. We will analyze the consequences of heterogeneity among working conditions and the quality of labor in the following sections.

1.1.1 Supply and Demand in a Simple Model of the Labor Market

Let us consider a market in which a representative firm produces a consumption good with a production function $F(L)$ where labor, denoted L, is the sole input. There is a large number of workers, all of whom supply a unit of labor and receive a wage w (expressed in units of the good produced) if they are hired. The welfare of a worker is evaluated using a utility function $u(R, e, \theta)$ with three arguments. Income R is equal to wage w when the worker is employed, and equal to 0 when he is not. For the sake

of simplicity, it is assumed that all income is consumed, so that there is no saving. Parameter e measures the effort (or the disagreeability) attached to each job. We assume that this disagreeability is identical for all jobs, and without any loss of generality, we will assume that parameter e is equal to 1 if there is a hire and equal to 0 if not. The parameter $\theta \geq 0$ represents the disutility (or the opportunity cost) of labor for the individual considered. In this model, all the jobs thus have the same "intrinsic" difficulty e, but individuals react differently to the difficulty of the tasks confronting them. Those with a low θ accept it more easily than those with a high θ. The cumulative distribution function of parameter θ will be denoted $G(.)$. Finally, in order to simplify, we will assume that an agent's utility function takes a linear form equal to the difference between the income and the opportunity cost of labor, or $u(R, e, \theta) = R - e\theta$.

In a competitive market, firms regard the wage as a given, and labor demand results from the maximization of profit $F(L) - wL$. It is thus defined by:

$$F'(L^d) = w \tag{3.1}$$

On the assumption that the marginal productivity of labor is decreasing ($F'' < 0$), labor demand is a decreasing function of the wage (see chapter 2 for a much fuller account of the theory of labor demand).

In addition, a worker with characteristic θ attains a level of utility equal to $w - \theta$ if she is hired, and 0 if she does not work. Consequently, only individuals whose opportunity cost θ is less than the wage decide to work. If we normalize the measure of the labor force to 1, then labor supply is equal to $G(w)$.

1.1.2 Equilibrium and Optimum

The functioning of the labor market is represented in figure 3.1, in which the quantity of labor is shown on the vertical axis and the wage on the horizontal axis. Labor demand is represented by the decreasing curve $L^d(w)$ and labor supply, equal to $G(w)$, is represented by an increasing curve passing through the origin. At labor market equilibrium,

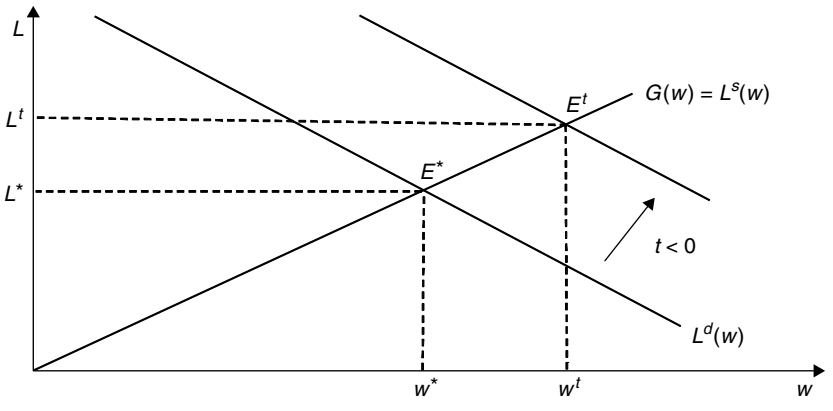


Figure 3.1
Market equilibrium with perfect competition.

supply is equal to demand. The equilibrium wage, at which labor demand and labor supply meet, is thus defined by the relation:

$$F'\left[G(w^*)\right] = w^* \tag{3.2}$$

and the equilibrium level of employment is equal to $G(w^*)$. Figure 3.1 shows that the labor supply and demand curves admit a sole intersection point E^*, the coordinates of which are the equilibrium wage $w^*$and the equilibrium level of employment $L^* = L^d(w^*) = G(w^*)$. Note that only individuals for whom the disutility of work θ is less than the equilibrium wage w^* decide to work. In the competitive equilibrium model, nobody is unemployed against his will: every worker who wishes to hold a job at the equilibrium wage w^* can do so. Those who choose not to are choosing not to participate in the labor market at all, and should be classified as "inactive" rather than unemployed.

One of the most striking results of microeconomic analysis is that the equilibrium of perfect competition yields a collective optimum. The reader can easily observe this well-known result in the model before us. More precisely, we can show that at market equilibrium, the allocation of individuals between employment and inactivity is efficient. To that end, let us consider an omniscient planner with the task of assigning workers to employment or inactivity so as to maximize the sum of individual utilities. This planner also sets consumption levels for both those in employment and those not participating. Let us assume that she decides to allot a quantity c of consumption goods to a working individual and a quantity z to a nonparticipant. Under these conditions, only workers whose opportunity cost θ verifies $c - \theta > z$ will agree to take jobs, while the rest will remain inactive. The planner's choice criterion is then written:

$$\Omega = \int_0^{c-z} (c-\theta)\mathrm{d}G(\theta) + \int_{c-z}^{+\infty} z\mathrm{d}G(\theta) \tag{3.3}$$

As well, the planner faces a resource constraint which simply states that the quantity of goods consumed cannot exceed the quantity produced. This resource constraint is written as follows:

$$F\left[G(c-z)\right] \geq cG(c-z) + z\left[1 - G(c-z)\right] \tag{3.4}$$

The program of our omniscient planner then comes down to choosing c and z in such a way as to maximize the criterion (3.3) under the resource constraint (3.4). If λ denotes the multiplier associated with the resource constraint, the Lagrangian of the planner's program is:

$$\mathcal{L} = \int_0^{c-z} (c-\theta)\mathrm{d}G(\theta) + \int_{c-z}^{+\theta} z\mathrm{d}G(\theta) + \lambda\left\{F\left[G(c-z)\right] - cG(c-z) + z\left[1 - G(c-z)\right]\right\}$$

The first-order conditions are obtained by canceling the derivatives of this Lagrangian with respect to c and z. After several simple calculations, we arrive at the two relations:

$$G(c-z) + \lambda\left\{G'(c-z)\left[F'(G(c-z)) - (c-z)\right] - G(c-z)\right\} = 0$$

$$1 - G(c-z) + \lambda\left\{-G'(c-z)\left[F'(G(c-z)) - (c-z)\right] - 1 + G(c-z)\right\} = 0$$

Note that only the difference $(c - z)$ appears in this system. Adding these last two equations member for member, we get $\lambda = 1$, which entails:

$$F'[G(c - z)] = c - z \tag{3.5}$$

Comparison of this equation to equation (3.2), which defines the competitive equilibrium, shows that the optimal value of the difference $(c - z)$ is equal to the equilibrium wage w^*. A perfectly competitive market thus yields the same allocation of resources that an omniscient planner would have chosen. In both cases, the level of employment is equal to $G(w^*)$ and only workers for whom $\theta < w^*$ hold jobs. At the competitive equilibrium, the allocation of individuals between employment and nonparticipation is efficient, for every worker takes up the occupation at which he is most productive. Workers whose opportunity cost θ is greater than marginal productivity $w^* = F'(L^*)$ remain outside the labor market, while all others enter it and find work. An omniscient planner with the task of assigning workers to employment or nonparticipation so as to maximize the sum of individual utilities would choose exactly the same allocation as the one that results from the competitive equilibrium.

The model of perfect competition is grounded in oversimplified hypotheses and is thus an imperfect representation of the functioning of many labor markets. Still, it is highly useful for analyzing the consequences of shocks, such as alterations in the tax regime, or demographic change, on wages and employment. Such shocks do in fact exert nontrivial effects, to the extent that labor demand and labor supply interact. The model of perfect competition allows us to understand such interactions, which are in fact similar in models of imperfect competition. Let us proceed to examine how the model of perfect competition makes it possible to grasp the effects of taxation on labor market equilibrium.

1.2 THE QUESTION OF TAX INCIDENCE

The fact that a tax is a charge upon the revenue of an agent (the payroll taxes paid by firms, for example) does not entail that the cost is borne by that agent. A firm might offset a rise in payroll taxes by lowering wages. In that case, the cost of labor to the firm remains the same, and it is the wage earners who finance the larger social security contributions by taking home smaller paychecks. The essential point about tax incidence is this: knowing who the *end* payer of the tax or the *end* recipient of the subsidy is. As we will see, the model of perfect competition enables us to answer that question; moreover it supplies predictions that are empirically pertinent.

1.2.1 WHO PAYS WHAT?

Let us consider a firm subject to a rate t of payroll tax on the net wage w. Its labor demand is defined by the equality $F'(L^d) = w(1 + t)$. When t is positive, it designates a tax paid by the firm; when t is negative, it designates a subsidy paid to the firm in the form, for example, of a reduction in social security contributions. Labor supply remaining equal to $G(w)$, the equilibrium wage on the labor market is always characterized by the equality of supply and demand which is now written:

$$L^d[w(1 + t)] = L^s(w) \tag{3.6}$$

Figure 3.1 illustrates the effect of a reduction in social security contributions ($t < 0$). Such a reduction corresponds to an upward shift in labor demand. Labor market equilibrium then goes from E^* to point E^t. We see that the upshot of this payroll tax reduction is a rise in both the wage and the level of employment. We see too that the respective amplitudes of these rises depend on the slopes of the curves of labor supply and demand.

This observation can be enhanced by differentiating both sides of relation (3.6) with respect to $(1 + t)$ and to w. After several calculations, we find that the elasticity of the net equilibrium wage with respect to $(1 + t)$, denoted η_t^w, is given by the formula:[1]

$$\eta_t^w = \frac{\eta_w^d}{\eta_w^s - \eta_w^d} \tag{3.7}$$

where η_w^s designates labor supply elasticity and $\eta_w^d < 0$ represents labor demand elasticity, taken here at point $w(1 + t)$. We saw in chapter 1 that under many circumstances labor supply has low elasticity. Let us take the extreme case of totally inelastic labor supply ($\eta_w^s = 0$). In our model, this situation arises when all individuals have the same parameter θ representing the opportunity cost of labor. Put another way, all individuals have the same reservation wage, denoted w_A, and they all offer an indivisible unit of labor for every wage that exceeds the reservation wage. For $w > w_A$, overall labor supply is then represented by a straight horizontal, the ordinate of which is the size of the active population, denoted N in figure 3.2. In this situation, we have $\eta_t^w = -1$, which means that any reduction in payroll taxes is fully passed on, in the form of a rise in the equilibrium wage that leaves the level of employment unchanged. This situation, portrayed in figure 3.2, is a good illustration of the main point regarding *fiscal incidence*: it is not the agent to whom the tax is charged (or the subsidy awarded) who is the real payer (or beneficiary). The equilibrium wage goes from w^* to w^{**} but the level of employment remains the same. When labor supply is inelastic, any lowering of payroll taxes meant in principle to aid the firm actually benefits the employee through a wage rise. In practical terms, then, knowledge of the elasticities of labor supply and demand proves to be of primary importance, since, as this example has just shown us, a policy of lowering payroll taxes with the aim of reducing the cost of labor in order to stimulate hiring may lead in the end to a wage rise that leaves the level of employment where it was.

In more general terms, knowledge of the elasticities of labor supply and demand makes it possible to calculate the impact of a change in payroll taxes on wages and employment. We know that on average the elasticity of labor supply at the extensive

[1] Differentiating the equality $L^d[w(1 + t)] = L^s(w)$ with respect to w and $(1 + t)$ we get:

$$\mathrm{d}w\left[(1+t)L^{d\prime}[w(1+t)] - L^{s\prime}(w)\right] = -\mathrm{d}(1+t)\left[wL^{d\prime}[w(1+t)]\right]$$

which may be written:

$$\frac{\mathrm{d}w}{w}\left[w(1+t)\frac{L^{d\prime}[w(1+t)]}{L^d[w(1+t)]} - \frac{wL^{s\prime}(w)}{L^s(w)}\right] = -\frac{\mathrm{d}(1+t)}{(1+t)}\left[w(1+t)\frac{L^{d\prime}[w(1+t)]}{L^d[w(1+t)]}\right]$$

$$\frac{\mathrm{d}w}{\mathrm{d}(1+t)}\frac{(1+t)}{w} = \frac{\left[w(1+t)\frac{L^{d\prime}[w(1+t)]}{L^d[w(1+t)]}\right]}{\left[\frac{wL^{s\prime}(w)}{L^s(w)} - w(1+t)\frac{L^{d\prime}[w(1+t)]}{L^d[w(1+t)]}\right]}$$

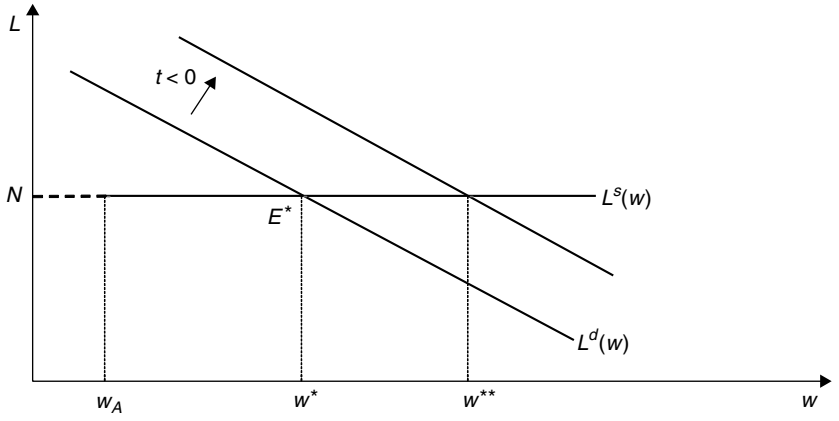


FIGURE 3.2
The effects of a reduction in payroll taxes with inelastic labor supply.

margin is on the order of 0.25 (table 1.4), while the elasticity of labor demand is of the order of -0.3 (chapter 2, section 2.2.1). This means that an increase in social security contributions that ex ante augments (i.e., at given net wage w) the cost of labor, equal to $w(1+t)$ by 1% leads to a wage variation of:

$$\eta_t^w = \frac{-0.3}{0.25 + 0.3} \simeq -0.55$$

Thus, when the cost of labor increases ex ante by 1%, the net wage shrinks by 0.55% and the cost of labor ex post (i.e., once the wage adjustment takes place) increases by 0.45%. Employment therefore shrinks by $0.45 \times 0.3 = 0.135$, since the elasticity of labor demand is equal to -0.3. For empirically pertinent average values, the model of perfect competition thus predicts that variations in payroll taxes have a strong impact on net wages, which move in the direction opposite to that of the payroll taxes. This impact on net wages can even be stronger in the case of low-skilled manpower, given that the elasticity of labor demand is of the order of -1 for this labor category. Here an increase ex ante of 1% in the cost of labor induces a net wage drop of 0.8% and a drop in employment of 0.2%. Bearing in mind that the elasticity of labor supply at the extensive margin of low-wage workers is higher (see chaper 1), the negative impact of tax rises on their net wages is damped [see equation (3.7)], but the negative impact on low-skilled employment is amplified.

The presence of a minimum wage changes these outcomes. To the extent that labor supply exceeds labor demand due to a minimum wage, the impact of payroll taxes on employment is entirely determined by changes in labor demand, for the same net wage. Under these conditions, an increase in payroll taxes leading to an ex ante rise of 1% in the cost of low-skilled labor entails a fall of 1% in the employment of low-skilled persons, since the elasticity of labor demand is of the order of -1 for this category of manpower.

1.2.2 Fiscal Incidence in Practice

Much research has been done on the incidence of payroll taxes. The earliest studies, which relied on correlations arising out of temporal series, yielded diverse results (Brittain, 1972; Feldstein, 1972; Holmlund, 1983). This early work did not allow researchers to correctly identify the impact of payroll taxes, changes in which may be correlated to events not observable by the econometrician. To properly identify the impact of payroll taxes, one must be in a position to analyze the behavior of comparable groups for whom the tax regime changes in different ways.

The contribution of Gruber (1997) adopts this method. He studied the incidence of a dramatic change in payroll taxation in Chile in 1981. Prior to this time, most social insurance programs in Chile were financed by a substantial payroll tax. In 1980 the average payroll tax rate for manufacturing firms was 30%, while the tax rate on workers averaged 12%. Then, in May 1981, Chile privatized its social security and disability insurance programs, as well as shifting the financing of most other social insurance programs from employer payroll taxes to general revenues. As a result, the average payroll tax rate for manufacturing firms dropped to 8.5% by 1982. Gruber shows that the diminution in payroll taxes led to an increase in the net wage of the same amount and had no impact on the level of employment. This result is compatible with a setting in which the elasticity of labor supply is close to zero. In this case, the elasticity of the net wage with respect to the payroll tax is equal to -1, and employment does not depend on the amount of the payroll tax. Anderson and Meyer (2000) have obtained results of the same kind in their study of the consequences of a change in the financing of unemployment insurance in the state of Washington in the middle of the 1980s.

1.3 The Effect of a Shock on Labor Supply

The model of perfect competition indicates that the relation between wages and employment depends on the characteristics of labor demand and labor supply. An increase in labor demand may have no impact on employment if labor supply is totally inelastic: in that case, it is the wage that rises, while the level of employment remains the same. Conversely, the impact of a change in labor supply on employment and wages depends on the properties of labor demand.

This means that wages and employment are intrinsically determined by interactions between demand and supply. In other words, correlations between wage and employment can originate from changes in demand, in supply, or in both. Accordingly, a negative correlation between wage and employment cannot be interpreted as reflecting a movement along the labor demand curve unless the researcher has ensured that there are changes in labor supply alone. The only way to know the slope of labor demand is to detect changes in labor supply that do not move the labor demand curve. Symmetrically, the slope of labor supply can only be identified from changes in labor demand that do not affect the labor supply curve. Events that move either labor supply or labor demand, but not both, are not easy to detect in the real world, inasmuch as most of the shocks that come to mind might very likely affect both supply and demand. It is, however, essential to implement empirical strategies that utilize such events, if we are to rigorously identify labor demand and labor supply elasticities.

In what follows, we present the contribution of Acemoglu et al. (2004), who apply this strategy to estimate the elasticity of labor demand. Their paper is presented in some detail for two reasons. First, it is a good illustration of how interactions between supply and demand can be used to estimate labor demand. Second, it is a good introduction to the so-called instrumental variable approach, which is widely used in empirical labor economics and which will be studied in more detail in chapter 4. The main results of this contribution, which are presented below, can be replicated using the database and the program available at www.labor-economics.org.

1.3.1 Theoretical Mechanisms

In 1940 between 40 and 55% (depending on the state) of all eligible males aged 18 to 44 years across the United States were mobilized for World War II, and 73% were deployed overseas. This tremendous shock on the supply of labor from males was partly compensated by an inflow of women into the labor market: the employment rate of women increased from 24% in 1930, to 28% in 1940, and to 34% in 1950. In fact, the decade of the 1940s saw the largest proportional rise in female labor force participation during the twentieth century.

This positive demographic shock due to an influx of supplementary population into the labor market leads to an upward shift of the curve of labor supply (suppose for example that function $G(w)$ has been multiplied by a coefficient greater than 1). In figure 3.3 the positive demographic shock identified by the symbol $\Delta N > 0$ shifts the equilibrium of the labor market from E^* to point E^{**}. The econometrician can profit from this shift to estimate the curve, and thus the elasticity, of labor demand.

To illustrate more precisely the impact of an increase in the amount of labor supplied by women, we can make use of a simple model with male labor and female labor. Consider a Cobb-Douglas function homogeneous of degree 1: $Y = AK^{\alpha}L^{1-\alpha}$, where K is capital and L labor. Now, consider that labor has two components, male labor M and female labor F, which are combined in production so that they display a constant

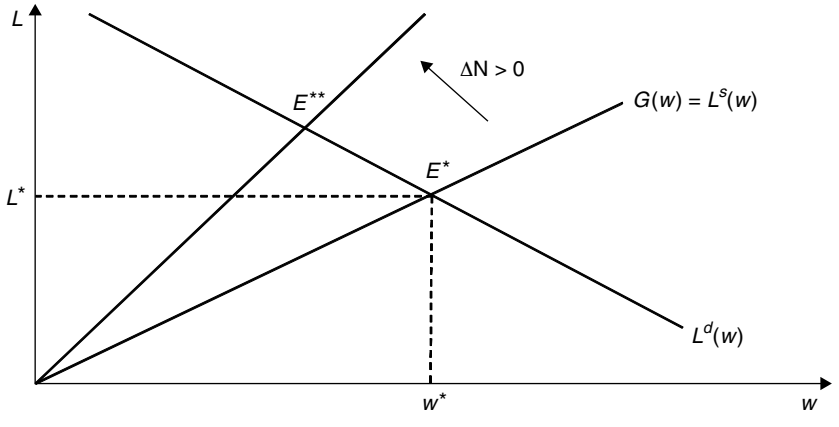


Figure 3.3
The effects of a demographic shock.

elasticity of substitution: $L = \left[(1-\lambda)(a_M M)^{\frac{\sigma-1}{\sigma}} + \lambda (a_F F)^{\frac{\sigma-1}{\sigma}}\right]^{\frac{\sigma}{\sigma-1}}$, where a_M and a_F are positive factor-augmenting productivity terms, σ is the elasticity of substitution between female labor and male labor, and λ is a shared parameter. Integrating this labor input into the production function gives a *nested CES function*:

$$Y = AK^{\alpha}\left[(1-\lambda)(a_M M)^{\frac{\sigma-1}{\sigma}} + \lambda (a_F F)^{\frac{\sigma-1}{\sigma}}\right]^{\frac{(1-\alpha)\sigma}{\sigma-1}}, \quad \sigma > 0, \quad a_i > 0$$

With W^i designating the unit wage cost of labor input i, equalizing wages with the marginal products of each labor input gives:

$$W^F = (1-\alpha)\lambda a_F A K^{\alpha}(a_F F)^{-\alpha}\left[(1-\lambda)\left(\frac{a_M M}{a_F F}\right)^{\frac{\sigma-1}{\sigma}} + \lambda\right]^{\frac{(1-\alpha)\sigma}{\sigma-1}-1}$$

and

$$W^M = (1-\alpha)(1-\lambda) a_M A K^{\alpha}(a_M M)^{-\alpha}\left[(1-\lambda) + \lambda\left(\frac{a_F F}{a_M M}\right)^{\frac{\sigma-1}{\sigma}}\right]^{\frac{(1-\alpha)\sigma}{\sigma-1}-1}$$

After log-linearizing these equations and holding capital constant, as it should be in the short run, the elasticity of female wages to female employment (which is the inverse elasticity of female labor demand) is:

$$\frac{\partial \ln W^F}{\partial \ln F} = -(1-s^m)\alpha - s^m \frac{1}{\sigma} \tag{3.8}$$

and the cross elasticities of male wage to female employment are:

$$\frac{\partial \ln W^M}{\partial \ln F} = -(1-s^m)\alpha + (1-s^m)\frac{1}{\sigma} \tag{3.9}$$

where $s^m = \left[(1-\lambda)(a_M M)^{\frac{\sigma-1}{\sigma}}\right] / \left[(1-\lambda)(a_M M)^{\frac{\sigma-1}{\sigma}} + \lambda (a_F F)^{\frac{\sigma-1}{\sigma}}\right]$ is the share of male labor in overall labor input and $(1-s^m)$ is the corresponding share of female labor. Equation (3.8) shows that when female employment increases, the female wage always decreases when capital is held constant. This is an illustration of the law of demand studied in chapter 2. Equation (3.9) shows that the impact of female employment on male wages is ambiguous. It is negative if σ, the elasticitity of substitution between male and female labor, is large enough, that is, larger than $1/\alpha$, where α is the share of capital in total production costs. When the elasticity of substitution between male and female labor is sufficiently large, the employment of women easily replaces male employment, which entails that growth in female employment provokes a contraction in the demand for labor by men; and that causes men's wages to fall. In the opposite case, where men's labor is not easily replaceable by that of women, a swell in the employment of women leads to increased demand for labor by men; and that causes men's wages to rise.

1.3.2 THE SUPPLY SHOCK

The idea of Acemoglu, Autor, and Lyle is to use the exogenous shock on the labor supply of women induced by the mobilization of men during World War II to estimate the elasticity of labor demand and the elasticity of substitution between working men and working women. In fact, women's participation increased steadily during the twentieth century for various reasons, especially related to the organization of the family, the education of children, and changing tastes for work. But the shock of World War II induced changes in female labor supply that had nothing to do with these demand-side factors. This increase was due to the lack of men in the labor market. What this meant was that the labor supply of women suddenly shifted leftwards, as in figure 3.3, and this shift lasted well after men came back from the war. Indeed, the increase in female employment rates did not recede in the aftermath of the war, as women got used to working and firms got used to employing them. As figure 3.3 shows, with labor demand left unchanged, and independently of any factor that could have influenced wages in other ways (such as an increase in tastes for work), this shift should have induced an increase in female employment and a decrease in female wages, hence revealing the slope (or elasticity) of the labor demand curve. Some evidence suggests that this is probably what happened. Figures 3.4 and 3.5 show that the mobilization rate across states in 1940 is indeed positively correlated with the observed changes in female weeks worked per year between 1940 and 1950, and it is negatively correlated with the change in the female weekly wage between 1940 and 1950. Such correlations cannot be observed between 1950 and 1960. This suggests that the mobilization rate had an impact on female employment and wages during the war, with some lasting effect.

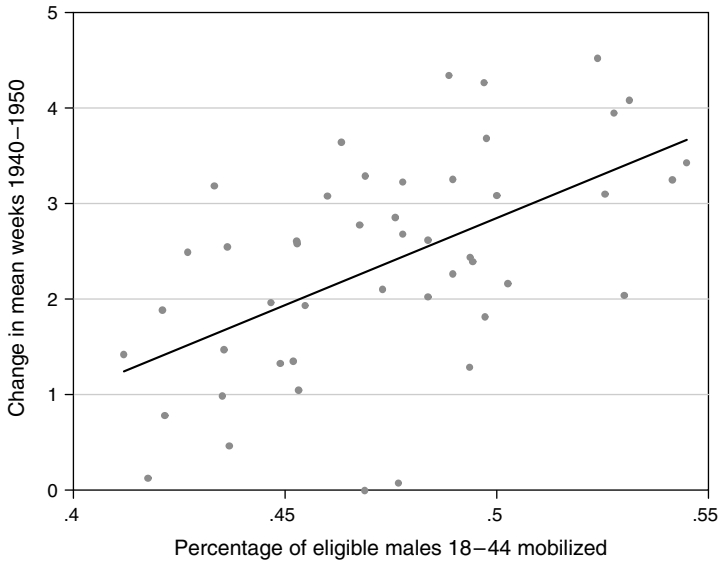


FIGURE 3.4
State World War II mobilization rates and change in female mean weeks worked per year in 1940–1950 and 1950–1960.

Source: Acemoglu et al. data set (2004).

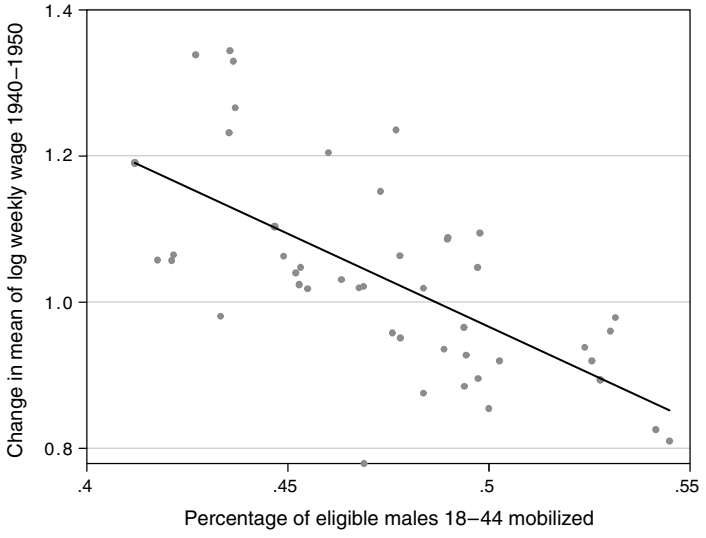


Figure 3.5
State World War II mobilization rates and change in mean female weekly wages, 1940–1950.

Source: Acemoglu et al. data set (2004).

Heterogeneity in Rates of Mobilization

The mobilization started before the United States actually went to war with Germany, Italy, and Japan in 1941. Soon after the defeat of France, the Selective Service Act of October 1940 initiated a mandatory draft of men, based on a series of lotteries to determine the order in which registrants were called to active duty. Exemptions were available based on marital status, fatherhood, medical disabilities, and skills needed for the civilian war effort (such as farmers for food production, leading to farm states showing lower mobilization rates). This left wide discretion to the members of local draft boards, which also led to lower mobilization rates in states with higher percentages of foreign-born residents (German, Japanese, Italian). Additionally, as military units were segregated at that time and the number of black units was low, the states with higher percentages of blacks also showed lower rates of enlistment. Table 3.1 presents a regression of the mobilization rate on a range of explanatory variables. Clearly the effect of ethnic origin is quite large compared to other variables: a segment of German-born folk in the population larger by 1 percentage point led to a mobilization rate lower by 3.19 percentage points, while a segment of farmers in the population higher by 1 percentage point led to a mobilization rate lower by 0.15 percentage points. Ethnic origin and farming, along with the age structure, race, and schooling, explain a substantial part of the cross-state variation in mobilization. About 30 to 40% corresponds to idiosyncratic variations.

Mobilization Rates and Labor Supply by Women

Now, how well are mobilization rates correlated with variations in female weeks worked? The authors pooled the data of 1940 and 1950 and regressed the following equation:

$$y_{ist} = \delta_s + \gamma d_{1950} + \mathbf{x}_{ist}\boldsymbol{\beta}_t + \varphi m_s d_{1950} + \varepsilon_{ist} \qquad \textbf{(3.10)}$$

TABLE 3.1

1940 state-level determinants of World War II mobilization rates (N = 47 states). The first column displays the mean value of each variable. The second column displays the results of the regression of the mobilization rate. Standard errors in parentheses.

	Mean	Mobilization rate
Share farmers	.15	−.15 (.05)
Share nonwhite	.10	−.01 (.05)
Average education	8.89	.02 (.01)
Share aged 13–24	.42	.25 (.34)
Share aged 25–34	.31	.15 (.48)
Share German origin	.007	−3.19 (.89)
Share Japanese or Italian origin	.010	1.70 (.52)
Share married	.50	−.10 (.17)
Share fathers	.47	.08 (.13)
R^2		.78

Source: Acemoglu et al. (2004, table 4).

where y_{ist} is weeks worked by woman i residing in state s in year t (1940 or 1950); δ_s denotes a full set of state-of-residence dummies; d_{1950} is a dummy variable equal to 1 in 1950 and to 0 in 1940; $\mathbf{x}_{ist}$ denotes other covariates including state or country of birth, age, marital status, race, share of farmers and nonwhites, and average schooling in the state in 1940; ε_{ist} is an error term. Although not shown in this equation (in order to lighten the notation), age, race, and marital status are interacted with a 1950 dummy to allow the returns to these variables to differ by a decade. The coefficient of interest is φ, which corresponds to the interaction term between the 1950 dummy and the mobilization rate m_s in state s. This variable measures whether states with higher rates of mobilization for World War II experienced a greater increase in female employment from 1940 to 1950.

This equation is like a difference-in-differences analysis of female employment before and after the mobilization, with the "treatment" intensity varying across groups depending on the mobilization rate. This rate cannot be included alone in this equation because it is constant over time (it is measured only in 1940, since the database only contains two years—1940 and 1950) and would thus be redundant with respect to the state dummy, which equals 1 for both decades. For white females, OLS gives this result: a 10 percentage point rise in the mobilization rate increased weeks worked by 1.1 annually in 1950 compared to 1940 (see column 1 of table 3.2, and note that the mobilization rate goes from 0 to 1), and this coefficient is very significant and stable when controlled individually for age, marital status, and state or country of birth to account for possible composition effects (notably due to migration across states) over the period (see column 2 of table 3.2). Moreover, as we saw in table 3.1, mobilization rates were lower in states where the share of German-born people was higher (a noneconomic factor) but also lower in agricultural states and in states where the share of nonwhite

Table 3.2

Impact of World War II mobilization rates on labor supply, 1940–1950. Dependent variable: Annual weeks worked. Column (1): no covariates; (2) with covariates (marital status, age, state of birth); (3) as in (2) plus two control variables: 1940 male share farmers × 1950 dummy and 1940 male share nonwhites × 1950 dummy; (4) as in (3) plus one additional control variable: 1940 male share average education × 1950 dummy (standard errors in parentheses).

	(1)	(2)	(3)	(4)
	White females (N = 530,026)			
$ms.d_{1950}$	11.2 (1.9)	9.9 (2.1)	10.6 (2.7)	8.5 (2.4)
	All females (N = 585,745)			
$ms.d_{1950}$	13.9 (1.8)	9.1 (2.4)	10.2 (2.6)	8.3 (2.4)
	White males (N = 441,343)			
$ms.d_{1950}$	.5 (6.0)	3.6 (5.4)	−6.6 (7.1)	−6.6 (7.6)

Source: Acemoglu et al. (2004, table 5).

people was higher. These latter factors have economic consequences and could also have had a direct impact on female employment. If that were the case, the coefficient measured between female employment and mobilization would be biased due to simultaneity. Hence it is necessary to verify that after including these factors, the coefficient of the mobilization rate is stable. Actually it is: when adding the share of male farmers in 1940, of male nonwhites in 1940 (column 3 of table 3.2), and even the average education level in 1940 (column 4 of table 3.2), the results remain broadly unchanged. The second row of table 3.2 also shows that the results are similar when considering all females instead of white females only.

So the mobilization rates are very significantly correlated with the rise in female employment, even when controlling for the influence of economic components correlated with the mobilization rate. But are we sure that the growth in female employment reflects a shift in labor supply and not a shift in labor demand? Maybe high-mobilization states experienced a higher demand for labor (of both male and female workers) in 1950 for reasons not accounted for in the model. In that case, we should observe a similar positive correlation between mobilization and male employment. Maybe high-mobilization states were also states where in 1950 men were supplying less labor because veterans were experiencing difficulties reentering the labor markets, which could have induced an increase in demand for female labor. In that case, we should observe a negative correlation between mobilization and male employment. As the last row of table 3.2 shows, neither of these assumptions can be verified: the coefficient rate on male weeks worked is statistically insignificant.

These results are also stable when instead of directly controlling economic factors correlated with the mobilization rate, the latter is *instrumented* by the noneconomic factors, notably the age structure of the population and the share of German-born people.[2] Why use "instruments" for the mobilization rate? As stated above, the mobilization rate may be partly driven by economic factors that also influenced female or male employment in the states. In that case the coefficient of the mobilization rate would not

[2] The instrumental variable method is presented in more detail in chapter 4.

reflect a causal relationship. A way to avoid this problem is to find instruments, that is, variables correlated to the mobilization rate (this can easily be checked) but which are not correlated to the error term of equation (3.10) or, in other words, which have no partial effect on female / male employment in general, and which are not correlated to the other explanatory variables of employment (this is called the *exclusion restriction*, which is usually more difficult to prove). A good candidate for such instrumentality is the share of German-born residents within states, recorded in table 3.1. Another variable could be the share of young males in the population. This instrumental variables method is implemented with a *two-stage least squares* estimation of the following system of two equations:

$$y_{ist} = \delta_s + \gamma d_{1950} + \mathbf{x}_{ist}\boldsymbol{\beta}_t + \varphi m_s d_{1950} + \varepsilon_{ist} \quad \textbf{(3.11)}$$

$$m_s = \alpha_0 + \mathbf{z}_s\boldsymbol{\alpha}_1 + v_s \quad \textbf{(3.12)}$$

where the first equation is the same as equation (3.10) except that the mobilization rate m_s is defined by its predicted value determined by the second equation, where $\mathbf{z}_s$ denotes the vector of the instruments and v_s is an error term. Results are shown in table 3.3: the impact of mobilization on labor supply is of similar magnitude, although estimates are less precise with this approach (standard errors are larger).

A final test is to reproduce this table for the years 1950 and 1960. Perhaps the mobilization rate merely captures secular cross-state trends in female employment, not linked to the war. In that case, the coefficients should be significant and of the same sign and magnitude as for the years 1940 and 1950. The authors show that this is not the case, meaning that the cross-state growth of female labor force participation was correlated with the mobilization rate only during the decade of the war and its immediate aftermath.

TABLE 3.3

IV estimates of the impact of World War II mobilization rates on labor supply, 1940–1950. Dependent variable: annual weeks worked. Each column is from a separate pooled 1940 and 1950 micro data 2SLS regression of weeks worked by female or male state of residence on the instrumented World War II state mobilization rate interacted with a 1950 dummy, year main effects, and dummies for age, marital status, state of residence, and state/country of birth. All individual variables, aside from state of residence/birth, are also interacted with a 1950 dummy. Instruments for the mobilization rate are the fraction of males aged 13–44 in 1940 who are German-born or who are in the listed age categories. All models are weighted by census sampling weights (standard errors in parentheses).

	White females (N = 530,026)		White males (N = 441,343)	
	(1)	(2)	(3)	(4)
$m_s d_{1950}$	13.19 (5.49)	11.42 (3.97)	−17.00 (13.98)	−.04 (11.94)
	First-stage coefficients			
1940 male share ages 13–24		.27 (.15)		.44 (.15)
1940 male share ages 25–34		−.22 (.25)		−.20 (.21)
1940 male share German	−1.83 (.39)	−1.33 (.46)	− 2.03 (.38)	−1.30 (.41)
p−value (first stage)	.00	.00	.00	.00

Source: Acemoglu et al. (2004, table 7).

We see that the authors analyzed in great detail the relationship between the mobilization rate and female employment to verify to the maximum possible extent their key identifying assumption that the war induced an *exogenous* shift in female labor supply, from which elasticities can be estimated.

1.3.3 THE ELASTICITIES OF FEMALE LABOR DEMAND

We may now focus on the impact of the variation in the labor supply of women between 1940 and 1950 induced by the rate of mobilization. First, the impact of women's labor supply on their wages may be analyzed on the basis of equation:

$$\ln W_{ist} = \delta_s + \gamma d_{1950} + \mathbf{x}_{ist}\boldsymbol{\beta}_t + \chi \ln F_{st} + u_{ist} \tag{3.13}$$

where W_{ist} is the wage of woman i in state s at date t, F_{st} is female labor supply measured by the number of average weeks worked in state s at time t. Other variables are the same as in equation (3.10). The female labor supply F_{st} can be instrumented by the mobilization rate using an equation of type (3.10) in the first stage:

$$F_{st} = \delta_s + \lambda d_{1950} + \mathbf{x}_{st}\boldsymbol{\rho}_t + \varphi m_s d_{1950} + \varepsilon_{ist} \tag{3.14}$$

where $\mathbf{x}_{st}$ includes the state's female age structure, its share of farmers, its share of nonwhite people, and its average education.

Column 1 of table 3.4 shows the results of the estimation of equation (3.13) with OLS, whereas column 2 displays the result of the two-stage least square estimates of the system of equations (3.13) and (3.14). The estimates of χ using OLS, displayed in column 1, are biased towards zero due to simultaneity (presumably because female employment increased relatively more in states with greater demand for female labor). When instrumented with the mobilization rate, the estimate finds that a one-week increase in female labor supply is associated with a 12.4% decline in female weekly earnings. As predicted by the model of labor demand presented in chapter 2, demand for female labor decreases with their wage.

In order to analyze with greater precision the impact of the augmentation of women's labor supply, it is useful to estimate the structural model explored in the

TABLE 3.4

OLS and IV estimates of the impact of female labor supply on log weekly earnings, 1940–1950. Dependent variable: log weekly earnings (standard errors in parentheses).

	White females (N = 69,335)	
	(1) OLS	(2) Two-stage least squares
F	−.002 (.011)	−.124 (.029)
		First-stage coefficients
$m_s d_{1950}$		10.22 (1.81)

Source: Acemoglu et al. (2004, tables 3 and 9).

previous section. This model can be estimated by regressing the log of wages onto the log of employment with the equation:

$$\ln W_{ist} = \delta_s + \gamma d_{1950} + \varphi f_i + \mathbf{x}_{ist}\boldsymbol{\beta}_t + \chi \ln\left(\frac{F_{st}}{M_{st}}\right) + \eta f_i \ln\left(\frac{F_{st}}{M_{st}}\right) + u_{ist}$$

where the sample now includes all individuals (male and female), f_i is a dummy equal to 1 if individual i is a female and to zero otherwise; $\ln(F_{st}/M_{st})$ is the log ratio of female to male labor supply (in weeks) in the state of residence on average over the period (1940 or 1950). Each of the individual and aggregate state controls included in $\mathbf{x}$ is permitted to affect male and female earnings differentially by gender and decade. The labor supply measure $\ln(F_{st}/M_{st})$ is instrumented by the state mobilization rate. There are two coefficients of interest in this equation, χ and η. According to equations (3.8) and (3.9), coefficient χ corresponds to the term $-(1-s^m)\alpha$, which measures the common effect of variation in female labor supply on both male and female earnings (note that $\frac{\partial \ln W^M}{\partial \ln F}$ holds M constant, so $\frac{\partial \ln W^M}{\partial \ln F} = \frac{\partial \ln W^M}{\partial \ln(F/M)}$). As for η, this is the differential effect of female labor supply on female wages. Hence it would correspond to the inverse elasticity of substitution $\frac{1}{\sigma}$. Altogether, an estimate of the inverse elasticity of female labor demand is given by $\chi + \eta$.

Results are presented in table 3.5, which shows that increased labor supply from females reduces female earnings as the theoretical model predicted: a 10% increase in relative female labor supply reduces female wages by 7 to 8% (summing up the first two lines of the table), which corresponds to an own-labor demand elasticity of -1.2 to -1.5 (fourth line of the table). Compared to the labor demand elasticities reported in chapter 2, this result suggests that female labor demand is quite sensitive to wages. Moreover a 10% increase in female labor supply lowers female wages relative to male wages by 3 to 4% (second line of the table); this corresponds to an elasticity of substitution of 2.4 to 3.2, which is also high. Acemoglu et al. verify that these results are compatible with the restrictions of the theoretical model, where equation (3.9) entails that $\hat{\chi} + \hat{\eta} =$

TABLE 3.5

IV estimates of the impact of female / male labor supply on log weekly earning, 1940–1950. Dependent variable: log weekly earnings (whites) (standard errors in parentheses).

	White females / males	
	(1)	(2)
$\ln\left(\frac{F}{M}\right)$	−.51 (.11)	−.25 (.20)
$f.\ln\left(\frac{F}{M}\right)$	−.31 (.13)	−.42 (.19)
Estimated σ	3.18	2.37
Estimated σ_F	−1.21	−1.48
	First-stage coefficients	
$m_s d_{1950}$	1.56 (.19)	1.14 (.30)
Includes share of farmers, share of nonwhites, and education	No	Yes

Source: Acemoglu et al. (2004, table 10).

$-(1-s^m)\alpha - s^m\frac{1}{\sigma}$. Knowing the values (which can be observed directly) of s^m and α, respectively equal to 0.82 and 0.33, it is possible to calculate the value of $1/\sigma$ compatible with $\hat{\chi} + \hat{\eta}$ and to verify that it is equal to $\hat{\eta}$. This exercise indicates that the estimates are indeed compatible with the restrictions of the theoretical model.

1.4 Other Evidence on the Impact of Massive Shocks

A number of other events that produced massive shocks on either side of the markets have confirmed that wages tend to react in ways consistent with the predictions of the perfect competition model.

As shown in chapter 11, section 3.3.2, certain exceptional flows of migration, most often due to political events, like the Cuban immigration to Miami in May 1980 (Card, 1990) and immigration to France in the wake of Algerian independence in 1962 (Hunt, 1992), amount to "natural experiments" that induce exogenous shocks on labor supply.

Demand-side shocks can also be very severe. In 1968 oil was discovered in Prudhoe Bay, Alaska, and the reserves were estimated to exceed 10 billion barrels. At that time, Alaska had a very small economy, with about 100,000 jobs in total in 1970, and a disproportionately young, male, and migrant workforce. Soon the oil companies developed a project to transport Alaskan oil to the U.S. mainland by building a pipeline from Prudhoe Bay in the north to the southern Alaskan port of Valdes, a distance of 1,300 km (800 miles). The Trans-Alaska Pipeline System (TAPS) was built between 1974 and 1977 and became one of the most expensive privately financed projects in U.S. history. It added some 50,000 workers each summer to the local labor market until the completion of the project, only a few of whom were Alaskan residents. This shock was only temporary, though. After the pipeline was built, the operating company quickly reduced its workforce to a small maintenance team. Carrinton (1996) shows that the average monthly earnings in Alaska (statewide) went from $2,600 at the end of 1973 to $4,100 by the end of 1976, a 56% increase, and went back to pre-TAPS levels as early as 1979, immediately after the end of the project. This shock drove wages far beyond construction-industry levels because many workers elsewhere in the market quit lower-paid jobs in order to benefit from the project. Hence labor supply decreased in the other sectors of the economy, boosting wages everywhere. These "natural experiments" show that the labor market can be very flexible and absorb large shocks in a relatively short time, as long as wages can adjust.

More generally, the model of perfect competition constitutes a useful tool for evaluating the impact on the labor market of shocks such as modifications of the tax regime or changes in labor supply and demand. As we will now see, this model also helps to account for wage differentials.

2 COMPENSATING WAGE DIFFERENTIALS AND THE HEDONIC THEORY OF WAGES

The previous section describes how a labor market would function if labor services were all perfectly homogeneous and the work were equally arduous no matter what job one held. In reality, there is an extremely wide range of working conditions across all jobs.

Perfect competition in the labor markets ought to lead to a wage heterogeneity, inasmuch as some jobs are harder to do than others and some suppliers of labor are more willing to accept hardship than others. Perfect competition would ensure that these differences were compensated for by wage differentials. This is the essence of the *hedonic theory of wages*. Equilibrium is still identified as a social optimum, and any measures aimed at reducing the difficulty of jobs do not ameliorate welfare.

2.1 A Simple Model of Compensating Wage Differentials

We begin by studying an equilibrium model of the labor market where jobs are arduous to varying degrees and workers also vary in their willingness to tolerate hard labor. In this setting, the equilibrium of perfect competition leads to an optimal allocation of resources, with those workers whose tolerance for hardship is greatest holding the hardest jobs and receiving higher wages in return.

2.1.1 Wages and the Difficulty of Jobs

Let us now introduce heterogeneity among jobs arising from the difficulty of the work to be done. To that end, we tangibly alter the way the production sector is formalized in the previous model: we now assume that there exists a continuum of jobs, each requiring one unit of labor but a different level of effort $e > 0$. This effort variable is a synthetic measure of the difficulty of jobs and so covers a number of dimensions like accident risk, hours of work, environment, and the advantages, whether in kind or in status, that flow from holding a particular job. Strictly speaking, e should thus be a vector with as many coordinates as there are characteristics to any job, but for the sake of simplicity, we reduce heterogeneity to a single dimension. Various aspects of the actual content of jobs will be examined in greater detail in section 2.2, in which we present the relevant empirical work.

The productivity of every sort of job is an increasing and concave function of effort, or $y = f(e)$ with $f'(e) > 0, f''(e) < 0$, and $f(0) = 0$. Productivity y here corresponds to production *net* of any costs occasioned by employment, except wages. For example, if we interpret e as a measure of industrial accident risk, it is generally possible to reduce these risks by reducing the intensity of work or by making outlays that achieve the same result. In either case, jobs that offer lower risk have less productivity in our model. As previously, we assume that the utility function of an agent takes the linear form $u(R, e, \theta) = R - e\theta$, where θ measures aversion to effort, and that effort e is strictly positive when the worker is employed and amounts to 0 when he is not participating.

Let us assume that every firm may be thought of as an occupational slot requiring one unit of labor with its own particular degree of effort. Let us assume further that there is a market for each of the kinds of job that correspond to each of these degrees of effort. In a setting of perfect competition, entrepreneurs keep on entering all markets until, for every type of work, profits fall to zero. If $w(e)$ denotes the equilibrium wage that applies to jobs that demand effort e, then wage equals productivity and we have $w(e) = f(e)$. A worker with information about all jobs at her disposal, and enjoying perfect mobility, is able to "visit" different markets and choose the job that gives her the greatest satisfaction. If she chooses a job in which effort equals e, she will receive wage $f(e)$. Hence the problem for a worker of type θ consists of selecting a value of effort

that maximizes her satisfaction $u[f(e), e, \theta] = f(e) - e\theta$. The first-order condition of this problem, necessary and sufficient as a consequence of the concavity of function f, gives:

$$f'(e) = \theta \Leftrightarrow e = e(\theta) \tag{3.15}$$

Equation (3.15) indicates that an agent chooses the job in which the marginal return to effort $f'(e)$ is equal to the disutility θ that it gives rise to. As $f'(e)$ is decreasing with e, optimal effort $e(\theta)$ diminishes with parameter θ measuring aversion to effort. Given that the equilibrium wage received by a worker of type θ amounts to $w[e(\theta)] = f[e(\theta)]$, the counterpart of tough jobs is a "compensating" wage differential, since wages increase with effort.

An additional requirement is to ensure that the participation constraint $u(w, e, \theta) \geq u(0, 0, \theta) = 0$ is met. This constraint signifies that the worker accepts a job if doing so makes her situation preferable to nonparticipation (where $R = e = 0$). When the effort function satisfies relation (3.15), we have $u(w, e, \theta) = f(e) - ef'(e)$. The latter quantity is positive, since function f is concave and thus the participation constraint is met. A further requirement is to ensure that relation (3.15) does indeed define positive values of effort. The concavity of function f entails that $e(\theta) > 0$ for values of θ such that $\theta < f'(0)$. Consequently, individuals with "weak" aversion to effort, that is, those for whom $\theta < f'(0)$, do participate in the labor market while the rest stay home. The size of the active population is thus equal to $G[f'(0)]$, where $G(.)$ still denotes the cumulative distribution function of parameter θ.

The relation between effort and wage is illustrated graphically in figure 3.6, which represents the choices of two types of worker. Type θ^+ is characterized by a stronger aversion for effort than type $\theta^- < \theta^+$. The effort is on the horizontal axis and the wage

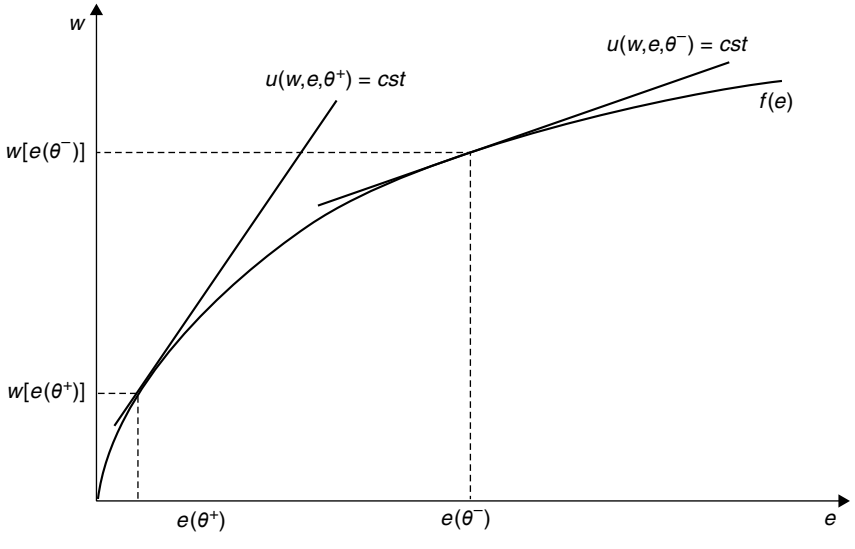


FIGURE 3.6
The hedonic theory of wages.

on the vertical axis. An indifference curve—which, let us recall, is the set of points (e, w) for which an individual obtains the same level of utility—is represented for both types of worker. The indifference curves are straight lines with slope θ. For given θ, an upward shift of the indifference curve corresponds to increased satisfaction. Hence each worker chooses a level of effort e such that one of her indifference curves is tangent to $f(e)$. In consequence, individuals with a strong aversion to effort choose low-effort jobs with correspondingly low wages. More generally, at equilibrium wages are given as a function of the θ type of each individual, according to the formula $w(\theta) = f\left[f'^{-1}(\theta)\right] = h_d(\theta)$. The h_d function is called the *hedonic wage function*. It gives the equilibrium value of the wage of a worker in line with that worker's characteristics.

In our model, scrutiny of figure 3.6 shows that all individuals of type $\theta > f'(0)$ prefer not to participate in the labor market, for they have indifference curves that are steeper at the origin than the slope of function $f(e)$. Individuals whose aversion to effort is too great, such that $\theta > f'(0)$, decide not to work.

2.1.2 Normative Implications of the Hedonic Theory of Wages

According to the hedonic theory of wages, the mechanisms of perfect competition allow workers to choose from a range of working conditions, with wage differentials "compensating" for the greater difficulty of some jobs. Moreover, competitive equilibrium allocations are efficient, furnishing each worker with an income $w[e(\theta)] = f[e(\theta)]$ and inducing a level of employment $G'(\theta)$ on the job market of type θ. This means that each worker is engaged in the task for which the difference between what he produces and the disutility that he undergoes is greatest. This result emerges if we look at the problem of a planner assigning workers to different jobs in such a way as to maximize the sum of utilities. For each worker with characteristic θ, such a planner would choose effort $e(\theta)$—potentially equal to 0, in which case the worker with characteristic θ does not participate in the labor market—and the consumption of good $c(\theta)$. The choice criterion of the planner then is written:

$$\Omega = \int_0^{+\infty} [c(\theta) - \theta e(\theta)]\, \mathrm{d}G(\theta)$$

The planner is moreover faced with a resource constraint stating simply that the quantity of goods consumed cannot exceed the quantity produced. It is written:

$$\int_0^{+\infty} f[e(\theta)]\, \mathrm{d}G(\theta) \geq \int_0^{+\infty} c(\theta) \mathrm{d}G(\theta) \tag{3.16}$$

The program of the omniscient planner is therefore to choose $e(\theta)$ and $c(\theta)$ in such a way as to maximize criterion Ω under the resource constraint. Let us denote by λ the multiplier associated to the resource constraint; the Lagrangian of the planner's program is written:

$$\mathcal{L} = \int_0^{+\infty} [c(\theta) - \theta e(\theta)]\, \mathrm{d}G(\theta) + \lambda \left\{ \int_0^{+\infty} f[e(\theta)]\, \mathrm{d}G(\theta) - \int_0^{+\infty} c(\theta) \mathrm{d}G(\theta) \right\}$$

The first-order conditions are obtained by canceling the derivatives of this Lagrangian with respect to $e(\theta)$ and $c(\theta)$. After several simple calculations, we arrive at the two following relations:

$$\frac{\partial \mathcal{L}}{\partial e(\theta)} = G'(\theta)\left\{-\theta + \lambda f'[e(\theta)]\right\} = 0$$

$$\frac{\partial \mathcal{L}}{\partial c(\theta)} = G'(\theta)(1-\lambda) = 0$$

Thus the effort function is again defined by the equality $f'[e(\theta)] = \theta$. So we come back to the competitive equilibrium allocation, in which the return to effort and its marginal cost are equal and in which only individuals of type $\theta \leq f'(0)$ participate in the labor market. At the collective optimum, the distribution of resources remains undetermined. The planner may choose any values of $c(\theta)$ that respect the resource constraint (3.16), which is saturated at the optimum, since $\lambda = 1$.

The efficiency of the competitive equilibrium has the corollary that steps taken by the public authorities to make jobs less demanding are undesirable if, and only if, markets function according to the principles of perfect competition. If there is no a priori restriction on the type of job on offer, the ones that are very difficult because highly dangerous, for example, are chosen and remunerated in full awareness of the risks and rewards, and any legal constraint that limits the difficulty of doing them results in a welfare loss. We can more clearly grasp the sense of this result if we ponder the impact of a policy aiming to reduce accident risk by putting security regulations in place. Let us assume that the variable of effort e simply equals accident risk, and that public policy consists of imposing an upper limit e^+ to this risk. The introduction of this constraint entails a welfare loss for all individuals whose disutility of labor θ is such that effort $e(\theta)$, defined by equation (3.15), is greater than e^+. Figure 3.7 shows us how the situation of

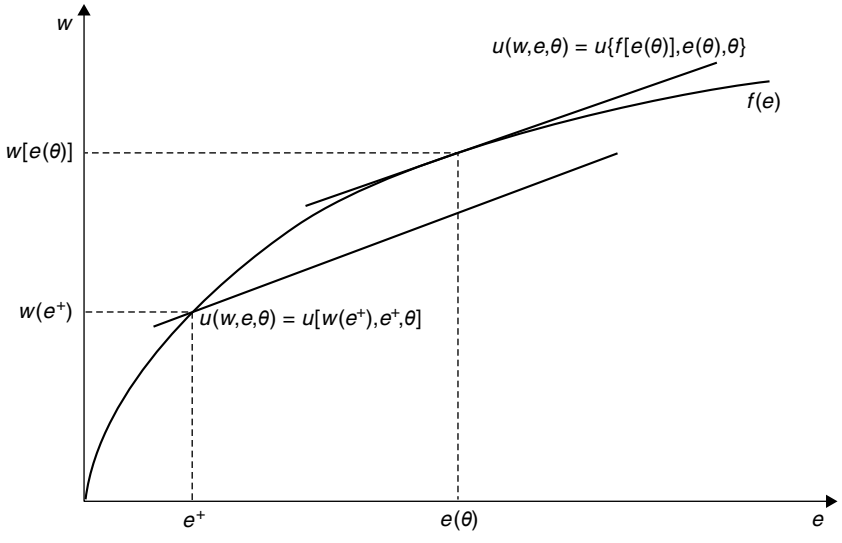


FIGURE 3.7
The impact of a legal constraint on accident risk.

such an individual changes. This situation corresponds to an indifference curve associated with a lower level of utility in the presence of the constraint on accident risk. The individual in this case also receives a lower wage, equal to $f(e^+)$.

It is worthwhile to insist on the fact that the uselessness of public interventions when it comes to the difficulty or danger of working conditions has been established only when markets function in accordance with all the principles of perfect competition (perfect information, free entry). In very many cases these conditions are not all met at the same time, especially as regards the quality of the information available about the dangers or hardship of jobs.

2.2 Does the Hedonic Theory of Wages Really Apply?

The main prediction of the hedonic theory of wages is that wage differentials compensate for the conditions in which a job is performed. Tests of this prediction run up against methodological difficulties having to do, on one hand, with unobserved characteristics, and on the other, with the heterogeneity of individual preferences about the attractive or unattractive features of doing any job. We illustrate these difficulties by presenting the application of the hedonic theory of wages to the problem of evaluating the price of a human life.

2.2.1 Considerations of Method

The method used to test the predictions of the hedonic theory of wages consists of estimating the wage w received by an individual as a function of his personal characteristics, represented by a vector $\mathbf{x}$, and the nonwage characteristics of the job, represented by a vector $\mathbf{e}$. In general, the equation estimated is of the form:

$$\ln w = \mathbf{x}\boldsymbol{\beta} + \mathbf{e}\boldsymbol{\alpha} + \varepsilon \tag{3.17}$$

In this expression, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are vectors of parameters to be estimated and ε is a disturbance term with zero mean that is assumed to be normally distributed. Vector $\mathbf{x}$ of personal characteristics generally includes age, sex, number of years of study or degree obtained, experience, seniority at work, ethnic origin, place of residence, family status, and trade-union membership. Vector $\mathbf{e}$ of the nonwage characteristics of jobs incorporates variables like the duration and the flexibility of hours worked, the repetitive aspect of tasks, the risk of injury, the level of ambient noise, the physical strength required by the job, the risk of job loss, the cost of health insurance, the cost of saving for retirement, and so on.

The Impact of Unobserved Individual Characteristics

Taking the nonwage aspects of jobs into account poses two delicate problems. The first arises from unobserved characteristics. In the hedonic model of wages just set forth, we have assumed that wage differences reflect differences in working conditions alone, for all individuals have the same efficiency (they all supply one unit of labor). The same holds good for firms: all jobs have identical productivity if the work performed is identical. This model thus predicts that wage differences reflect differences in working conditions, as long as the efficiency of workers and jobs is held constant. This is

why the characteristics of workers and jobs are included in equation (3.17) utilized to test the empirical predictions of the hedonic model. Nevertheless, individual efficiency depends on factors such as motivation and talent that as a general rule are not observed by the econometrician. If talent is unobservable, and if it influences the choice of working conditions, equation (3.17) does not permit us to estimate correctly the impact of working conditions on remuneration, for the nonwage characteristics of the job, represented by vector e, are correlated with the error term ε. For instance, good working conditions are likely to be normal goods, the "consumption" of which increases as income rises. If the income effect is sufficiently strong,[3] then the most efficient individuals choose the less laborious jobs, which entails a negative relation between wages and the laboriousness of jobs.

This point is illustrated in figure 3.8, which represents, in the plane (w, e), the choices of two individuals having different levels of efficiency. In this figure, parameter e is a unidimensional measure of the degree of laboriousness of tasks. In conformity with the theoretical elements developed in section 2.1, the equilibrium corresponds to a tangency point between one of his indifference curves, denoted u^+, and the frontier f^+ of possible combinations of wage and task laboriousness. The less efficient worker has lower productivity and finds himself facing a set of trade-off possibilities between wages and task laboriousness, of which the frontier f^- is situated beneath frontier f^+. Figure 3.8 represents a situation in which the wage obtained by the efficient worker is higher, but the degree to which his tasks are laborious is lower than that chosen by

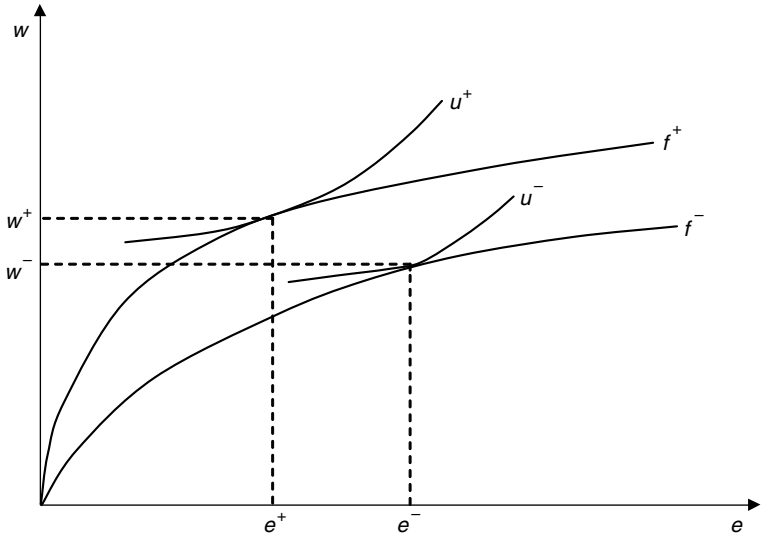


Figure 3.8
Compensating differentials and the unobserved characteristics of jobs.

[3] The utility function $w - \theta e$, which we have used in presenting the hedonic theory of wages, does not incorporate any income effect (see chapter 1).

the less efficient worker. So what we have is a negative relation between wages and the degree of laboriousness of jobs. If the difference between the frontiers f^+ and f^- is ignored by the econometrician, the negative correlation between wages and laboriousness of tasks will be underestimated. To escape this type of difficulty, it is preferable to make estimates using longitudinal data that allow us to follow individuals and thus control for their observable and unobservable time invariant personal characteristics (Brown, 1980; Duncan and Holmlund, 1983; Hwang et al., 1992).

The Importance of the Heterogeneity of Individual Preferences

The second problem encountered when estimating the impact of the nonwage elements of jobs on wages arises from the heterogeneity of individual preferences. There is not necessarily unanimous agreement that certain characteristics of jobs, like repetitiveness, use of physical strength, or flexible work schedules, are disagreeable, so the predictions of the hedonic theory of wages can only focus on certain elements that are clearly identifiable as drawbacks or advantages for all workers.

2.2.2 Empirical Evidence on Compensating Differentials

Empirical tests of the hedonic theory of wages give qualified results, which nevertheless, in many cases, highlight compensating wage differentials linked to the nonwage aspects of jobs. The first studies that focused on this area found that nonwage characteristics—like the repetitive content of jobs, bad working conditions, job security, freedom for persons to organize their own work, the opportunity for them to assist their fellow employees, the degree of supervision, the mortality rate, and the intensity of the work—had a sign that conformed to theoretical predictions (Brown, 1980). Nevertheless, Brown points out that numerous studies arrive at results that lack significance, or even contradict theoretical predictions, and suggests that these problems derive from the fact that all the studies conducted down to 1980 utilized cross-section data from which biases linked to the existence of unobserved variables might have arisen. For this reason, it is preferable to make use of longitudinal data. With that in mind, two identification strategies are available. The first consists of following workers who change jobs and identifying the compensating differences on the basis of the relation between the wages these persons earn and the jobs they hold. The second strategy consists of utilizing natural experiments, meaning events that modify, in an uneven (i.e., differentiated) manner, the characteristics of jobs held by similar workers.

Changes of Job

In order to identify compensating differentials, several studies focus on relations between job characteristics and the pay of workers who change jobs. This approach has the merit of controlling for unobservable, time-invariant individual characteristics by introducing individual fixed effects into the wage equation (3.17). However, if it is to pinpoint compensation differentials, this approach must overcome the endogeneity of job switching. This requirement is hard to meet, essentially for two reasons. First, since many characteristics of jobs are not observed by the econometrician, omitted variables may bias estimates. Second, given that job search bears a cost, movements from one job to another allow wage earners to increase their well-being. This means that they move from "worse" jobs combining a low wage and poor working conditions to "better" jobs that offer higher wages and good working conditions. Hence the changes in

remuneration observed when jobs are changed reflect both differences in job quality and perhaps differences in compensation (Hwang et al., 1998; Lang and Majumdar, 2004). These difficulties explain why this approach yields modest results that do not make it possible systematically to detect compensation differences.

Villanueva (2007) furnishes a good illustration of this approach. This paper uses longitudinal data on job characteristics and wages in Germany in 1984–2001. Its aim is to shed light on the arbitrage that occurs between rates of pay and disamenities by examining the characteristics of the different jobs held when jobs are changed through voluntary quits. The disamenities taken into account are heavy workload on a job, badly scheduled hours, skill requirements poorly matched to the skills possessed by the new worker, and job insecurity. Villanueva uses the model of Hwang et al. (1998) to show that when workers voluntarily quit one job in order to take another with greater (or less) disamenity, the average change in their wage sets an upper (or lower) boundary to the "value," or "return," attaching to this disamenity on the labor market. Villanueva finds that the market return to jobs with a heavy workload lies between 3.5% and 4.8%. He also finds that the market return to badly scheduled hours lies between 0 and 5.1%, and that jobs where the skills required are a poor match for the skills possessed by the typical worker pay a wage premium bounded by 0 and 6.1%. In contrast, Villanueva finds no significant relation between wages and job security. Overall, Villanueva's results lend some support to the theory of compensating wage differentials. Bonhomme and Jolivet (2009), in a similarly oriented study bearing on European data, are unable to detect compensation differences associated with nonwage amenities. They find strong preferences for some amenities, which are not reflected in wage/amenity correlations.

Other studies use panel data on displaced workers who switch jobs for exogenous reasons (Simon, 2001; Lehrer and Pereira, 2007). For instance, Lehrer and Pereira (2007) examine the experience of displaced workers who change jobs for arguably exogenous reasons on U.S. data for 1984–2002. They do not, however, succeed in showing that wages fall to compensate for the provision of health insurance.

Using data on displaced workers does have the advantage of enabling researchers to observe workers who switch jobs for exogenous reasons. But it has limitations arising from the fact that displaced workers generally suffer sizable losses of income, which suggests that the next job they find is generally of lower quality than the one they lost. There are in addition potentially important biases induced by variables that have been left out.

Natural Experiments

If researchers resort to natural experiments instead, they look for events constituting a credible source of variation in job characteristics, variation which is arguably exogenous to workers' employment decisions. These studies generally find evidence consistent with the theory of compensating differentials. For example Gruber (1994), exploiting changes in state legislation that made coverage for childbirth a mandatory part of all health insurance policies, finds substantial shifting (between 59% and 90%) of the cost burden of this mandate onto the wages of the targeted group.

DeSimone and Schumacher (2004) obtain results of the same kind. The authors examine the effects of the AIDS epidemic on the wages of nurses in the United States. AIDS being transmissible through contact with blood and other body fluids, nurses practicing their profession in zones where the epidemic is especially prevalent have

greater risk of contracting the disease than nurses in zones where AIDS is less prevalent. According to the theory of compensation differentials, their wages thus ought to be higher. DeSimone and Schumacher compare the earnings of nurses, the test group in this case, with those of college-educated non-health-care workers, the corresponding control group. They find that a 10% increase in the prevalence of AIDS leads to an increase in nurses' wages of slightly less than 1% relative to the wages of the control group.

Along the same lines, Rao et al. (2003) have estimated the compensation differential for condom use, exploiting data from a random sample of sex workers in Calcutta. In India, as in many developing countries, the propagation of condom use encounters resistance from men who strongly dislike using them. In September 1992 the All India Institute of Public Health and Hygiene began a program that attempted to provide basic health care facilities to sex workers and their families while also educating them about HIV-AIDS and the methods to prevent it. A group of sex workers were recruited to become peer educators. They were given intensive training on AIDS and other aspects of health care, provided with green coats to identify them as medical workers, and sent into the community to promote safe sex practices. The primary tool they employed for this purpose was a flip chart that used a series of pictures to explain the nature and progression of the HIV virus, its effect on the human immune system, and how the use of condoms was the most effective method of preventing the disease. The peer educators also carried condoms with them to distribute to the sex workers free of cost while demonstrating their proper use. Sex workers who wished to use condoms could also pick them up for free from nearby locations; thus condoms were available in virtually unlimited supply at zero cost. The sex workers who benefited from this program were selected progressively, independently of individual characteristics that might have influenced their adoption of the condoms. Consequently, by comparing the extra remuneration earned for having unprotected sex by the sex workers who had benefited from the information campaign with the extra earned the same way by those who had not, it is possible to identify the impact of the risk awareness imparted to the former. Rao et al. find that the sex workers who had benefited from the program had protected sex more often, and they estimate that insisting on the use of condoms entailed a significant loss ranging from 66% to 79% of the price they could have charged for agreeing to unprotected sex. That there does exist a remuneration compensating sex workers for the risk of sexually transmitted disease is confirmed by Arunchalam and Shah (2012), who utilize transaction-level data and biological sexually transmitted infection markers from sex workers in Ecuador. Arunchalam and Shah find that locations with lower disease prevalence exhibit lower premiums for unprotected sex. In the approximately 10% of locations with zero disease, the risk premium is close to zero and not statistically significant. Overall, a 1 percentage point increase in the local disease rate increases the premium for sex without a condom by 33%.

2.2.3 An Application to the Evaluation of the Value of a Statistical Life

The seminal contribution of Thaler and Rosen (1976) estimated a weekly wage equation similar to equation (3.17) in which the explanatory variables were age, level of education, geographic location, amount of time worked, the presence of a trade union, and the risk of a fatal industrial accident per year multiplied by 10^5. The main results are

presented in the first column of table 3.6. These figures led Thaler and Rosen to calculate the statistical value of a life saved in the following manner: "Suppose 1,000 men are employed on a job entailing an extra death risk of .001 per year. Then, on average, one man out of the 1,000 will die during the year. Since we know the amount of the average wage, the regression indicates that each man would be willing to work for $176 (in 1967 dollars) per year less if the extra death probability were reduced from .001 to 0. Hence, they would collectively pay $176,000 to eliminate that death: the value of the life saved must be $176,000" (Thaler and Rosen, 1976, p. 292). If we divide this figure by the average value of annual wages (given by $\$132.65 \times 50 = \$6,633.5$), we obtain the value of a statistical life expressed in years of wage. This price is given in the first column of table 3.6.

Obviously, the weight of unobserved variables renders estimates of wage equations fragile. To show the importance of these variables, Hwang et al. (1992) correct the estimates of Thaler and Rosen by making hypotheses, supported by a number of empirical studies, about the values of unobserved variables capable of biasing the estimate. Hwang et al. (1992) accounted for biases that depend on three variables: the heterogeneity in unobserved productivity, the percentage of earnings paid in nonwage form (advantages in kind, health insurance, retirement, etc.), and the dispersion of the preferences of workers when it comes to trading off between remuneration in the form of wages and in other forms.

Hwang et al. (1992) show that the work of Thaler and Rosen probably leads to a considerable underevaluation of the price of a statistical life. This point is illustrated in column 2 of table 3.6, which corresponds to a situation in which the heterogeneity of unobserved productivity (measured by the ratio of the variance of unobserved productivity to the variance of observed productivity) is equal to 0.395, the portion of earnings

TABLE 3.6
Estimates of the price of a life saved. The figures in parentheses are standard errors.

Weekly wage (in levels)	Model 1 (Thaler and Rosen)	Model 2 (Hwang et al.)
Age	3.89 (.80)	4.50
$(\text{Age})^2$	−.0479 (.0092)	−.0965
Education	3.40 (.55)	4.87
Risk	.0353 (.0210)	.3020
R^2	.41	.31
Price of a life saved (in years of average wage)	26.54	227.67
Values of variables (907 observations)	**Average**	**Standard error**
Age (years)	41.8	11.3
Education (years)	10.11	2.73
Weekly wage (1967 dollars)	132.65	50.80
Risk (probability $\times 10^5$)	109.8	67.6

Source: Hwang et al. (1992, table 1).

paid in wage form is equal to 0.80, and the dispersion of wages due to the heterogeneity of preferences (measured by the ratio of the variance of wages conditional on observed productivity to the total variance of wages) is equal to 0.106. The "corrections" made by Hwang et al. lead to an evaluation of the value of a statistical life almost 10 times higher than that obtained by Thaler and Rosen! So biases created by variables that have been left out are potentially very large, which ought to make us cautious in dealing with the results of studies of this type.

As matters stand, it is not surprising to find the wide variation in estimates of the value of a statistical life that we do find in articles on the subject. In a recent survey, Kniesner et al. (2012) have examined in great detail the problems posed by the econometric estimation of the value of a statistical life; these include the measurement of risk, the composition of samples, and the difficulty of taking into account the heterogeneity of individuals to which Hwang et al. (1992) had already drawn attention. For the United States, they estimate that the value of a statistical life in 2012 lies between $4 million and $10 million.

3 ASSORTATIVE MATCHING

The models examined so far in this chapter have assumed the existence of a large potential number of suppliers and demanders for every type of service traded. So, in the hedonic wage model of the previous section, there are as many markets as there are degrees of hardship, and on each of these markets there are implicitly a multitude of suppliers and demanders who are price takers. In addition, in this model the hypothesis of free entry into each market amounts to the assumption that it is possible to transform jobs in order to adapt them to the preferences of workers. For example, a modification of preferences corresponding to a greater aversion for bad working conditions induces a diminution of the number of jobs that offer bad working conditions and an increase in those that offer good conditions.

Such adjustments are pointers to a long-term phenomenon, the potential transformation of jobs. In the shorter term, it is also of interest to gain an understanding of the functioning of a market where jobs and workers all have different characteristics and where the distributions of these characteristics are exogenous functions. Under these circumstances, we must account not only for how wages are formed but also for how workers distribute themselves into the array of jobs they hold. In other words, we must explain how the characteristics of each worker are associated with the characteristics of each job. To analyze this problem, we resort to assortative matching models. These models are relevant for understanding the functioning of a market in which the heterogeneity of actors is enduring and plays an important role. Such is the case in particular for the markets for "superstars," whether they be sports figures, artists, journalists, lawyers, doctors, scientists, or managers of large companies, who possess specific talents hard to replicate.

We will study the functioning of a market of this type on the basis of an assortative matching model that associates chief executive officers (CEOs) who have different talents with firms of varying size. This model explains how the remuneration of CEOs is formed, as well as the manner in which they are allocated among the firms. As we

will see, the model allows us to understand why the remunerations of CEOs of closely similar talents may vary steeply and why the pay they earn can be extremely high and yet be socially efficient. The reason is that the most "talented" managers are to be found in the largest companies, which maximizes the global output of the economy. As we will also see, this model allows us to explain the very strong rise in the remuneration of the CEOs of large companies since the start of the 1980s.

The earliest models of assortative matching were advanced by Becker (1973) and Rosen (1981, 1982). Assortative matching models with hedonic wages were developed by Ekeland et al. (2004) and Chiappori et al. (2010). Tervio (2008) and Gabaix and Landier (2008) have applied assortative matching models to CEO remuneration.

3.1 A Competitive Equilibrium with Assignment

3.1.1 A Simple Model

Take the case of a continuum of workers (CEOs for present purposes) who differ in talent and productivity (ability), denoted $p \geq 0$. The distribution of talents is characterized by a cumulative distribution function (CDF) $F(.)$ with a smooth density function F' on $[p_0, +\infty)$. Take as well a continuum of firms with varying capacities to produce wealth. We may assume that this capacity is represented by the stock market value of each firm, which we will call its "size," denoted $\gamma > 0$, in order to simplify the vocabulary. Their size distribution is characterized by a CDF $G(.)$ with a smooth density function G' on $[\gamma_0, +\infty)$. There is the same number, or more exactly the same mass, of workers and firms. This mass is normalized to 1.

Most of the time the talent of a CEO is not objectively measurable. Falato and Milbourn (2012), for example, try to capture CEO talent using a battery of indicators like the number of times a CEO is mentioned positively in the business press, his educational attainment, and so on. They then construct a synthetic index which, albeit labeled CEO talent, actually has no more than ordinal value, meaning simply that it allows each CEO to be ranked on a talent scale. In practice, it does appear more useful to pursue this line of reasoning on the basis of a CEO's rank rather than his talent. There is no loss of generality when a CEO is indexed this way, that is, by the fraction of CEOs who are less productive than he is. Formally we may denote a the rank of a CEO in the distribution of abilities. By definition, the rank falls in the interval $[0, 1]$. Similarly, we can index each firm by its rank, denoted s, in the distribution of firm sizes.

A firm of size s matched to a CEO of talent a produces an output $Y(a, s) \geq 0$ (from now on, for the sake of simplicity, we will refer indifferently to talent or rank in the distribution of talents and adopt the same convention for firms). It is assumed that production function $Y(a, s)$ is increasing with the size of the firm and the talent of the CEO. If Y_i designates the partial derivative of Y with respect to its argument i, we then have $Y_1 > 0$ and $Y_2 > 0$. We also assume that CEOs who do not get matched obtain a payoff of zero.

The equilibrium of this model is described by an assignment function (or matching function) $\alpha(s)$ that defines the talent of the CEOs who head firms of size s and by a compensation function $w(a)$ that defines the remuneration of a CEO of talent a. More precisely, in this model a competitive equilibrium is made up of a compensation function $w(a)$, taken as given by each firm and each CEO, and an assignment function $\alpha(s)$,

such that no CEO-firm pair could do better by matching up with each other than they are doing with their current partners, and no CEO and no firm prefers to remain single.

3.1.2 The Equilibrium Assignment Function

The assortative matching model assumes that the mobility of CEOs occurs without friction and without cost and that information is perfect for all agents. The talent of CEOs and the size of firms in particular are perfectly observable. A CEO of talent a obtains a wage $w(a)$ and the firm of size s that employs a CEO of talent a obtains a profit:

$$\pi(a,s) = Y(a,s) - w(a) \tag{3.18}$$

The composite of functions $\{w(a),\alpha(s)\}$ is an equilibrium if there is no CEO-firm pair that could do better by matching amongst themselves than they are doing with their current partners. In other words, the assignment function $\alpha(s)$ yields the maximum value of the profit to each firm, and no CEO of talent $\alpha(s)$ can find another firm willing to pay him a higher wage than the wage he gets in firms of size s.

The assignment function is obtained by maximizing profit (3.18) with respect to a. The first-order condition is then obtained by canceling the derivative of $\pi(a,s)$ with respect to a, or

$$Y_1(a,s) = w'(a) \tag{3.19}$$

This condition indicates that, at the optimum, the marginal gain from increasing the talent of the firm's CEO, $Y_1(a,s)$, is equal to the marginal cost, $w'(a)$, incurred by having to pay the higher wage that would be needed to attract a CEO of higher talent. The second-order condition imposes:

$$Y_{11}(a,s) - w''(a) < 0 \tag{3.20}$$

At the competitive equilibrium, the assignment function, which describes the relation between a and s, must verify (3.19) for all s. We thus have:

$$Y_1[\alpha(s),s] = w'[\alpha(s)], \quad \forall s \tag{3.21}$$

Deriving this equation with respect to s, we have:

$$\alpha'(s) = \frac{Y_{12}[\alpha(s),s]}{w''[\alpha(s)] - Y_{11}[\alpha(s),s]}, \quad \forall s \tag{3.22}$$

Taking into account the second-order relation (3.20), we arrive at:

$$\alpha'(s) \lesseqgtr 0 \Leftrightarrow Y_{12}[\alpha(s),s] \lesseqgtr 0, \quad \forall s \tag{3.23}$$

This last inequality links the direction of variation of the assignment function to the cross derivative of the production function. By definition, the latter is said to be *supermodular* if $Y_{12} \geq 0$ and *submodular* if $Y_{12} \leq 0$. In assignment models of CEOs with firms of different sizes, it is assumed that the production function is supermodular

over the whole of its support. This amounts to stating that the marginal productivity of talent increases with the size of the firm or, to put it another way, that talent and firm size are complementary factors of production. That being the case, (3.23) shows that the assignment function is increasing: the "best" CEO (the one with the most talent) is assigned to the largest firm, the one whose talent ranks just below is assigned to the firm whose size ranks just below, and so on down to the least talented CEO, who is assigned to the firm of smallest size. Allocation of this kind is called positive assortative matching. This result is inverted if the production function is submodular: the most talented persons are then assigned to the least profitable firms (a configuration which might come about, for example, when individuals are called in as consultants or to turn a struggling firm around).

When $\alpha'(s) > 0$, all CEOs whose talent is inferior to given talent a find themselves in firms the size of which is inferior to $\alpha^{-1}(a)$. The market-clearing condition for CEO talent then entails that the "number" (more precisely, the mass) of CEOs whose talent is inferior to a must be equal to the "number" (more precisely, the mass) of firms the size of which is inferior to $s = \alpha^{-1}(a)$. In the reverse case, where $\alpha'(s) < 0$, the CEOs of greatest talent are paired with the firms that are smallest in size. Since a and s represent ranks, the market-clearing condition then is written:

$$\alpha(s) = \begin{cases} s & \text{if } Y_{12}(a,s) > 0, \quad \forall (a,s) \\ 1-s & \text{if } Y_{12}(a,s) < 0, \quad \forall (a,s) \end{cases} \tag{3.24}$$

In this context, the assignment function and the wage function define a competitive equilibrium, since each firm possesses a CEO whose talent maximizes its profit. No firm then has an interest in separating from the CEO it has. Reciprocally, no CEO can find another CEO of greater talent willing to change places with him.

An immediate consequence of the assignment rule is that the competitive equilibrium is efficient. In this model, the task of an omniscient planner would be to allocate talents according to the size of firms in such a way as to maximize the total output of the economy. This is exactly what the market accomplishes by establishing a bijective correspondence between the talents of CEOs and the size of firms: the CEOs with the most talent go to the firms of the largest size, and the CEOs with less talent go to the firms of smaller size, when the production function is supermodular. Under the hypothesis that the production function is supermodular, this process of resource allocation maximizes the overall production of the economy.

3.1.3 The Wage Rule and the Superstars Phenomenon

The compensation function $w(a)$ defined by equation (3.21) shows that the wage is increasing with talent, for $Y_1 > 0$. Note that this result holds good whatever hypotheses are adopted about the cross derivative Y_{12}. Thus greater talent is always compensated by more wage, whether the production function is supermodular or submodular.

We may go a bit further by integrating this wage rule. It is written as follows, denoting $\sigma(\cdot)$ the reciprocal of function α:

$$w(a) = w_0 + \int_0^a Y_1 [x, \sigma(x)] \, \mathrm{d}x \tag{3.25}$$

where w_0 is a constant representing the remuneration of the CEO of least talent. When $Y(0,0) = 0$, this remuneration is necessarily null, and the wage function is uniquely determined. Otherwise, there is a continuum of competitive equilibria associated with different w_0 which can be determined by some exogenous bargaining rule. For example, $w_0 = 0$ if the smallest firm can make a take-it-or-leave-it offer to the CEO of least talent, whose external option is here equal to zero by hypothesis.

This equation shows that the remuneration of each CEO depends on his own marginal productivity, as well as on the marginal productivity of all the CEOs of less talent. An increase in the marginal productivity of the less talented CEOs boosts the remuneration of the CEOs with more talent.

The equilibrium wage function of the assignment model entails that small differences in talent may give rise to large differences in wage. The impact of talent differences on wages may be amplified by the assignment rule. For example, when there is positive assortative matching, the most efficient CEOs are hired by the largest firms, which enables them to benefit from wages that are all the higher. To illustrate this property, we may posit a multiplicative production function $Y(a,s) = a \cdot s$. That being the case, the production function is supermodular and the equilibrium assignment rule (3.24) entails that $\sigma(a) = a$, and we then have $Y_1(a,\sigma(a)) = \sigma(a) = a$. The integration of equation (3.25) then gives the wage of a CEO of talent a. Assuming that $w_0 = 0$, we find $w(a) = \frac{a^2}{2}$. The wage is a convex function of talent, which means that small differences in talent between low-talent individuals give rise to slender differences in remuneration, but small differences in talent between very talented individuals give rise to wide differences in remuneration. As Rosen (1981) points out, this property is characteristic of the remuneration of superstars, whether they be CEOs of large companies or sports figures, journalists, or lawyers.[4] It should be noted that this result does not depend on the hypothesis that the production function is supermodular. If the technology is not supermodular, those with the most talent are not assigned to the most profitable firms, but remuneration may nevertheless be a convex function of talent. To confirm this, take the production function $Y(a,s) = a \cdot (1 - s)$. The equilibrium assignment rule implies that $\sigma(a) = 1 - a$. The wage of a CEO of talent a is always equal to $\frac{a^2}{2}$, assuming $w_0 = 0$.

Falato et al. (2012) have corroborated these predictions by constructing a synthetic scale of the talents of American CEOs based on a number of objective criteria such as their career paths, the reputation the press awards them, and their educational attainment. Their data comprise a sample of 2,195 CEO successions between 1993 and 2005. They show that remuneration at hiring increases on average by $280,000 for each decile of the talent distribution and that the remuneration function is effectively convex above a certain level of pay.

3.2 An Illustration: The Upswing in CEO Remuneration

Frydman and Saks (2010) have reconstituted the historical series of the remunerations of the three top-paid managers of the 50 largest American firms between 1936 and 2005 (according to the availability of data, the size of firms is defined by their turnover or

[4] For a more general analysis of the inequalities explained by the assortative matching model, see Kremer (1993) and Saint-Paul (2001).

their stock market value). Figure 3.9 shows the median level of total compensation, composed of salary, bonuses, long-term bonus payments (including grants of restricted stock), and stock option grants. We see that the remuneration of CEOs literally soared starting in the 1970s and especially from the 1980s on.

Gabaix and Landier (2008) have explained this upswing in the remuneration of the CEOs of the largest companies with the help of an assignment model analogous to the one set forth above. To that end, they utilize explicit functions for the distributions of talent and firm size and a production function with constant returns to scale, the parameters of which they estimate. The calibration of the model entails that the elasticity of average CEO compensation to average firm size at a given point in time should be equal to 1. This result explains very well the path followed by the remuneration of the CEOs of the 500 largest market capitalizations in the United States, which grew by 500% between 1980 and 2003 (see also figure 3.9), while the average market value of the 500 largest firms in the United States likewise grew by 500% over this period. When stock market valuations increase by 500%, under constant returns to scale, CEO "productivity" increases by 500%, and equilibrium CEO pay increases by 500%.

The calibration exercise of Gabaix and Landier suggests that the dispersion of talent is very narrow at the top end of the distribution. After classing the CEOs by decreasing order of talent (number 1 is the most talented), they calculate that if one were to replace CEO number 250 by CEO number 1, the value of his firm would increase by only 0.016%. However, these very small differences in talent translate into considerable

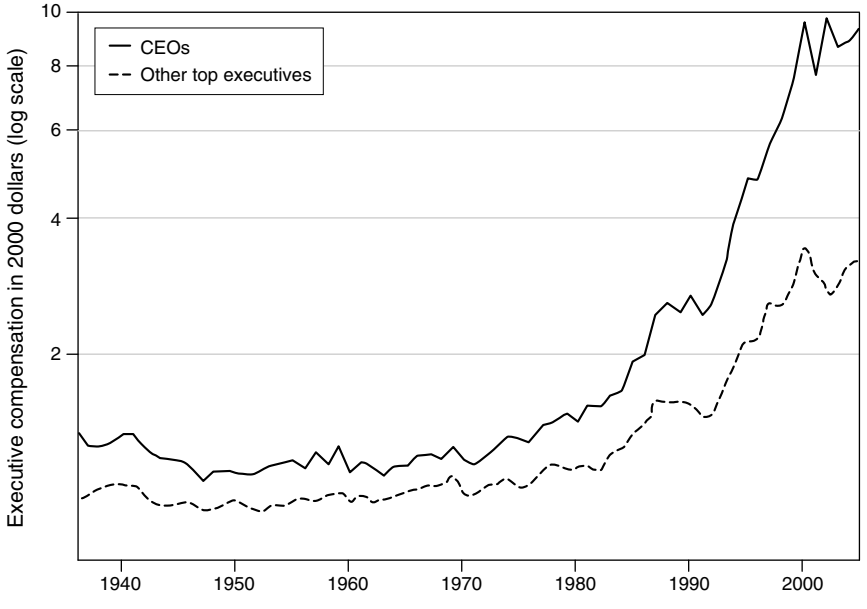


Figure 3.9
Median compensation of CEOs and other top officers from 1936 to 2005. The CEO is identified as the president of the company in firms where the CEO title is not used. "Other top executives" includes any executives among the three highest paid who are not the CEO. All dollar values are in inflation-adjusted 2000 dollars.

Source: Frydman and Saks (2010, figure 1).

compensation differentials, as they are magnified by firm size. Indeed, the same calibration delivers the result that CEO number 1 is paid over 500% more than CEO number 250. Tervio (2008) obtains results of the same kind with data on the 1,000 largest publicly traded companies in the United States in 1994–2004. The variation in CEO pay is found to be mostly due to variation in firm characteristics, whereas implied differences in managerial ability are small and make relatively little difference to shareholder value. Tervio estimates that the value added of scarce CEO ability within the 1,000 largest firms in the United States was about $21 to $25 billion in 2004, of which the CEOs received about $4 billion as ability rents while the rest was capitalized into market values. Llense (2010) arrives at analogous results on French data.

This assignment model offers a particularly simple and convincing explanation of the formation of CEO remuneration in recent years. But figure 3.10 shows clearly that the 1950s and 1960s were marked by a substantial increase in the size of firms without a simultaneous increase in CEO remuneration. Hence other complementary or competing theories like managerial rent extraction, greater power in the managerial labor market, or increased incentive-based compensation, also have a part to play in explaining the formation of CEO remuneration, depending on the epoch in question (see the survey of Bertrand, 2009).

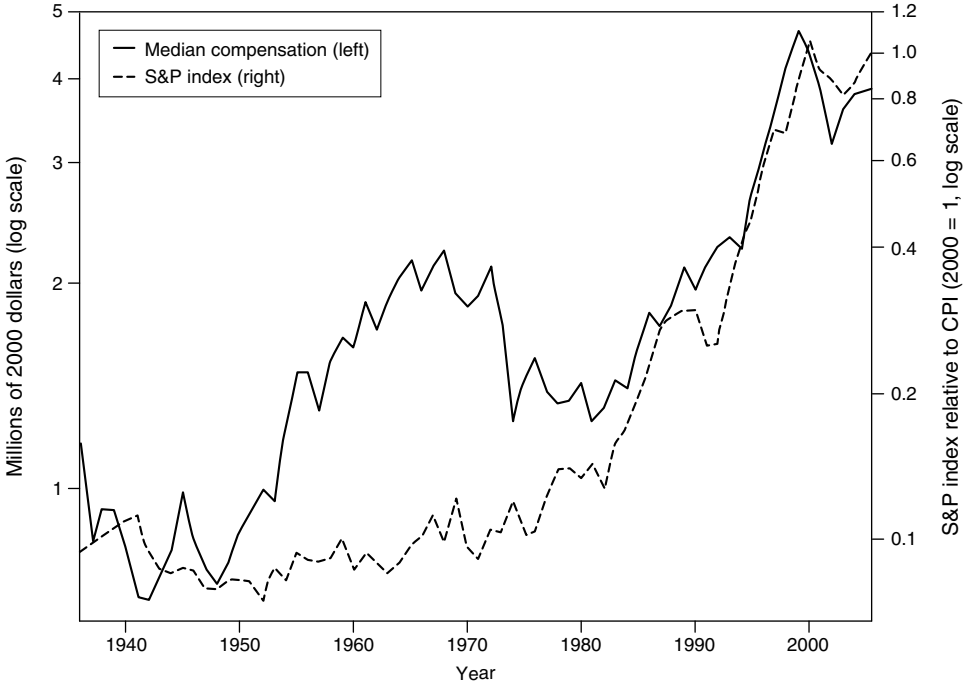


FIGURE 3.10

Total compensation and the Standard and Poor's index. Total compensation is composed of salary, bonuses, long-term bonus payments, and stock option grants. Based on the three highest-paid officers in the largest 50 firms in 1940, 1960, and 1990 (a total of 101 firms). The S&P index is expressed relative to the Consumer Price Index and equals 1 in 2000.

Source: Frydman and Saks (2010, figure 5).

4 SUMMARY AND CONCLUSION

- A perfectly competitive equilibrium on the labor market is characterized by wages that match supply and demand. Workers receive wages equal to their marginal productivity. When all jobs are equally hard, supply is principally determined by the disutility of work, which varies from one individual to another. This model also shows that a change in taxes does not necessarily bear on the agent on whom the tax is imposed, depending on how wages adjust.
- The hedonic theory of wages shows that the mechanism of perfect competition allows agents to choose different working conditions, and wage differentials "compensate" for the laboriousness or danger of tasks. Testing the extent of such wage compensation runs into difficulties having to do, on one hand, with unobserved individual characteristics and, on the other, with the heterogeneity of individual preferences when it comes to the advantageous or disagreeable aspects of the working conditions of a job. Empirical studies do, however, bring to light phenomena of wage compensation in many circumstances. But the orders of magnitude obtained must be interpreted with caution on account of the weight of the omitted variables.
- The assortative matching model explains how firms and workers with different characteristics match up in the same market. The equilibrium of a model of this kind describes an assignment function (or matching function) that indicates the firm to which each worker is assigned as a function of the characteristics of the firm and the worker, as well as the remuneration of each worker as a function of her characteristics and those of the firm that employs her. This model shows that the process of matching may provoke very steep inequalities of remuneration among workers of closely similar characteristics. In particular, this model explains the extremely high remunerations of superstars and explains in part the remuneration of CEOs in recent years.

5 RELATED TOPICS IN THE BOOK

- Chapter 5, section 4: Search frictions and wage differentials
- Chapter 6, section 2: Risk sharing
- Chapter 8, section 2: Theories of discrimination
- Chapter 9, section 2: The competitive model with labor adjustment costs
- Chapter 10, section 2.2: A model with skills and tasks
- Chapter 11, section 3: Migrations
- Chapter 12, section 2.2: Minimum wage and employment

6 FURTHER READINGS

Acemoglu, D., Autor, D., & Lyle, D. (2004). Women, war and wages: The effect of female labor supply on the wage structure at mid-century. *Journal of Political Economy*, *112*(3), 497–551.

Arunachalam, R., & Shah, M. (2013). Compensated for life: Sex work and disease risk. *Journal of Human Resources*, *42*(2), 345–369.

Bertrand, M. (2009). CEOs. *Annual Review of Economics*, *1*, 1–29.

Gabaix, X., and Landier, A. (2008). Why has CEO pay increased so much? *Quarterly Journal of Economics*, *123*, 49–100.

Gruber, J. (1997). The incidence of payroll taxation: Evidence from Chile. *Journal of Labor Economics*, *15*(S3), S72–S101.

Kniesner, T., Kip Viscusi, W., Woock, C., & Ziliak, P. (2012). The value of statistical life: Evidence from panel data. *Review of Economics and Statistics*, *94*(1), 74–87.

Rosen, S. (1986). The theory of equalizing differences. In O. Ashenfelter & R. Layard, (Eds.), *Handbook of labor economics* (vol. 1, chap. 12, pp. 641–692). Amsterdam: Elsevier Science.

REFERENCES

Acemoglu, D., Autor, D., & Lyle, D. (2004). Women, war and wages: The effect of female labor supply on the wage structure at mid-century. *Journal of Political Economy*, *112*(3), 497–551.

Anderson, P., & Meyer, B. (2000). The effects of the unemployment insurance payroll tax on wages, employment, claims and denials. *Journal of Public Economics*, *78*, 81–106.

Becker, G. (1964). *Human capital*. New York, NY: National Bureau of Economic Research.

Becker, G. (1973). A theory of marriage, part I. *Journal of Political Economy*, *81*, 813–846.

Bertrand, M. (2009). CEOs. *Annual Review of Economics*, *1*, 1–29.

Bonhomme, S., & Jolivet, G. (2009). The pervasive absence of compensating differentials. *Journal of Applied Econometrics*, *24*(5), 763–795.

Brittain, J. (1972). *The payroll tax for social security*. Washington, DC: Brookings Institution.

Brown, C. (1980). Equalizing differences in the labor market. *Quarterly Journal of Economics*, *94*, 113–134.

Card, D. (1990). The impact of the Mariel boatlift on the Miami labor market. *Industrial and Labor Relations Review*, *43*, 245–257.

Chiappori, P.-A., McCann, R., & Nesheim, L. (2010). Hedonic price equilibria, stable matching, and optimal transport: Equivalence, topology, and uniqueness. *Economic Theory*, *42*(2), 317–354.

DeSimone, J., & Schumacher, E. (2004). Compensating differentials and AIDS risk (Working Paper 10861). National Bureau of Economic Research.

Duncan, G., & Holmlund, B. (1983). Was Adam Smith right after all? Another test of the theory of compensating wage differentials. *Journal of Labor Economics*, *1*, 366–379.

Ekeland, I., Heckman, J., & Nesheim, L. (2004). Identification and estimation of hedonic models. *Journal of Political Economy*, *112*(1), S60–S105.

Falato, A., Li, D., & Milbourn, T. (2012). CEO pay and the market for CEOs. Finance and Economics Discussion Series, Divisions of Research & Statistics and Monetary Affairs, Federal Reserve Board, Washington, DC.

Feldstein, M. (1972). Comment on Brittain. *American Economic Review*, *62*, 735–738.

Frydman, C., & Saks, R. (2010). Executive compensation: A new view from a long-term perspective, 1936–2005. *Review of Financial Studies*, *23*(5), 2099–2138.

Gabaix, X., & Landier, A. (2008). Why has CEO pay increased so much? *Quarterly Journal of Economics*, *123*, 49–100.

Gruber, J. (1994). The incidence of mandated maternity benefits. *American Economic Review*, *84*, 622–641.

Gruber, J. (1997). The incidence of payroll taxation: Evidence from Chile. *Journal of Labor Economics*, *15*(S3), S72–S101.

Holmlund, B. (1983). Payroll taxes and wage inflation: The Swedish experience. *Scandinavian Journal of Economics*, *85*, 1–15.

Hunt, J. (1992). The impact of the 1962 repatriates from Algeria on the French labor market. *Industrial and Labor Relations Review*, *45*(3), 556–572.

Hwang, H., Mortensen, D., & Reed, R. (1998). Hedonic wages and labor market search. *Journal of Labor Economics*, *16*, 815–847.

Hwang, H., Reed, R., & Hubbard, C. (1992). Compensating wage differentials and unobserved productivity. *Journal of Political Economy*, *100*, 835–858.

Kniesner, T., Kip Viscusi, W., Woock, C., & Ziliak, P. (2012). The value of statistical life: Evidence from panel data. *Review of Economics and Statistics*, *94*(1), 74–87.

Kremer, M. (1993). The O-ring theory of economic development. *Quarterly Journal of Economics*, *108*(3), 551–575.

Lang, K., & Majumdar, S. (2004). The pricing of job characteristics when markets do not clear: Theory and policy implications. *International Economic Review*, *45*, 1111–1128.

Lehrer, S., & Pereira, N. (2007). Worker sorting, compensating differentials and health insurance: Evidence from displaced workers. *Journal of Health Economics*, *26*(5), 1034–1056.

Llense, F. (2010). French CEO compensations: What is the cost of a mandatory upper limit? *CESifo Economic Studies*, *56*(2), 165–191.

Rao, V., Gupta, I., Lokshin, M., & Jana, S. (2003). Sex workers and the cost of safe sex: The compensating differential for condom use in Calcutta. *Journal of Development Economics*, *71*(2), 585–603.

Rosen, S. (1974). Hedonic prices and implicit markets. *Journal of Political Economy*, *82*, 34–55.

Rosen, S. (1981). The economics of superstars. *American Economic Review*, *71*(5), 845–858.

Rosen, S. (1982). Authority, control and the distribution of earnings. *Bell Journal of Economics*, *13*, 311–323.

Roy, A. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers*, *3*(2), 135–146.

Saint-Paul, G. (2001). On the distribution of income and worker assignment under intrafirm spillovers, with an application to ideas and networks. *Journal of Political Economy*, *109*(1), 1–37.

Simon, K. (2001). Involuntary job change and employer provided health insurance: Evidence of a wage/fringe benefit tradeoff? *International Journal of Health Care Finance and Economics*, *1*, 249–271.

Tervio, M. (2008). The difference that CEOs make: An assignment model approach. *American Economic Review*, *98*(3), 642–668.

Thaler, R., & Rosen, S. (1976). The value of saving a life: Evidence from the labor market. In N. Terleckyj (Ed.), *Household production and consumption*. New York, NY: Columbia University Press.

Villanueva, E. (2007). Estimating compensating wage differentials using voluntary job changes: Evidence from Germany. *Industrial and Labor Relations Review*, *60*(4), 544–561.

CHAPTER 4

Education and Human Capital

In this chapter we will:

- See that education represents a significant and rising investment in the OECD countries
- Study how the theory of human capital explains the choice of how much education to get
- Understand why time spent on acquiring education can also serve to signal individual abilities to future employers
- Grasp how the returns, individual and social, to education are estimated
- Work through an example of the estimation of the returns to education based on the contribution of Angrist and Krueger (1991), who use the method of instrumental variables (Data and programs allowing readers to replicate the main results of this contribution are available at www.labor-economics.org.)
- Gain an overview of the principal results concerning the returns to education

INTRODUCTION

A decent amount of education, certified by a recognized diploma, is often seen as a basic necessity for winning a well-paid job. There may be several reasons for this. According to the theory of human capital which became popular following Becker (1964), education is an investment, producing knowledge acquisition and increased productivity, leading in turn to higher earnings. Some economists, though, see this concept of education as very reductive. Much of what is taught in primary, secondary, and post-secondary institutions brings no immediate payoff in the labor market (and seems not to have the virtue of promoting socialization either). Studying mathematical functions, for example, is of practical value in only a handful of professions, so why inflict it on vast numbers of students who will never need it? Some justify this kind of study by arguing that it

develops a capacity for abstract thought and therefore promotes higher productivity. Others, however, take the view that the essential virtue of this type of learning is to select students. From this perspective, first formulated by Spence (1973), the education system plays the role of a filter: it selects individuals on the basis of their intrinsic efficiency, allowing them to signal their abilities to potential employers. If education serves both to acquire knowledge and select individuals, then we must try to determine the respective weight of each of these dimensions, not only to understand the impact of education on earnings and growth but also to assess the effectiveness of expenditure on education, a large portion of which is paid by the state in all OECD countries. To enable us to grasp the exact role of education and then if possible quantify it, we will need a precise conceptual structure, capable of representing the consequences of both knowledge acquisition and selection. This is what the economic analysis of education aims at.

Following a review of the main features of the education systems in the OECD countries, we will see how the theory of human capital accounts not just for the relationship between education and earnings but also for the choice of how much education to get. Individual choices will be seen to be socially efficient if the labor market is competitive and if education produces no externalities. We will then see how, when information asymmetries on the labor market were taken into account, Spence was led to emphasize the role played by the education system as a selection mechanism. In this context, individual choices about education are generally socially inefficient and may lead, in certain circumstances, to overeducation—something which may appear paradoxical, given the degree to which the state strives to promote access to education. The final sections of this chapter are devoted to empirical studies that attempt to estimate the returns to education and assess the causal linkage between education and earnings. These studies suggest that the education system does make a significant contribution to improving the efficiency of individuals in the labor market by imparting knowledge to them. Thus they highlight the relevance of the model of human capital as a tool for analyzing problems arising from education and the labor market. They also show that education gives rise to externalities that justify, to a certain extent, state intervention in this area.

1 SOME FACTS

This section brings together the principal descriptive data regarding the extent of spending on education in OECD countries and some non-OECD countries and the impact of the education system on wages and employment for those who pass through it.

1.1 Spending on Education

On average, the OECD countries spend 6.3% of their GDP on educational institutions. Non-OECD countries such as Brazil or the Russian Federation spend comparable amounts (see figure 4.1). According to the OECD definition (OECD, 2012, p. 214), spending on educational institutions includes expenditure on instructional educational institutions as well as expenditure on noninstructional educational institutions. Noninstructional educational institutions are educational institutions that provide administrative, advisory, or professional services to other educational institutions,

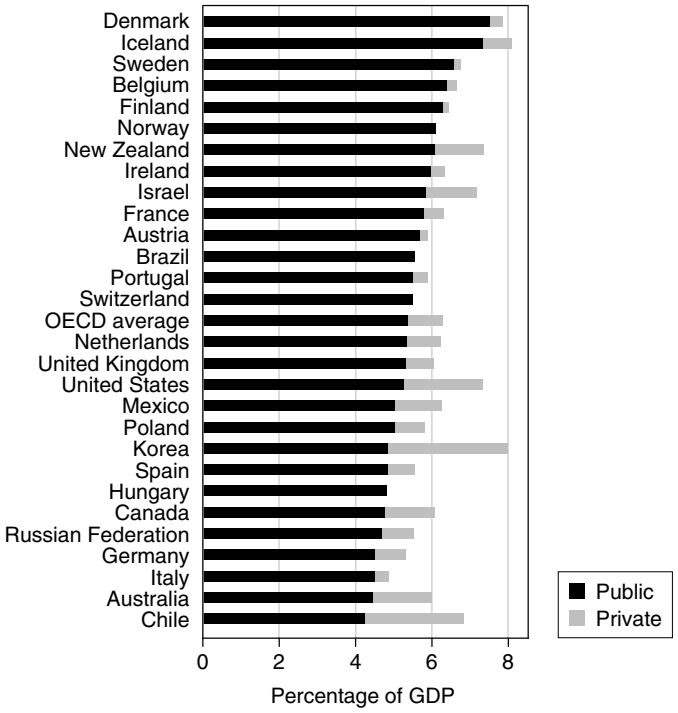


Figure 4.1

Expenditure on educational institutions as a percentage of GDP, 2009. The OECD average is the nonweighted average of the 34 OECD countries, including those not represented in this figure. Brazil and the Russian Federation are not part of the OECD. Private expenditure is missing for Brazil, Hungary, Norway, and Switzerland. Data are missing for China, Greece, and Turkey.

Source: OECD (2012, table B2.3, p. 246).

although they do not enroll students themselves. Examples include national, state, and provincial ministries or departments of education; other bodies that administer education at various levels of government or analogous bodies in the private sector; and organisations that provide such education-related services as vocational and psychological counseling, placement, testing, financial aid to students, curriculum development, educational research, building operations and maintenance services, transportation of students, and student meals and housing. Among the OECD countries, expenditure on educational institutions as a percentage of GDP runs from 5% in Italy to 8% in Denmark and Korea.

In most countries, education is financed primarily from the public purse, with the consequence that expenditure on education today constitutes a significant budget item. Even in the United States, private expenditure came to only 39% of public expenditure in 2009. Exceptions are Korea and Chile, where private expenditure represents about 60% of public spending on education. In Finland, where the portion of private expenditure directed to education is the lowest, it comes to only 2.2% of public expenditure. For the OECD countries, this ratio averages 17%.

1.2 GRADUATION RATES

At the dawn of the 21st century, a majority of the population in the majority of OECD countries has obtained a diploma signifying the completion of upper secondary education. According to the definition of the OECD, upper secondary education corresponds to the final stage of secondary education in most OECD countries. The entrance age to this level is typically 15 or 16 years. There are substantial differences in the typical duration of programs both across and between countries, typically ranging from two to five years of schooling. Upper secondary education may be either "terminal" (i.e., preparing the students for entry directly into working life) or "preparatory" (i.e., preparing students for post-secondary education). Figure 4.2 shows that the average percentage of the working-age population that has completed secondary schooling in 2010 is 74% for the OECD countries. Educational levels are advancing, for in all countries the proportion of the population with at least secondary schooling is higher in the age range

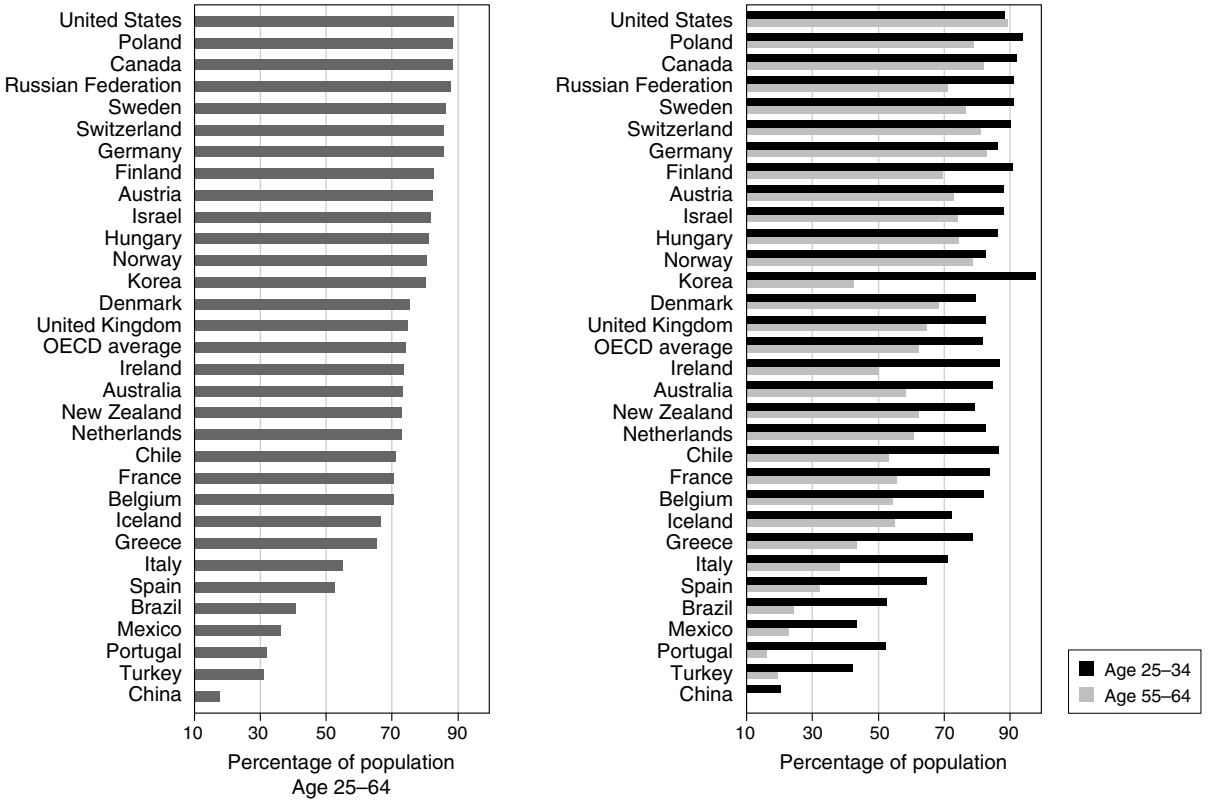


FIGURE 4.2

Percentage of the population that has attained at least upper secondary education, by age group, 2010. The OECD average is the nonweighted average of the 34 OECD countries, including those not represented in this figure. Brazil, China, and the Russian Federation are not part of the OECD.

Source: OECD (2012, table A1.2a, p. 35).

from 25 to 35 than it is in the age range from 55 to 64. In this area a convergence phenomenon is observable, inasmuch as countries where the rates of secondary schooling were lower to start with have advanced more rapidly than the others.

Figure 4.3 shows that the percentage of those with a diploma signifying the completion of tertiary (or in common parlance, post-secondary) education in 2010 is, on average, 31% in the OECD countries. This figure is about two thirds smaller than that for those with upper secondary diplomas. The proportion of individuals with tertiary education is 1.7 times as high in the age range from 25 to 35 as it is in the age range from 55 to 64. Tertiary education, like secondary education, is thus clearly on the rise, but between these two age ranges secondary education has been advancing more rapidly than tertiary education, since the difference in educational level is 25% for secondary and 15% for tertiary. Here again convergence is observable, for the countries where the rates of tertiary education have advanced most rapidly are the ones where these rates were lower to start with.

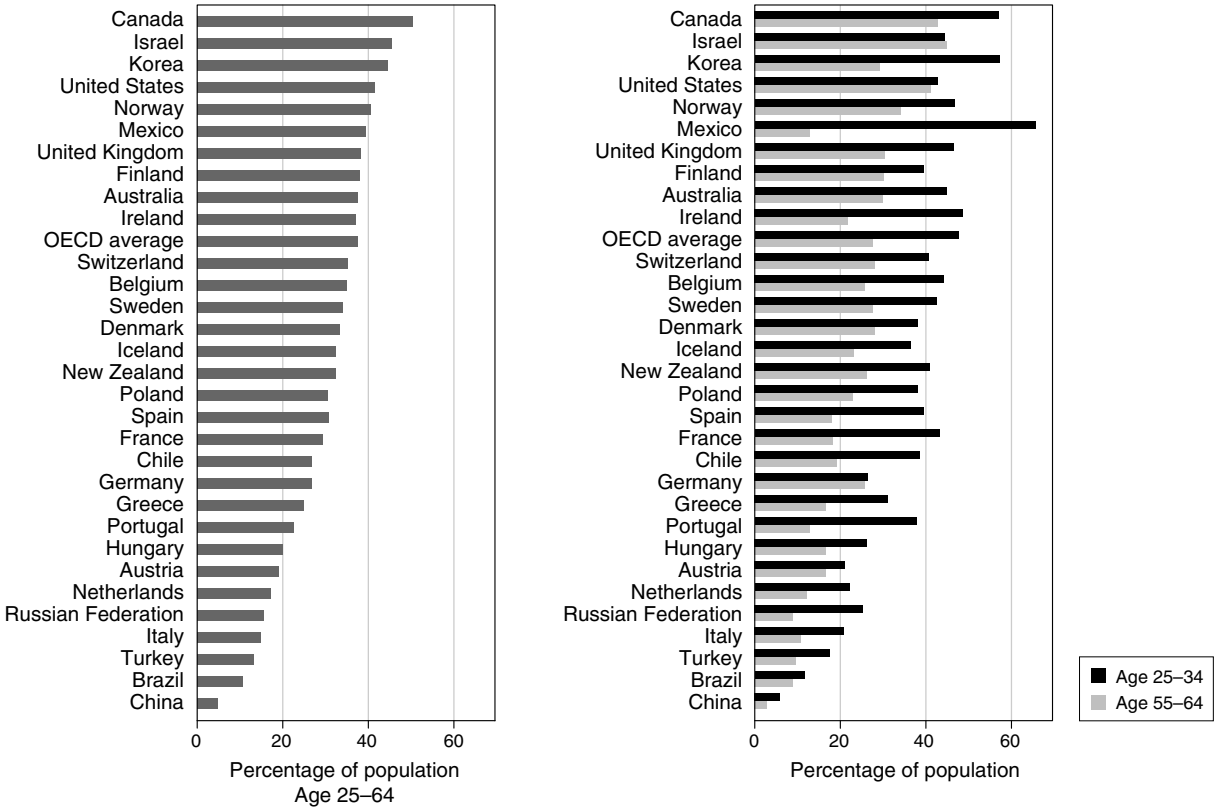


Figure 4.3

Percentage of the population that has attained at least tertiary education or advanced research programs, by age group, 2010. The OECD average is the nonweighted average of the 34 OECD countries, including those not represented in this figure. Brazil, China, and the Russian Federation are not part of the OECD.

Source: OECD (2012, table A1.3a, p. 36).

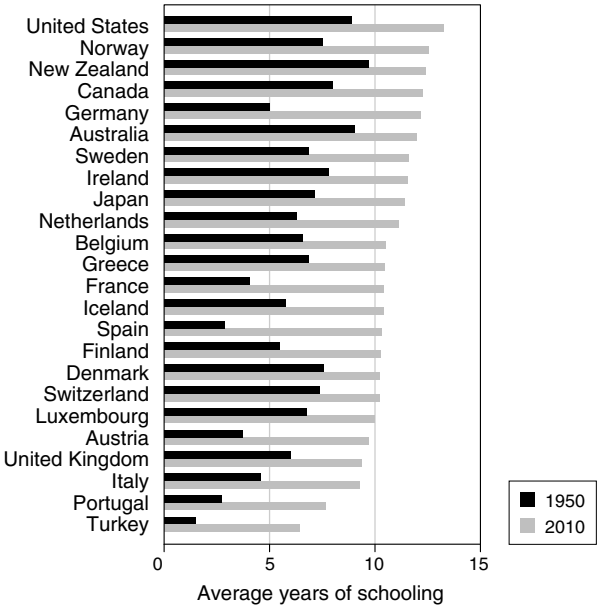


FIGURE 4.4
Years of schooling of the total population aged 25 and older.

Source: Barro and Lee (2010, education data set, available at www.barrolee.com/data).

There is thus a significant increase in the duration of schooling in the OECD countries as a whole. For persons aged 25 and older, the average duration of schooling went from 5.7 to 10.7 years between 1950 and 2010 in the advanced economies reported in figure 4.4. Moreover, the duration of schooling is increasing in all these countries, without exception. Yet in 2010 average durations of schooling were still widely dispersed: the United States had the highest figure, 13.3 years, and Turkey had the lowest, 6.5 years.

1.3 EDUCATION AND PERFORMANCE ON THE LABOR MARKET

Higher levels of education are positively correlated to greater labor market participation and to better performance in this market. Figure 4.5 shows that wages rise with educational level in all the countries considered. On average, in the OECD in 2010 a worker with less than upper secondary level receives a wage equal to 77% of the wage of a worker who has reached upper secondary education. Wage earners with a tertiary level diploma receive wages 54% higher than those with an upper secondary diploma. This suggests that to acquire education is a way to elevate one's wages. As well, figure 4.6 shows that, on average, rates of unemployment fall off as educational level rises. In 2010 the average rate of unemployment for those with a tertiary diploma was 4.7% in the OECD countries reviewed in figure 4.6. Persons with an educational attainment falling below upper secondary level have a probability more than twice as large of experiencing unemployment as do those with a tertiary diploma, since their unemployment rate is 12.5%. During the great recession of 2008, the rate of unemployment for those

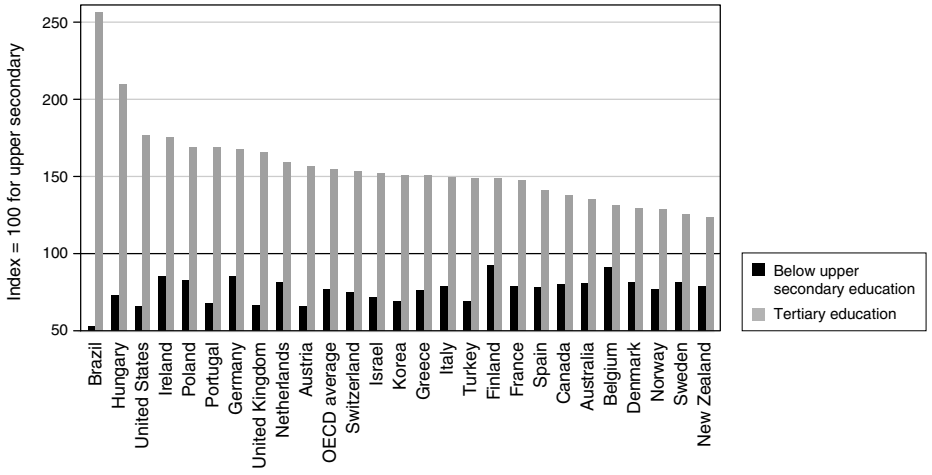


Figure 4.5

Relative earnings from employment among 25- to 64-year-olds, by level of educational attainment (2010 or latest available year). Upper secondary and post-secondary nontertiary education = 100. The OECD average is the nonweighted average of the 34 OECD countries, including those not represented in this figure. Brazil is not part of the OECD. Data are missing for Chile, China, Iceland, Mexico, and the Russian Federation.

Source: OECD (2012, chart A8.1, p. 140).

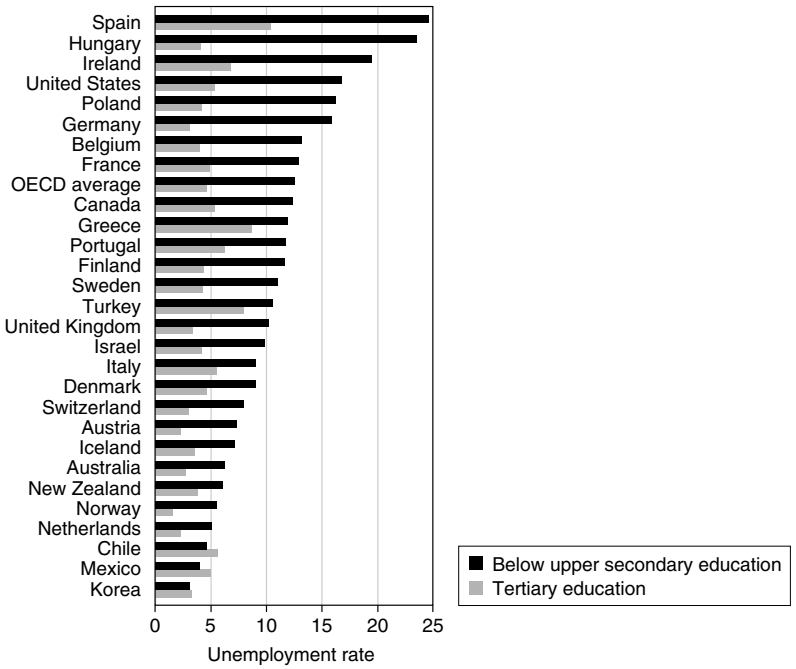


Figure 4.6

Unemployment rates by level of educational attainment for 25- to 64-year-olds, 2010. The OECD average is the nonweighted average of the 34 OECD countries, including those not represented on this figure. Data are missing for non-OECD countries.

Source: OECD (2012, table A7.4a, p. 133).

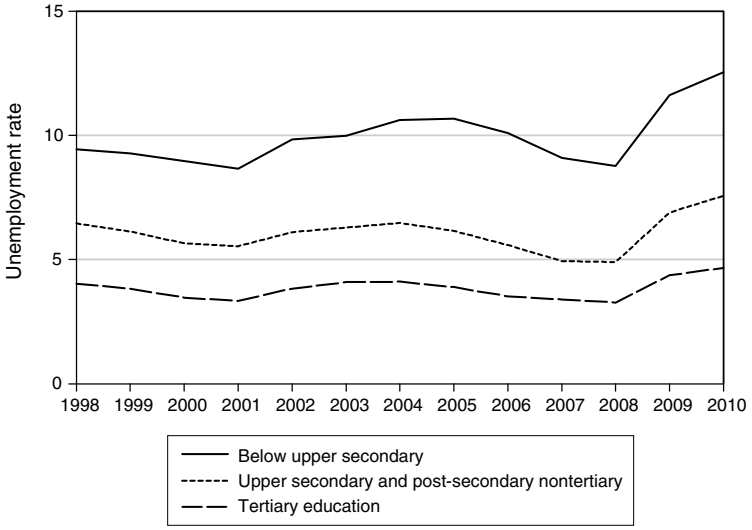


FIGURE 4.7
Unemployment rates by level of educational attainment for 25- to 64-year-olds, 2010. The OECD average is the nonweighted average of the 34 OECD countries.

Source: OECD (2012, table A7.4a, p. 133).

with education below upper secondary level rose much more than for those with tertiary education, as shown in figure 4.7.

The factual elements just reviewed lead us to three essential conclusions. First, every country dedicates an important share of its total expenditures to education. Second, the majority of persons in the OECD countries which we reviewed stay in school long enough to reach the upper secondary level. Finally, higher levels of education are linked to better labor market performance. The rest of this chapter is devoted to exploring and explaining this state of affairs. More precisely, we shall see how economic analysis can clarify not only the linkage between education and labor market performance but also the factors that determine individual decisions about education.

2 THE THEORY OF HUMAN CAPITAL

The theory of human capital, inaugurated by Becker (1964), starts with the hypothesis that education is an investment that will produce earnings in the future. In this context, wage differentials are influenced by differences in individual productivity, which are themselves influenced by investments in education or training (the two terms are used indifferently here) made by individuals throughout their lives. To acquire competences that the labor market will reward brings "training costs" comparable to investments that will be sources of future earnings. These costs include the expenses of study (fees to enroll in specialized establishments, costs of lodging and travel, purchase of materials, etc.), potential loss of earnings due to the fact that time spent on study is not devoted to remunerated activity, and the psychological costs arising from stress and possibly the

sheer difficulty of studying. Investments in education may pay off when they produce an accumulation of competences, "human capital" as it is called, which brings returns in the form of higher remuneration.

We begin by showing how the mechanisms of competition allow individuals to make their investments in training pay off. We will prove that individual choices about training are socially efficient if markets are perfectly competitive. Next we will analyze the dynamic dimension of educational choices using a simple model in which individuals receive education only at the outset of their active lives; in other words, education is taken as equivalent to schooling (primary, secondary, and post-secondary). In this setting, the number of years spent in schooling is conditioned by individual characteristics and influences future earnings. We then extend this model by assuming that agents have the opportunity to add to their education over the course of their entire professional lives. We will see that a simulation of this model conforms very closely to the path of earnings over life cycle observed in the real world.

2.1 The Relation Between Earnings and Human Capital

From Becker's perspective, education can only be a source of future earnings if wages reflect differences in productivity. Now it is not at all self-evident that improved productivity on the part of a wage earner does lead systematically to an increase in her wage, even in a perfectly competitive labor market in which firms have perfect knowledge of workers' characteristics, and workers and jobs are both perfectly mobile.[1] In reality, a worker who has acquired competences and expertise that improve her productivity can only make them pay off if she is able to play two or more employers off against one another. A single employer would indeed have no reason to raise the wage of a worker whose productivity had improved if that worker could not credibly threaten to take a better-paid job elsewhere. This observation led Becker to adopt the distinction between *general training*, which enhances the productivity of the individual concerned for all types of job, and *specific training*, which only enhances her productivity for one particular type of job. This distinction is clearly theoretical, to the extent that all training has a certain degree of specificity, but it is analytically useful. General training is fundamentally associated with the worker, who can make it profitable in different types of job and so bring employers to compete for her services. Specific training is associated with a particular type of job.

The link between wages and human capital can be highlighted in a very simple two-period model in which the labor force is made up of a continuum of identical workers, the size of which is normalized to 1. Each worker lives 2 periods. The first period of life is devoted to education only and the second period to production. For the sake of simplicity, there is no preference for the present so that the discount rate is equal to zero.

If he has had the advantage of *general* training equal to i, he is capable of producing a quantity of goods $y(i)$ in the second period if he occupies any job whatever. On the other hand, if he has had the advantage of *specific* training equal to i for a particular job, he is capable of producing a quantity of goods $y(i)$ if he occupies that particular job. Whenever he is not holding a job in the second period of life, each worker obtains z units of goods. The production function $y(i)$ is assumed to be increasing, concave,

[1] The mechanisms of perfect competition are presented in detail in chapter 3, section 1.

and such that $y(0) \geq z$. For simplicity's sake, the amount of time needed to make an investment in training is assumed to be zero.

2.1.1 Competitive Equilibrium with General Training

In a situation of perfect competition, all suppliers of labor who have made an investment i in general training are employed if they want to be. The condition of free entry into the market ensures that the profits of the entrepreneurs who employ trained individuals are zero, that is, $y(i) = w(i)$, where $w(i)$ designates the wage received by a worker who has level i of general training.

In such a case, a worker cannot make a credible promise to share the fruits of an investment in general training with the first employer she encounters (the wage earner receiving less than $y(i)$ if the entrepreneur participates in the investment in training), because once the investment has been made, the worker has an interest in quitting that employer, knowing that she will immediately find another firm to offer her wage $y(i)$. The upshot is that suppliers of labor are the only real beneficiaries of investments in general training and so must bear the entire cost of it themselves.[2] Optimal investment maximizes $y(i) - i$, and is thus defined by relation:

$$y'(i) = 1 \tag{4.1}$$

This result signifies that each individual has an interest in investing to enhance her general training as long as the marginal return $y'(i)$ of this investment is greater than its marginal cost, here equal to 1. Employers for their part have no incentive to finance this type of training because every worker can obtain a wage increase by offering her services to competing bidders as soon as her productivity increases.

2.1.2 Competitive Equilibrium with Specific Training

By definition, when training is specific, workers can only make their training pay off in a particular job. Once trained, they are unable to demand wage increases from their employer by making him bid against other employers. Hence employers may have an incentive to invest in this type of training. This conclusion will emerge more clearly if we represent the decisions of the second period of life by a two-stage game. In the first stage, employers freely enter the market and compete through the wages they offer to workers. In the second stage, each employer chooses the level of investment in specific training that maximizes his profit. Given wage w offered in the first stage, this profit is written $y(i) - w - i$. Profit maximization then gives an investment i^* satisfying $y'(i^*) = 1$. Free entry in the first stage of the game entails zero profit, and thus wage $w = y(i^*) - i^*$. As in the case of general training, workers obtain an income equal to their productivity minus the cost of investment in training.

2.1.3 The Social Optimum

Choices made by individuals within the framework of perfect competition lead to social efficiency. This can be verified by writing the problem of a planner seeking to determine

[2] We will see in chapter 14 that there are situations of imperfect competition in which employers may have an interest in investing in general human capital.

optimal investment in training, whether general or specific. Since $y(0) \geq z$, the planner decides to assign all of them to the technology $y(\cdot)$ in use in the market rather than let them produce z domestically. If the planner dedicates an amount i of resources to the training of an individual, his problem is written as follows:

$$\max_i y(i) - i$$

The solution of this problem is again given by the equality (4.1). Thus, in a perfectly competitive economy, individual choices regarding training are socially efficient.

The theory of human capital suggests that the mechanisms of competition give individuals an incentive to become educated for the purpose of acquiring knowledge or skills on which the market sets a premium. Moreover, it shows that individual educational choices are socially efficient if the labor market is perfectly competitive.

Evidently in reality, markets are not perfectly competitive, as we will see more precisely in the second part of the present book. That being the case, wages and productivity differ, and educational choices are no longer efficient. For example, if wages are lower than productivity because firms dispose of monopsony power, the investment in human capital is less than the social optimum. We will also see at the end of this chapter that there are externalities, for the most part positive, associated with education. In particular, better-educated persons transmit part of their knowledge, which boosts the productivity of those around them. Education also reduces criminality for that matter. The collective return to education is thus greater than the individual return. Now, individuals do not take the positive externalities of education into account when they choose how much effort to put into their education. That implies that educational effort is generally insufficient in the absence of any intervention by the public authorities.

In sum, the results obtained in this section signify, more generally, that competition on the labor market allows workers to derive value, in the form of earnings, from knowledge that improves their productivity. Conversely, in the absence of competition, the incentives to invest so as to improve productivity disappear. For this reason it is generally firms that invest in specific training, which wage earners cannot exploit to increase their market value. We now examine in greater detail the determinants of educational investment in a context in which improved productivity leads to a higher wage.

2.2 Schooling and Wage Earnings

The theory of human capital throws light on the choice of the duration of studies. It shows that the length of time spent in school is influenced by individual characteristics such as aptitude and inherited human capital, by the discount rate, and by the productivity achieved thanks to the accumulation of human capital.

2.2.1 The Choice Between Getting Educated and Getting Paid

To illustrate these propositions, we examine the choices of an individual who can acquire education starting at date $t = 0$ and whose life in the labor force ends at date $T > 0$. The retirement period is set aside for the sake of simplification. We work with a continuous time model in which the preferences of an agent are represented by an instantaneous utility function equal to his current earnings and by a discount factor $r > 0$. At every moment it is possible to study or work but not to do both at the

same time. Education allows the accumulation of "human capital," that is, it allows the agent to increase his stock of knowledge. We assume from this point forward that over every interval of time $[t, t+\mathrm{d}t]$, it is possible for an individual to dedicate a fraction $\sigma(t) \in [0, 1]$ of this interval to training. The law of motion of human capital, denoted $h(t)$, is defined by the differential equation:[3]

$$\dot{h}(t) = \theta\sigma(t)h(t) \tag{4.2}$$

The parameter θ represents the efficiency of the effort made by the agent to become educated, so it reflects his aptitude. Relation (4.2) simply means that if an individual decides to become educated, the relative increase $\dot{h}/h$ in his human capital is proportional to his individual efficiency θ and to his effort in education $\sigma(t)$. Let us assume that an individual endowed with a stock of human capital $h(t)$ at date t produces a quantity of goods $Ah(t)$, $A > 0$, at this date, and that there is free entry into any type of job. Then profits are zero and the wage received at date t by this person will simply equal $Ah(t)$ when that person works. It follows that if an individual dedicates a fraction $\sigma(t)$ of period $[t, t+\mathrm{d}t]$ to education, he works during a fraction $1 - \sigma(t)$ of this period, and so receives earnings $A[1 - \sigma(t)]h(t)\mathrm{d}t$. Thus his gain discounted over the whole of his life cycle is defined by:

$$\Omega = \int_0^T A[1 - \sigma(t)]h(t)e^{-rt}\mathrm{d}t \tag{4.3}$$

To define the optimal choice of schooling, it is useful to compute the marginal returns to education effort at time t. We get:

$$\frac{\partial\Omega}{\partial\sigma(t)} = -Ah(t)e^{-rt} + \int_0^T A[1 - \sigma(z)]\frac{\partial h(z)}{\partial\sigma(t)}e^{-rz}\mathrm{d}z$$

Since the integration of the differential equation (4.2) yields:

$$h(t) = h_0 \exp\theta\int_0^t \sigma(z)\mathrm{d}z$$

where h_0 denotes the stock of human capital for zero year of schooling, we have:

$$\frac{\partial h(z)}{\partial\sigma(t)} = 0 \quad \text{if } z < t \quad \text{and} \quad \frac{\partial h(z)}{\partial\sigma(t)} = \theta h(z) \quad \text{if } z \geq t$$

Therefore,

$$\frac{\partial\Omega}{\partial\sigma(t)} = -Ah(t)e^{-rt} + \int_t^T \theta A[1 - \sigma(z)]h(z)e^{-rz}\mathrm{d}z \tag{4.4}$$

[3] The time derivative of $h(t)$ is denoted $\dot{h}(t)$.

Then we can compute the derivative of the marginal returns to education effort with respect to t, or, formally:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{\partial\Omega}{\partial\sigma(t)}\right] = -A\dot{h}(t)e^{-rt} + rAh(t)e^{-rt} - \theta A[1-\sigma(t)]h(t)e^{-rt}$$

This last equation implies, together with (4.2):

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{\partial\Omega}{\partial\sigma(t)}\right] = Ah(t)e^{-rt}(r-\theta) \tag{4.5}$$

This equation shows that the marginal return to educational effort increases over time if $r > \theta$. Since equation (4.4) entails that the marginal return to education is negative at date T, the marginal return to education is necessarily negative in the interval $[0, T]$ if $r > \theta$. In consequence, $\sigma(t) = 0$ for all $t \leq T$ if $r > \theta$. In other words, there is never an interest in educating oneself if the discount rate r is greater than the efficiency θ of educational effort. So if one is to acquire education, one must be sufficiently patient and the returns to education must be sufficiently high.

2.2.2 The Optimal Duration of Schooling

If $r < \theta$, relation (4.5) tells us that the marginal return to educational effort decreases over time. Since at date T, we have, according to equation (4.4), $\frac{\partial\Omega}{\partial\sigma(T)} = -Ah(T)e^{-rT} < 0$, there may exist a date s such that $\frac{\partial\Omega}{\partial\sigma(s)} = 0$. As $\frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{\partial\Omega}{\partial\sigma(t)}\right] < 0$, the marginal return to education is positive for $t < s$ and negative for $t > s$. This means that educational effort is necessarily null after date s. Before date s, the fact that $\partial\Omega/\partial\sigma(t) > 0$ entails that it is optimal to furnish maximum effort, or $\sigma(t) = 1$. There is thus an interest in devoting all one's time to becoming educated, $\sigma(t) = 1$, before date s and to acquire no further education, $\sigma(t) = 0$, after date s. In this case, we have $h(s) = Ah_0e^{\theta s}$ and $h(t) = h(s)$ for $T \geq t \geq s$. Since date s is defined by $\frac{\partial\Omega}{\partial\sigma(s)} = 0$, relation (4.4) allows us to obtain an explicit expression of s.

We have:

$$s = \begin{cases} T + \dfrac{1}{r}\ln\left(\dfrac{\theta - r}{\theta}\right) & \text{if } \theta \geq \frac{r}{1-e^{-rT}} \\ 0 & \text{otherwise} \end{cases}$$

This equation shows that the duration of schooling increases with the duration of life T and with the efficiency parameter θ. Hence the most efficient individuals spend the longest amount of time on education.

We can also see that the duration of schooling decreases with the discount rate r. This means that more impatient individuals, or ones facing higher financial hurdles that drive up the cost of borrowing, must study for shorter periods. We also note that s is positive only if $r < \theta(1 - e^{-rT})$, in other words, if the efficiency of education and the age of retirement are sufficiently large with respect to the discount rate. Hence it might be optimal not to get any training or education when the efficiency parameter is too small,

in which case the agent preserves the same stock of knowledge h_0 throughout her life, which procures for her a discounted gain equal to $Ah_0\left(1 - e^{-rT}\right)$.

The law of motion of the stock of knowledge (4.2) entails that human capital accumulated at the end of the training period is equal to $h_0 e^{\theta s}$, and thus that the wage of an individual of type θ is worth $Ah_0 e^{\theta s}$ at all dates $t \geq s$. This wage increases with the efficiency θ of the educational investment for two reasons. For one thing, each period of education augments the stock of human capital to a greater degree, the more efficient the individual is, and for another, more efficient individuals study longer. We also see that the wage depends on the initial stock of knowledge h_0. In this sense, "inherited" human capital influences earnings from work.

2.3 EDUCATION, TRAINING, AND LIFE-CYCLE EARNINGS

In all developed countries, for all professions, the relationship between age and annual income from employment over the life cycle presents the same characteristics (Psacharopoulos, 1985). After an initial period of education during which no wage income is received, this curve is concave and reaches a maximum between the ages of 45 and 60, before gradually tailing off. Figure 4.8 portrays this relationship for holders of high school diplomas (12 years of schooling) and college degrees (16 years of study) in the United States.

Ben-Porath (1967), Heckman (1976), and Weiss (1986) have shown that the theory of human capital explains the relationship between age and labor earnings very naturally. These authors have enriched the basic model just laid out in various ways. For example, Heckman (1976) introduced a trade-off between consumption and leisure. In what follows, we will limit ourselves to expounding the seminal model of Ben-Porath (1967), on the assumption that the marginal return to education effort is decreasing. This model in fact arrives at an earnings profile analogous to the one represented in figure 4.8.

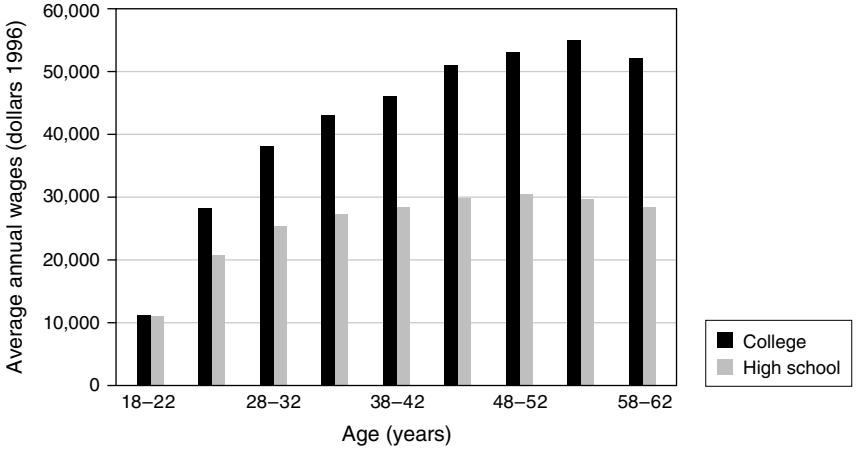


FIGURE 4.8
Average wage gains for college and high school graduates in the United States in 1996.

Source: Ashenfelter and Rouse (1999).

2.3.1 A Model with Training Over the Life Cycle

To describe the evolution of the stock of human capital, we adopt a more general representation than the law of motion (4.2) from the previous basic model. Let $\delta \geq 0$ be the rate of depreciation of knowledge; this law of motion is now defined by:

$$\dot{h}(t) = \theta g\left[\sigma(t)h(t)\right] - \delta h(t), \quad g' > 0, \quad g'' < 0 \tag{4.6}$$

In this equation, it is assumed that the efficiency of educational effort is proportional to the stock of human capital. Moreover, the previous benchmark model assumed that function g was linear. Here it is assumed that the accumulation of human capital is a concave function of effort. The purpose of this hypothesis is to obtain solutions for which $\sigma(t)$ is *strictly* comprised between 0 and 1, which signifies that at each period of his life an individual may spend part of his time receiving training and part of it working. When $\delta > 0$, an individual's human capital depreciates as his knowledge and skills become obsolete.

In this environment, a supplier of labor must choose for each date t the fraction $\sigma(t) \in [0,1]$ of his time to be dedicated to training. His problem thus consists of maximizing his discounted gains (4.3) subject to the law of motion of human capital given by equation (4.6). Let $\lambda(t)$ be the multiplier associated with this last equation; the Hamiltonian[4] of this problem is written:

$$H = A\left[1-\sigma(t)\right]h(t)e^{-rt} + \lambda(t)\left\{\theta g\left[\sigma(t)h(t)\right] - \delta h(t)\right\}$$

If we limit ourselves to interior solutions for which $\sigma(t) \in (0,1)$, the first-order conditions take the form:

$$\frac{\partial H}{\partial \sigma(t)} = 0 \Leftrightarrow -Ae^{-rt} + \lambda(t)\theta g'\left[\sigma(t)h(t)\right] = 0 \tag{4.7}$$

$$\frac{\partial H}{\partial h(t)} = -\dot{\lambda}(t) \Leftrightarrow A\left[1-\sigma(t)\right]e^{-rt} + \lambda(t)\left\{\sigma(t)\theta g'\left[\sigma(t)h(t)\right] - \delta\right\} = -\dot{\lambda}(t) \tag{4.8}$$

Optimal solutions must also satisfy the transversality condition which, in this problem with a finite horizon, comes down to $\lambda(T)h(T) = 0$. Since $h(T) > 0$,[5] the transversality condition is verified if and only if $\lambda(T) = 0$. Now, at date T of the end of life, the time dedicated to education is necessarily null. In fact, at final date T, the time spent on education shows only loss of earnings without any future gains, which implies that it is not worth spending time on education. In this case, (4.7) with $\sigma(T) = 0$

[4] See mathematical appendix B on dynamic optimization at the end of this book.

[5] Equation (4.6) can be written:

$$\dot{h}(t) + \delta h(t) = \theta g\left[\sigma(t)h(t)\right]$$

Under this form, (4.6) is a first-order differential equation where the right side is $\theta g\left[\sigma(t)h(t)\right]$. The integration of this equation gives the solution:

$$h(t) = e^{-\delta t}\left\{h_0 + \theta\int_0^t g\left[\sigma(x)h(x)\right]e^{\delta x}\mathrm{d}x\right\}$$

This relation shows clearly that $h(T) > 0$.

yields $\lambda(T) = Ae^{-rT}/\theta g'(0)$. For $\lambda(T) = 0$ to obtain, it is therefore necessary to assume that $g'(0) = +\infty$, which is a standard hypothesis.

If we substitute the expression of $\lambda(t)$ defined by (4.7) in (4.8), we arrive at the linear differential equation $\delta\lambda(t) - \dot{\lambda}(t) = Ae^{-rt}$. It appears that $\lambda(t) = Ae^{-rt}/(r+\delta)$ is a particular solution of this equation. $\lambda(t) = ce^{\delta t}$, where c is any constant, is a solution of the homogeneous equation $\delta\lambda(t) - \dot{\lambda}(t) = 0$. The general solution is obtained by adding the particular solution to the solution of the homogeneous equation, which gives us $\lambda(t) = ce^{\delta t} + Ae^{-rt}/(r+\delta)$. Finally, $\lambda(T) = 0$ yields the value of the constant c. After some calculations, we find $c = -Ae^{-(r+\delta)T}/(r+\delta)$, and the multiplier $\lambda(t)$ is thus expressed:

$$\lambda(t) = \frac{Ae^{-rt}}{r+\delta}\left[1 - e^{-(r+\delta)(T-t)}\right] \tag{4.9}$$

The multiplier $\lambda(t)$ represents the marginal value of human capital at date t. Relation (4.9) indicates that this value decreases with age to reach zero value at date T, symbolizing the end of working life. The terminal condition $\sigma(T) = 0$ and the expression (4.9) of the marginal value of capital allow us to determine the values of $\sigma(t)$ and of the stock of human capital $h(t)$ thanks to the first-order condition (4.7) and the law of motion of human capital (4.6). Wage earnings $w(t) = A[1 - \sigma(t)]h(t)$ are immediately deducible.

2.3.2 Calibration Exercises

It is not possible to arrive at completely explicit analytical expressions for functions $h(t)$ and $\sigma(t)$. Still, by taking simple functional forms and reasonable values for the parameters, this model enables us to reproduce wage earnings over the life cycle similar to those generally observed in reality. By way of illustration, figure 4.9 represents the evolution of $\sigma(t)$, $w(t)$, and $h(t)$, assuming $g(s) = s^{0.71}$, $A = 0.75$, $\delta = 0.06$, $r = 0.05$, $h_0 = 5$, $T = 60$, and $\theta = 0.5$. The model is thus calibrated on annual data with a discount factor r worth 5%. The 60-year horizon of working life is justified by the age of retirement, which is 65 in many countries, and the onset of schooling, which normally occurs at

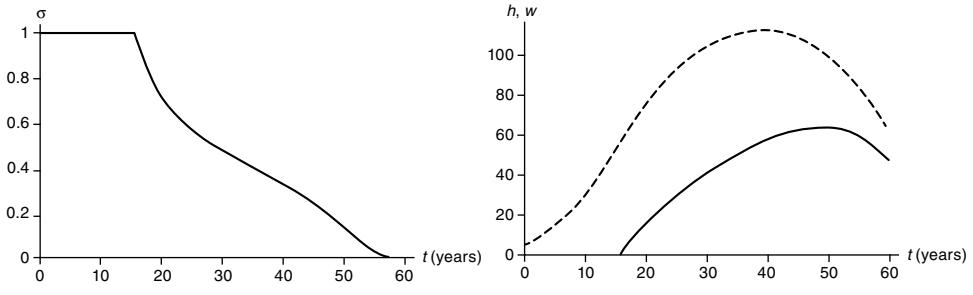


FIGURE 4.9

The law of motion of time dedicated to education (graph on the left), stock of human capital (dotted line in the graph on the right), and wage gains (solid line in the graph on the right) in the human capital model for an efficiency coefficient $\theta = 0.5$.

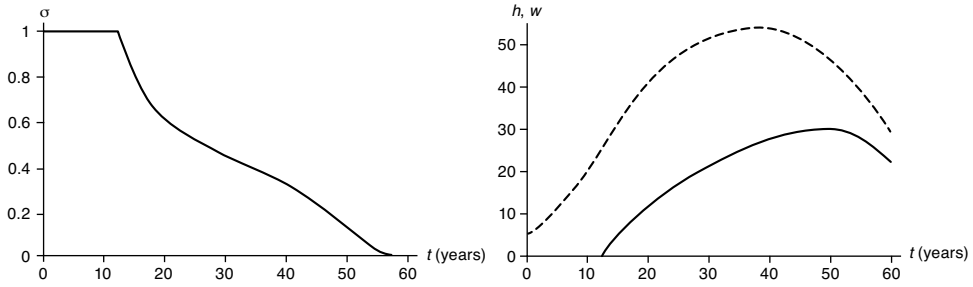


FIGURE 4.10
The law of motion of time dedicated to education (graph on the left), stock of human capital (dotted line in the graph on the right), and wage gains (solid line in the graph on the right) in the human capital model for an efficiency coefficient $\theta = 0.4$.

around age 5. Figure 4.9 reproduces very accurately the duration of schooling and the evolution of wage earnings for holders of a degree from a college in the United States. It shows that individuals follow a full-time course of studies—$\sigma(t) = 1$—for 16 years, but after that they invest less and less in training. The profile of wage earnings is increasing and concave and reaches a maximum of \$60,000 at around 10 years before retirement.

Interestingly, it is possible to represent the difference between the behaviors and the earnings of college and high school graduates by modifying the value of the efficiency parameter θ exclusively. Figure 4.10 does indeed show that when θ has the value 0.4 (the values of all the other parameters remaining unchanged), we obtain a wage profile and a duration of full-time study corresponding to those of a high school graduate: schooling lasts only 12 years, and the wage reaches a maximum of a little under \$30,000 at around 10 years before retirement.

These results show that in this model of human capital, the heterogeneity in abilities reflected by parameter θ explains to a large extent both educational behavior and the labor earnings that flow from it.

2.3.3 EXTENSIONS OF THE HUMAN CAPITAL MODEL

The model of human capital in which individuals choose the time they wish to dedicate to training reproduces very well the time path of earnings over the life cycle. Various extensions of this model have been proposed for the purpose of explaining other characteristics of the professional life of an individual.

For example, the amount of hours worked and hourly earnings vary over the course of the life cycle. In a typical profile, the hourly wage begins by increasing and reaches a maximum before retirement. The amount of hours worked also increases at the outset but peaks earlier than the hourly wage. By introducing hours worked into the human capital model, we are able to take these characteristics into account. To that end, we must assume, as in chapter 1, that the preferences are represented by a utility function $U(C, L)$ increasing with consumption C and leisure L. It is then possible to show

that choices among consumption, leisure, and investment in human capital lead to profiles of hourly wages and length of time worked similar to those observed in reality (see Weiss, 1986, for a synthesis). If we take "learning by doing" into account, these effects are reinforced. Under these conditions, hours spent working are also a method of learning and thus of improving productivity. There is therefore an interest in working a great deal at the outset of the life cycle so as to build up experience, then reducing one's working schedule at the end of it.

In practice, choices about education and training are made in an uncertain environment. Intuition suggests that these choices depend on the manner in which uncertainty affects the returns to education in relation to other possible sources of income. If the returns to education are little affected by uncertainty compared with other modes of earning, a supplementary investment in human capital becomes a way of hedging against risk. Rising uncertainty can thus augment the accumulation of human capital in certain cases (see Weiss, 1986).

As we pointed out at the beginning of this section, the theory of human capital rests on the hypothesis that wage differences reflect productivity differences, which are themselves influenced by the acquisition of competences by workers. The theory of human capital should thus allow us to gain insight into numerous aspects of individual decisions about education. But this conception of education is not uncontested: another theory assigns it the mere function of sending a "signal."

3 EDUCATION AS A SIGNALING DEVICE

The positive correlation, highlighted in figure 4.5, between duration of studies and earnings does not prove the existence of a causal impact of education on productivity. It is not, in fact, beyond dispute that education permits the accumulation of directly productive knowledge. The ability to resolve differential equations or to understand all the subtleties of Keynesian macroeconomics does not necessarily increase the productivity of a person working in a firm or an agency. On this basis, Spence (1973) put forward the idea that education also—and perhaps even primarily—serves to select individuals, without really influencing the productive efficiency that they will display in their future professional lives. The productive efficiency of a person is seen as a sort of intrinsic quality, which may certainly depend on a wide range of factors (family milieu, personal history, innate qualities or talents, etc.), but over which education exerts little influence.

The premise of Spence's theory is that those persons who perform most effectively in active life are also the ones who perform best while studying. If productive efficiency is not observable by potential employers, then success as a student simply serves to signal the presence of such productive characteristics—hence the term *theory of signaling* given to this view of education. From this standpoint, a person pursues education in order to signal her efficiency, without her studies really modifying this efficiency. If education serves only to signal intrinsic individual qualities, then the real significance of the positive correlation between duration of studies and earnings is just that more efficient individuals have higher earnings. The standpoint of the theory of human capital is

completely at odds with that of signaling theory because for the latter a prolongation of one's studies does nothing to increase one's productive capacity; all it does is to send out a signal to employers. Signaling theory also arrives at very different conclusions concerning the efficiency of investments in education. Whereas the theory of human capital indicates that individual decisions with regard to education are socially efficient under perfect competition, Spence (1973) shows that workers have a tendency to overeducate themselves with respect to the standard of social efficiency, if education does serve to signal their productive capacities to employers.

In this section, we present a model in which employers observe the productivity of workers imperfectly but view an educational degree, or the length of time spent in schooling, as an indicator of potential productivity. In this context, workers may have an interest in investing in education in order to "signal" their abilities to employers. This aspect of education may lead, under certain circumstances that we will highlight, to overinvestment in training.

3.1 A Model with Signaling

We here consider a labor market made up of a continuum of individuals whose productive abilities are different. The size of the continuum is normalized to one. A worker with ability h can produce h units of a good. For simplicity, we will now assume that there are only two levels of personal ability, h^+ and h^- with $0 < h^- < h^+$.

Workers do have the possibility to achieve a level of education $s \geq 0$ that is observed by employers. A level of education s bears a cost equal to s/h. Thus, the weaker the productive abilities of workers are, the more it costs. It should be noted that in this model education does not improve individual productivity; it can serve only to signal ability when it is not observed by employers. At a later stage, we will examine the consequences of education when it fulfills more than one function. The preferences of workers are represented by a utility function $u(R, s, h) = R - (s/h)$, where R designates earnings, equal to wage w if the individual is employed and to 0 otherwise.

We assume that decisions unfold in the following sequence:

1. workers, knowing which of the two types they belong to, choose their level of education s,

2. firms enter the labor market freely, observe the signals s, and make simultaneous wage offers to workers, and

3. workers accept or refuse the offers made to them.

Let us first consider a situation of perfect competition in which individual characteristics are perfectly observed. The hypothesis of free entry entails $w(h) = h$, for $h = h^-, h^+$. Since we have assumed that workers get zero earnings when they do not work and that the disutility of working is zero, hypothesis $h^- > 0$ entails that all workers are employed independently of the signal s which they may send. In consequence, in the first stage of the sequence of decisions, no one has any interest in using resources

to send a signal $s > 0$ and so they all choose a zero level of education. This situation is efficient, for $s > 0$ does not augment productivity.

3.1.1 Equilibrium When Ability Is Unobservable

When abilities are unobservable, on the other hand, the signal becomes a way for the most efficient workers to bring themselves to the attention of firms. To that end, it is sufficient for them to choose a level of education that is too costly for inefficient workers, given the wage differential $w(h^+) - w(h^-)$. In that case, firms are capable of distinguishing between the two types of workers according to their respective signals, and the equilibrium is called *separating equilibrium*. In this situation, the condition of free entry entails $w(h) = h$, for $h = h^-, h^+$, and workers with low efficiency send the signal $s = 0$, since a positive signal brings them no gain. For equilibrium actually to be separating, it must be verified that no person of type h^- has an interest in deviating by choosing a signal identical to that sent by more efficient persons. By sending a zero signal, a worker of low efficiency obtains a utility $u[w(h^-), 0, h^-] = h^-$, while by sending a signal s^+ identical to that of efficient workers, he obtains $u\left[w(h^+), s^+, h^-\right] = h^+ - (s^+/h^-)$. Hence a worker of low efficiency has no interest in sending a signal identical to that of more efficient workers if $h^+ - (s^+/h^-) \leq h^-$, which is equivalent to $s^+ \geq h^-.(h^+ - h^-)$. Knowing that, workers of type h^+ have an interest in sending the weakest signal possible, which workers of type h^- have no interest in imitating. This signal thus has the value $s^* = h^-.(h^+ - h^-)$. Evidently efficient workers prefer $s = s^*$ to $s = 0$, since workers of type h^-, whose signaling costs are greater, are indifferent between these two values of s. So in this economy there does exist a separating equilibrium in which workers of low efficiency do not seek education and obtain a wage $w(h^-) = h^-$, and in which efficient workers become educated to a level $s^* > 0$ and obtain a wage $w(h^+) = h^+$.

It is important to emphasize that, even in this simple model, the separating equilibrium just described is not the sole equilibrium possible. In fact, the definition of equilibria in signaling games raises difficulties having to do with the beliefs of agents (for an accessible and very thorough discussion of this subject, see Mas-Colell et al., 1995, chapter 13). In general, it is necessary to choose a very restrictive concept of equilibrium in order to eliminate outcomes which appear to have no relevance. It is also necessary to know that Cho and Kreps (1987) have proposed a criterion to be applied in situations of this type and known as the *intuitive criterion*, which results in only the separating equilibrium described here being maintained. In our elementary model, we implicitly selected the most efficient separating equilibrium, the one that corresponds to the smallest value of the signal that still makes it possible to distinguish between the two types of worker. Other separating equilibria exist in which the values of the signal are greater than s^*. Equilibria of this kind are eliminated if the intuitive criterion is used.

3.1.2 The Inefficiency of Education as a Signaling Device

In the example we gave, it is easy to show that education is a waste of resources that has no social utility. To reach that conclusion, it is enough to compare the allocations

obtained with and without the opportunity to become educated when individual abilities are not observable.

Let λ be the proportion of efficient workers, and let us begin by analyzing the situation in which education is absent and workers are indistinguishable. Since the opportunity cost of labor is assumed to be zero, and since $h^+ > h^- > 0$, everyone participates in the labor market and obtains an identical wage w given by $w = E(h) = \lambda h^+ + (1-\lambda)h^-$. Normalizing the number of workers to 1, total output is then equal to $E(h)$.

Now let us introduce the opportunity to get an education. At the separating equilibrium, in which the efficient workers get educated, overall production *net* of the costs of education is equal to the difference between gross production $E(h)$ and the costs of education, equal to $\lambda s^*/h^+$. In this case, education is clearly a waste of resources, one moreover that has detrimental redistributive effects for the least efficient individuals. These obtain a utility equal to $u(w, 0, h^-) = E(h)$ or $u[w(h^-), 0, h^-] = h^-$, in the absence and presence respectively of education. Workers with low productivity are thus systematically disadvantaged by education. On the other hand, education has an ambiguous effect on the welfare of the most productive persons, who obtain a utility equal to $u(w, 0, h^-) = E(h)$ or $u\left[w(h^+), s^*, h^+\right] = \left[(h^+)^2 - h^+.h^- + (h^-)^2\right]/h^+$ in the absence and presence respectively of education. What this means is that education improves the situation of efficient workers if and only if $u\left[w(h^+), s^*, h^+\right] > u(w, 0, h^-)$, which is equivalent to $\lambda < (h^+ - h^-)/h^+$. Efficient workers thus benefit from education if their proportion is sufficiently small with respect to the efficiency gap between them and the less productive workers.

So the model of Spence (1973) portrays the role played by education in a very negative light: all it serves to do is select workers according to their efficiency, without improving the allocation of resources. This result is not a general one, however, and the model that follows offers a case in which signaling activity makes it possible, under certain circumstances, to improve the allocation of resources.

3.1.3 The Efficiency of Education as a Signaling Device

For education to become an efficient signaling device, all we have to do is adjust the preceding model at the margin by assuming that the opportunity cost of labor is something other than zero. The preferences of workers are now represented by the utility function $u(R, s, d, h) = R + d - (s/h)$, where R designates earnings, equal to wage w if the individual is employed and 0 otherwise, d is an indicator function amounting to 0 if the individual is employed and 1 if not, and the signal s still stands for the level of education. Let us further assume that the individual characteristic h takes only two values, h^- and h^+, such that $0 < h^- < 1 < h^+$, with $E(h) < 1$. Under these hypotheses, when abilities are not observable and there is no signaling activity, nobody enters the labor market, since the wage compatible with free entry, $w = E(h)$, is less than the opportunity cost of labor. Such a situation arises when the proportion of workers whose productivity h is less than the opportunity cost of labor is large. The opportunity of using a costly signaling device may then allow the most efficient persons to enter the market and so improve the allocation of labor. Let us take a closer look at this situation.

When the equilibrium is separating, workers with low efficiency stay out of the market because their productive ability h^- does not permit them to obtain a wage greater

than the opportunity cost of labor (free entry dictates $w(h^-) = h^- < 1$). These workers therefore send a zero signal s, since a positive signal brings them no gain. For equilibrium actually to be separating, it must be verified that individuals of low efficiency have no interest in choosing a signal identical to that of more efficient workers. By sending a zero signal, a low-efficiency worker attains utility $u(0, 0, 1, h^-) = 1$. By sending a signal s^+ identical to that of efficient workers, he or she obtains $u\left[w(h^+), s^+, 0, h^-\right] = h^+ - (s^+/h^-)$. Consequently, a low-efficiency person has no interest in sending a signal identical to the one sent by an efficient person if $h^+ - (s^+/h^-) \leq 1$, which is equivalent to $s^+ \geq h^-(h^+ - 1)$. Knowing that, workers of type h^+ have an interest in sending the smallest signal that workers of type h^- have no interest in imitating. This signal is given by $s^* = h^-(h^+ - 1)$. As in the preceding model, it is clear that efficient workers prefer $s = s^*$ to $s = 0$, since individuals of type h^-, for whom signaling is more costly, are indifferent between these two values of s. This separating equilibrium dominates, according to the Pareto criterion, the equilibrium without signaling, since the less efficient workers obtain the same level of gain in the two equilibria—equal to $u(0, 0, 1, h^-) = 1$—while the more efficient workers obtain $u\left[w(h^+), s^*, 0, h^+\right] = \left[(h^+)^2 - h^+h^- + h^-\right]/h^+$ in separating equilibrium, which procures them a gain exceeding the opportunity cost of labor when $h^+ > 1 > h^-$.

3.2 Overeducation or Undereducation?

The previous example has shown that education might, through its role as a signal, improve the allocation of resources in certain circumstances. But this signaling role may also lead to "too much" education in relation to what the collective optimum requires. In this case, it is generally desirable to reduce signaling through cross-subsidization, financed by lump-sum taxes. This policy consists of reducing the earnings differential between workers with different signals so as to reduce the incentive to acquire education, while preserving positive levels of education.

3.2.1 A Model with Cross-Subsidies

To grasp the effect of cross-subsidies, a graphic representation of the model just laid out will be helpful. In the plane (w, s), the indifference curves identified by u^+ and u^- in Figure 4.11 apply respectively to workers of type h^+ and type h^-. As the slopes of the indifference curves, dw/ds, are equal to $(1/h)$, less efficient workers have more steeply sloped indifference curves. Moreover, the upward shift of an indifference curve corresponds to an improvement in satisfaction. In the absence of cross-subsidization, the separative equilibrium of the previous subsection corresponds to situation A, in which the most efficient individuals obtain a wage h^+ and choose a level of education s^*, and less efficient individuals stay out of the labor market and obtain a gain of $d = 1$.

It is possible to improve this situation by declaring that workers whose level of education is at least equal to s^1 receive wage w^1 and that workers whose level of education is less than s^1 receive a subsidy of amount x if they do not work. This situation, labeled B in Figure 4.11, is preferred by both types of workers to situation A. What is more, it limits the expenditures arising from signaling while allowing firms to make the distinction between the two types of wage earners, since the less efficient workers have no interest in imitating the more efficient workers by getting an education.

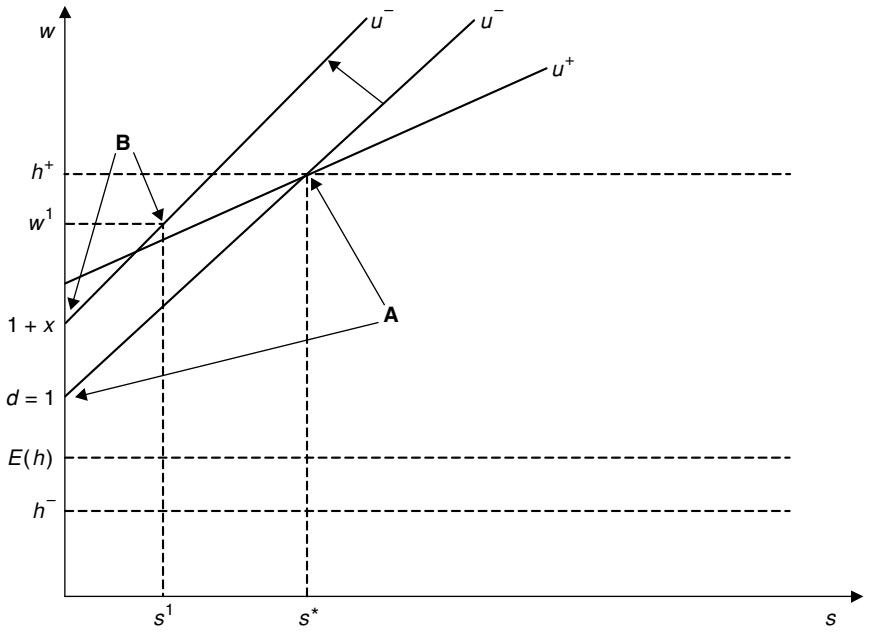


FIGURE 4.11
Overeducation in the model of Spence (1973).

Cross-subsidies are thus a means of limiting the incentives to overeducation. In our model, it is even possible to curb these incentives very drastically by causing outlays on education to remain positive but tend to 0.

To be compatible with a separating equilibrium that allows employers to distinguish among types of workers, the wage $w(s)$ linked to a level of education s and the subsidy x accorded to individuals having a level of education inferior to s and not participating in the labor market, must satisfy the conditions $w(s) - (s/h^-) \leq 1 + x$ for inefficient workers not to seek education, and $w(s) - (s/h^+) \geq 1 + x$ for efficient workers to do so. In consequence, a separating equilibrium is compatible with a value of x lying in the interval $\left[w(s) - (s/h^+) - 1, w(s) - (s/h^-) - 1\right]$, and it proves possible to define a value of s arbitrarily close to 0, such that there exists a value of x falling within this interval. When $s \to 0$, we get $x = w(s) - 1$, which means that the choice of any wage $w(s) \in \left[1, h^+\right]$ and a subsidy $x = w(s) - 1$ leads to a separating equilibrium with a signal the cost of which is arbitrarily low. Such cross-subsidies, tied to an infinitesimal signal cost, allow us to obtain, at the limit, an efficient equilibrium similar to the one that arises in the absence of the problem of adverse selection.

3.2.2 Is There Really Overeducation?

In practice, limiting investment in education through cross-subsidies is only desirable if education is doing nothing but offer a signaling service, and if the most efficient workers are getting overeducated. Two sorts of reasons make it doubtful that this is a valid representation of education.

For one thing, it is evident that the signaling services supplied by education do not necessarily lead to overeducation of the most efficient workers. The result that there is overeducation rests on the hypothesis that employers do not hire individuals while they are still in school. The model of Spence (1973) does indeed assume that individuals must necessarily finish their schooling before presenting themselves on the labor market. This hypothesis has been criticized on the grounds that employers may have an interest in "intercepting" good students and hiring them before the completion of their schooling (Weiss, 1983). Intuition then suggests that if all employers offer to hire students enrolled in long and difficult courses of study on the same day they enroll, education can no longer play any signaling role. Swinkels (1999) has shown that there are probably good grounds for this intuition: introducing an opportunity to make confidential hiring offers to students before they have completed their course of study into Spence's model entails a *pooling equilibrium* characterized by an absence of outlay on education and workers obtaining a remuneration equal to their expected productivity. Nonetheless, when education increases individual productivity (as above, unobservable by employers), individuals may have an interest in acquiring education. Swinkels (1999) shows that it is persons endowed with less ability who have a tendency to overeducate themselves in order to mimic more efficient workers. The latter never overeducate themselves and may even choose a level of education inferior to the one they would opt for in a situation of perfect information if they cannot be distinguished from workers of low efficiency. These results, which are at variance with those originally obtained by Spence (1973), show that the education system does not lead to a systematically excessive use of resources, even when it is simply acting as a signaling device. They also bring out the fragility of the predictions of models with asymmetric information, the properties of which appear highly sensitive to the manner in which the strategies of agents are represented.

In light of the theory of human capital, there is reason to doubt that public interventions to limit outlays on education are required because according to this theory, education makes it possible to accumulate knowledge and thus supplies other services besides that of sending a signal. These two dimensions of education are in fact difficult to separate, but the numerous empirical studies dedicated to the problem suggest that education does improve individual efficiency (see section 4.3 below).

4 IDENTIFYING THE CAUSAL RELATION BETWEEN EDUCATION AND INCOME

The estimation of earnings functions, the goal of which is to evaluate the returns to education, constitutes the basis of empirical work dedicated to education. This type of estimate, which brings out a *correlation* between education and income earned through work, has stimulated a large quantity of research aimed at finding out whether this correlation betrays a *causal link* between education and earnings. This research tries to determine if education serves to accumulate knowledge that has value in the labor market—as in the theory of human capital—or if its main function is to select the most efficient individuals, without teaching them a great deal.

4.1 The Theory of Human Capital: From the Model to Estimates

The main prediction of the theory of human capital is that education is the source of an accumulation of competences that make it possible to increase earnings. The assessment of this result is done by estimating earnings functions, which relate earnings to investment in education. Mincer (1974) proposed a form of earnings function deduced from the theoretical model presented in section 2.2 that arrives at an estimate of the internal rate of return to educational investment. The precision of these estimates is noticeably increased by taking experience into account.

4.1.1 The Internal Rate of Return to Education

The internal rate of return on an investment is the rate of return that makes the net present value of all benefits and costs from a particular investment equal to zero. In other words, the internal rate of return of an investment is the discount rate at which the net present value of costs equals the net present value of the benefits of the investment. Internal returns are commonly used to evaluate the desirability of investments. Imagine that you must borrow to finance a project. It is worth investing in this project if its internal rate of return is higher than the interest rate at which you must borrow. Since education is an investment, it makes sense to evaluate the internal rate of return to education as for any other investment. This is exactly what empirical studies do: they estimate the internal rate of return to education, which is generally merely called the rate of return to education. Let us now show how the internal rate of return to education depends on the benefits and on the costs of education.

Let us first put ourselves in the position of a person who acquires education at the outset of his life but ceases to do so for good once he starts to work. By definition, the internal rate of return to education, denoted ρ, represents the discount rate that equalizes the cost and the expected gain of investing in education.

Let $w(s)$ be the potential income associated with an amount of time s spent in school. If we assume, for the sake of simplicity, that the cost of education is identical to the loss of potential earnings undergone during the time s spent in school, the cost of education at date s is simply equal to $w(s)$. This cost makes it possible to increase earnings by an amount $\dot{w}(s)$ at every future date. Let T be the date at which working life comes to an end; the present value at date s of the incremental gain $\dot{w}(s)$ discounted at rate ρ is given by $\dot{w}(s)\int_s^T e^{-\rho(\tau-s)}d\tau = \dot{w}(s)\left[1-e^{-\rho(T-s)}\right]/\rho$. The internal rate of return to education equalizes the gain and the cost and is thus defined by the equation:

$$\frac{\dot{w}(s)}{w(s)} = \rho\frac{1}{1-e^{-\rho(T-s)}} \tag{4.10}$$

If we assume that T is much greater than s, the right-hand side of this equation is approximately equal to ρ, and we see that earnings satisfy the differential equation $\rho = \dot{w}(t)/w(t)$. Integrating this last equation, we finally get:

$$\ln w(s) = \ln w(0) + \rho s \tag{4.11}$$

This equation defines a linear relation between the log of earnings and the duration of education. If time is expressed in years, the internal rate of return ρ can be interpreted as the relative increase in earnings flowing from an extra year of schooling.

In practice, we can observe, for each individual i, his earnings w_i and his number of years of education t_i. However, his potential earnings if he had had no education, $w_i(0)$, are not observed. With this information, Mincer estimated equation (4.11) under the following form:

$$\ln w_i = a + \rho s_i + \varepsilon_i \tag{4.12}$$

In this expression, w, s, and ε designate respectively the earnings of individual i, his duration of studies, and an error term of zero mean reflecting the heterogeneity of individuals. The coefficients a and ρ are parameters to be estimated. The ordinary least squares (OLS) estimator of the returns to education, ρ, is unbiased if s and the error term ε are independent, which means formally that $Cov(s, \varepsilon) = 0$. But as we have just seen, the theoretical models suggest that individual capacities (measured by the term ε) influence the duration of studies, so the two terms s and ε are not independent. Therefore the estimator of the returns to education by ordinary least squares is biased. We will see below that labor economists have devoted much effort to dealing with this issue and have been able to imagine clever solutions. But let us look, in a first step, at the results obtained by Mincer in his seminal study.

The first line of table 4.1 presents the estimate of equation (4.11) obtained by Mincer (1974) using data concerning white men in the United States in 1959. It is clear that the length of time spent in school has a significant positive effect on earnings. The rate of return to an extra year of schooling is 7%. Still, the coefficient of determination, R^2, indicates that this equation explains less than 7% of the variation of the logarithm of earnings. Mincer suggests that it is possible to improve this performance by accounting for professional experience and the accumulation of human capital that takes place after leaving formal schooling behind.

4.1.2 The Importance of Experience

To improve his estimates, Mincer makes the assumption that it is possible to acquire education while employed. The life-cycle model of human capital accumulation set out in section 2.3 does in fact suggest that it is optimal to begin with full-time schooling, then gradually diminish the proportion of one's time dedicated to schooling from the point one enters the labor force. Let $t(\tau) \in [0, 1]$ be the portion of time dedicated to further training by a person with τ years of experience who has already spent s years in school. As in the theoretical model of section 2.3, we assume that the law of motion of the human capital $h(s + \tau)$ of this person is described by the differential equation:

$$\dot{h}(s + \tau) = \rho_x t(\tau) h(s + \tau), \quad \forall \tau \in [0, T - s]$$

TABLE 4.1

Estimates of wage equations; s designates the duration of schooling, x experience (measured by age minus the duration of schooling minus 6 years), and w the annual earnings of white men working in the nonagricultural sector in the United States in 1959 (t-statistics in parentheses).

$\ln w = \underset{}{7.58} + \underset{(43.8)}{0.070}s$	$R^2 = 0.067$
$\ln w = \underset{}{6.20} + \underset{(72.3)}{0.107}s + \underset{(75.5)}{0.081}x - \underset{(-55.8)}{0.0012}x^2$	$R^2 = 0.285$
$\ln w = \underset{}{4.87} + \underset{(23.4)}{0.255}s - \underset{(-7.1)}{0.0029}s^2 + \underset{(63.7)}{0.148}x - \underset{(-66.2)}{0.0018}x^2 - \underset{(-31.8)}{0.0043}xs$	$R^2 = 0.309$

Source: Mincer (1974, table 5.1).

In this expression, the constant coefficient ρ_x is interpretable as the rate of return to training after leaving school. The integration of this differential equation between dates $\tau = 0$ and $\tau = x$, then gives $h(s+x) = h(t)e^{\rho_x \int_0^x t(\tau)\mathrm{d}\tau}$. Assuming again that earnings $w(s+\tau)$ are equal to $A[1-t(\tau)]h(s+\tau)$, the earnings $w(s+x)$ of a person with x years of experience depend on her earnings $w(s)$ upon leaving school and on her time devoted to further training according to the formula:

$$w(s+x) = [1-t(x)]w(s)e^{\rho_x \int_0^x t(\tau)\mathrm{d}\tau} \tag{4.13}$$

To arrive at an explicit wage equation, Mincer assumes $t(x) = t_0 - t_0(x/T)$. Under this hypothesis, the fraction of time dedicated to the accumulation of human capital decreases in linear fashion with the amount of time passed since leaving school. We then have $\int_0^x t(\tau)\mathrm{d}\tau = t_0 x - (t_0/2T)x^2$. Taking the logarithms of the two sides of relation (4.13) and bearing in mind that income $w(s)$ after s years of schooling satisfies the law of motion (4.11), we arrive at the wage equation:

$$\ln w(s+x) = \ln w(0) + \rho s + \rho_x t_0 x - \rho_x (t_0/2T)x^2 + \ln[1-t(x)] \tag{4.14}$$

It should be noted that the variable x representing experience has an ambiguous status, for experience can result not only from—as we assume here—an investment that eats into efficient working time (*learning or doing*) but also from an accumulation of know-how that the worker builds up during her efficient working time (*learning by doing*). In the latter case, we can make the assumption that a worker acquires a significant amount of supplementary knowledge on the job at the beginning of her career and that such supplements in knowledge then tail off over time. That being so, it is sufficient to assume $t(x) = 0$ in (4.14).

The second line of table 4.1 presents the results of the estimation of equation (4.14) leaving out the term $\ln[1-t(x)]$. It indicates that bringing experience into the mix considerably improves the explanatory power of the earnings function. This function now explains around 30% of the variation of the logarithms of earnings, as opposed to 7% earlier. Further, comparison of the first two lines of table 4.1 shows that the rate of return to formal schooling is greater than that obtained by leaving experience out. Leaving experience out biases the estimate of the returns to formal schooling downward because schooling and experience are *negatively* correlated (those with the most experience are also those who leave school earliest). Hence, to estimate the return to education while leaving out the return to experience amounts to neglecting the fact that at a certain age an extra year of schooling means one less year of experience. This omission leads to an estimate of the return to education from which the return to experience is *subtracted*, since the fact that persons who dedicate an extra year to schooling necessarily have one less year of experience is not taken into acccount.

4.1.3 The Importance of the Duration of Schooling

The earnings function defined by equation (4.14) is grounded on the hypothesis of a constant rate of return to formal schooling, equal to ρ. This hypothesis is debatable, for the impact of education very likely varies with the duration of schooling. The third line of table 4.1 takes this possibility into account by introducing a quadratic term s^2 and a term of interaction sx between experience and the duration of schooling. We see that

the rate of return to education decreases with the duration of schooling. We also see that there is a negative interaction between the duration of schooling and experience, which would tend to prove that the return to experience decreases with the duration of schooling. Mincer (1974) shows, however, that this result is not significant when income is measured in weekly earnings.

This presentation of estimation procedures of the returns to education gives us an overview of the method followed by the seminal work of Mincer. This method has been refined in several respects, in particular in order to analyze in more depth the causal relation between education and income.

4.2 The Selection Problem

The correlation between duration of schooling, or more generally investments in training, on one hand, and earnings on the other, of the kind revealed in table 4.1, does not signify that there exists a *causal* relation between these two variables. Indeed, the model of human capital presented in section 2.2 shows that individual capacities (measured by the parameter θ in this model) influence both wages and the duration of studies. In addition, according to the theory of *signaling* (see section 3.1), education plays a filtering role, serving to select those workers who are innately efficient and to signal productive characteristics of workers that employers cannot directly observe. That being so, the correlation between duration of schooling and earnings would stem from the fact that the most efficient individuals have higher earnings and stay in school longer.

4.2.1 Ability Bias

The theory of human capital and signaling theory both do predict that the most productive individuals have an interest in studying for the longest period. This entails the possibility of the so-called ability bias, which means that the return attributed to education may come, in fact, from individual capacities. In these cirumstances, the returns to education estimated with the ordinary least squares (OLS) method may be overestimated.

In order to observe this, and to understand how to pin down a cause-and-effect relation between duration of study and earnings, it is helpful to start once more with the gains equation (4.12). By definition, the population regression coefficient ρ minimizes the expected squared errors in the population as a whole, $\mathbb{E}(\mathrm{w}-\rho s)^2$ where $\mathrm{w} = \ln w$. This coefficient is given by

$$\rho_{OLS} = \frac{Cov(s, \mathrm{w})}{Var(s)} \tag{4.15}$$

where $Var(s)$ stands for the variance of s and $Cov(s,\mathrm{w})$ the covariance of s and w. The OLS estimator of the population regression coefficient is obtained using the sample analog of the population regression coefficient:

$$\hat{\rho}_{OLS} = \frac{\sum_i (s_i - \bar{s})(\mathrm{w}_i - \overline{\mathrm{w}})}{\sum_i (s_i - \bar{s})^2}$$

where index i stands for the observation of variables belonging to individual i present in the sample and $\bar{s}$ and $\overline{\mathrm{w}}$ designate the sample mean of s and w respectively.

On the other hand, equation (4.12) implies:

$$Cov(s, \mathrm{w}) = \rho Var(s) + Cov(s, \varepsilon)$$

and then, together with equation (4.15):

$$\rho_{OLS} = \rho - \frac{Cov(s, \varepsilon)}{Var(s)}$$

This equation shows that the population regression coefficient ρ_{OLS} is equal to the parameter ρ of the Mincer equation only if the length of education is independent of the error term ε. Now, the Mincer equation incorporates individual capacities that are not observed into this error term. Hence the hypothesis of independence between the error term and duration of study is highly unlikely to be verified because, as the human capital and signal models suggest, the most capable persons have an incentive to pursue lengthier studies.

This selection problem, which has classic status in econometrics, has been addressed in numerous contributions aiming to evaluate the causal impact of education on earnings. Card (1999), Blundell et al. (2005), and Blundell and Costa Dias (2009) present syntheses of the methods employed. Here we present the common method of instrumental variables, basing ourselves on the widely cited article of Angrist and Krueger (1991). Data and programs allowing readers to replicate the main results of this paper are available at www.labor-economics.org.

4.2.2 The Instrumental Variable Method

The instrumental variable method consists of estimating the returns to education using a variable that influences the duration of studies while remaining independent of individual capacities. Let us assume that we know a variable z, which is correlated with the duration of study, that is, $Cov(z, s) \neq 0$, but that is independent of the error term ε in equation (4.12), that is, $Cov(z, \varepsilon) = 0$. A variable possessing these properties is called an instrumental variable or simply an instrument. The assumption $Cov(z, \varepsilon) = 0$ is called the *exclusion relation* because it assumes that the intrumental variable is excluded from the causal relation to be estimated. Since equation (4.12) implies that:

$$Cov(z, \mathrm{w}) = \rho Cov(z, s) + Cov(z, \varepsilon)$$

we get, with the assumption that $Cov(z, \varepsilon) = 0$, the (population) instrumental variable regression coefficient:

$$\rho_{IV} = \frac{Cov(z, \mathrm{w})}{Cov(z, s)} = \frac{Cov(z, \mathrm{w})/Var(z)}{Cov(z, s)/Var(z)}$$

This coefficient is simply equal to the ratio between the covariance between earnings and the instrumental variable and the covariance between duration of study and the instrumental variable. The second equality of the last equation shows that ρ_{IV} is also equal to the ratio between the population regression coefficient of the earnings on the

instrument and the population regression coefficient of the duration of schooling on the instrument.

The estimator of the instrumental variable regression coefficient is obtained using the sample analog of ρ_{IV}:

$$\hat{\rho}_{IV} = \frac{\sum_i (z_i - \bar{z})(w_i - \overline{w})}{\sum_i (z_i - \bar{z})(s_i - \bar{s})}$$

It must be emphasized that the instrumental variable method is valid only if the instrumental variable is indeed independent of the error term yet correlated with the duration of studies. The difficulty of this approach thus lies in finding such a variable. In this respect, Angrist and Krueger (1991) have made an interesting contribution, which consists of exploiting the existence of events that are much like natural experiments.

4.2.3 A Natural Experiment Inducing Changes in Compulsory School Attendance

Angrist and Krueger (1991) noted that individuals born early in the calendar year have shorter durations of schooling than those born later. This effect is owing to the compulsory duration of schooling. In the United States, school districts typically require a student to have turned age 6 by January 1 of the year in which he enters school. Compulsory schooling laws generally require students to remain in school until their 16th or 17th birthday. Therefore, two persons born in the same year begin school on the same date, but the one born earlier is authorized to quit school earlier than the other. If we assume that the date of one's birth is independent of factors influencing abilities and preferences, this phenomenon can entail an exogenous variation in the duration of schooling, which may be used as an instrumental variable. At first sight, this assumption, which is the so-called exclusion restriction, seems to make sense to the extent that it is unlikely that one's birthday is correlated with individual characteristics other than age at school entry. We will see below how we can provide evidence on this issue.

To implement the instrumental variable method, it is important to show first that the instrument is correlated with the endogenous variable or in other words that the season of birth is correlated with the duration of schooling. Angrist and Krueger rely on a variety of data sets constructed from the U.S. Public Use Census data in 1960, 1970, and 1980. The samples consist of men born between 1920 and 1959. Here, we focus on men born between 1930 and 1939 whose income is observed in the 1980 census. The duration of education is measured with *completed* years of schooling. Thus, individuals who were born before the month of June may leave school before the end of their 10th year of schooling, and individuals born after the month of September must attend school until the start of their 11th year of schooling but have no obligation to stay on until the end of their 11th year. Still, since they must at least start their 11th year of schooling, they may have greater incentive to complete it once they have begun it, and even to stay on and finish high school by completing their 12th year, than individuals born before the last quarter, who are allowed to quit school before they even start their 11th year.

Figure 4.12 shows that individuals born early in the year study for shorter lengths. The differences in duration of study between successive quarters of birth are significant at the 1% threshold for the first three quarters and at the 5% threshold for the last two. To verify that these differences are indeed linked to the obligatory duration of schooling, it is possible to compare the linkage between the duration of study and the

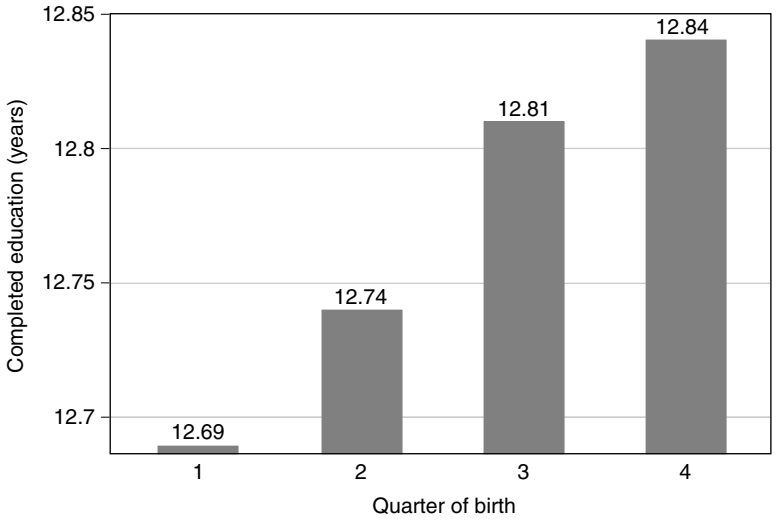


Figure 4.12
Quarter of birth and years of completed education for men born in 1930–1939.

Source: Angrist and Krueger (1991) data set.

quarter of birth of persons who left school at around age 16, and the duration of study and the birth quarter of persons who left school later and who ought not to be affected by the constraint imposed by obligatory attendance at school. Figure 4.13 shows that individuals born at the start of the year and who study for shorter periods, less than 12 years, leave school earlier than individuals born at the end of the year. On the other hand, this relation is not observed for individuals who study for longer than 12 years. This suggests that the relation between birth quarter and length of schooling is indeed induced by the regulations governing compulsory school attendance and not by other factors.

Still, to strengthen these results, we must take into account the fact that duration of study increases over time, as figure 4.14 shows. Such a trend implies that persons born early in the year may, on average, study for less time than those born toward the end of the year, independently of the effect of the regulations. To remove the trend in years of education across cohorts, Angrist and Krueger (1991) substract a moving average of the surrounding birth cohort's average education. They define a two-period, two-sided moving average of men born in year c and quarter j as follows:

$$MA_{cj} = \frac{S_{-2} + S_{-1} + S_{+1} + S_{+2}}{4}$$

where S_q is the average years of schooling attained by the cohort born q quarters before or after cohort (c, j). The detrended education series is simply $S_{cj} - MA_{cj}$. To quantify the effect of season of birth on years of education, they estimate regressions of the form:

$$S_{icj} - MA_{cj} = \alpha + \sum_{j}^{3} \beta_j Q_{ijc} + \varepsilon_{icj}, \quad \text{for } j = 1, 2, 3$$

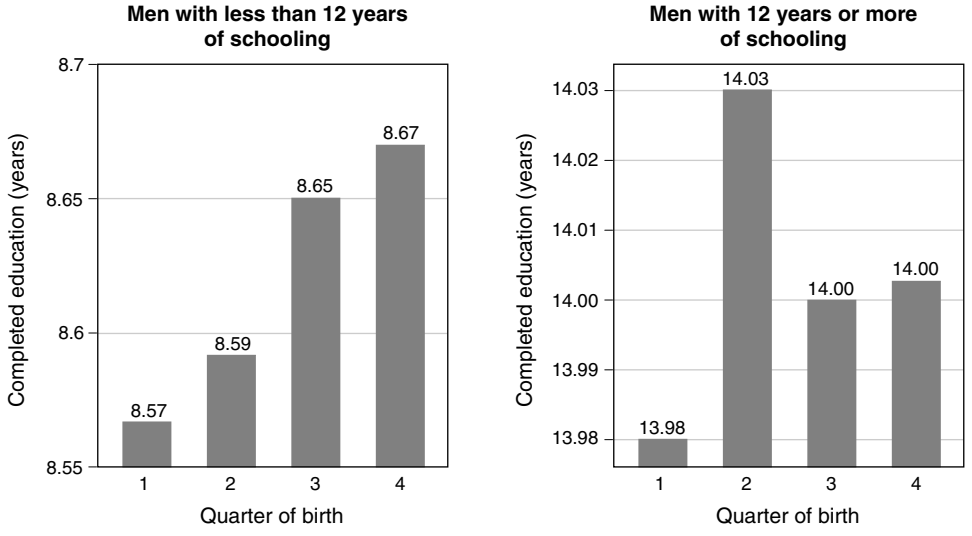


FIGURE 4.13
Quarter of birth and years of completed education for men born in 1930–1939.

Source: Angrist and Krueger (1991) data set.

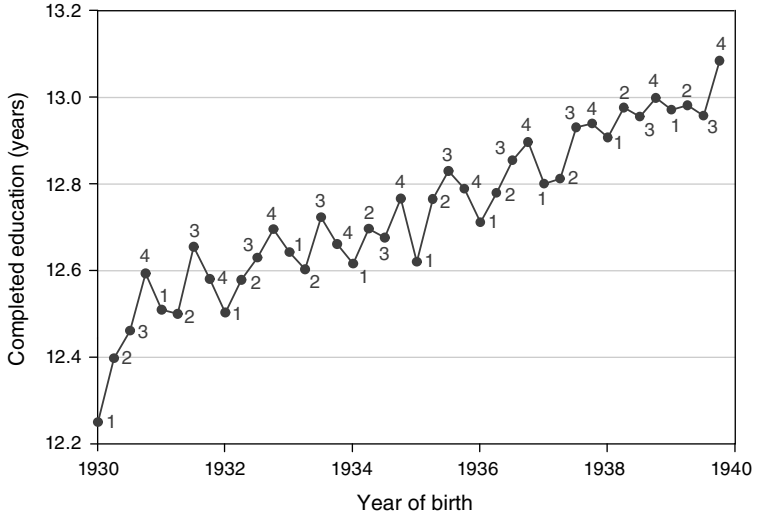


FIGURE 4.14
Years of education and quarter of birth. 1980 Census. Note: Quarter of birth is listed below each observation.

Source: Angrist and Krueger (1991, figure 1).

where Q_{ijc} is a dummy variable equal to one if person i was born inthe jth quarter of the year and equal to zero otherwise and ε_{icj} is an error term with zero average. Since $j = 1, 2, 3$, coefficient β_j measures the impact of a person's quarter of birth on years of education relative to a person born in the fourth quarter. The results are presented in

TABLE 4.2
The effect of quarter of birth on years of education (standard errors in parentheses).

	Quarter of birth effect		
	1	2	3
Total years of education	−.124 (.017)	−.086 (.017)	−.015 (.016)
Years of education for high school graduates	−.004 (.014)	.051 (.014)	.012 (.014)

Source: Angrist and Krueger (1991, table 1).

table 4.2. The first line, column 1, shows that the average number of years of education is about one tenth of a year lower for men born in the first quarter of the year than for men born in the last quarter of the year. The effect is highly significant. Its size is close to that depicted in figure 4.12. The size of the effect is lower for quarter 2 but still significantly different from zero and becomes not statistically different from zero for the third quarter. The second line of table 4.2 indicates that there is no more relationship between years of education and quarter of birth for high school graduates who have at least 12 years of education. All in all, these results support the hypothesis that the effect of the quarter of birth on the duration of schooling is due to compulsory schooling laws. In their paper, Angrist and Krueger (1991) provide further evidence, which readers may peruse, on the effects of compulsory schooling laws.

Wald Estimate, OLS Estimate, and 2SLS Estimate

Following this first stage, in which they bring out the pertinence of the hypothesis that there exists a correlation between birth quarter and duration of study (first stage assumption), Angrist and Krueger set about evaluating the returns to education. The variable explained is the (logarithm of) weekly wages in 1980. Since men born at the beginning of the year are older than those born at the end of the year and will have higher earnings if they are on the upward-sloping portion of the age-earning profile, Angrist and Krueger focus on 40- to 49-year-old men, born between 1930 and 1939, whose wages are hardly related to age, because the age-earnings profile is flat for their cohort.

To begin, a simple way to proceed is to compute the returns to education as the ratio of the difference in earnings by quarter of birth to the difference in years of education by quarter of birth. This is the Wald estimator of the returns to education. Angrist and Krueger present estimates that compare earnings and education between men born in the first quarter of the year and men born in the last three quarters of the year. Formally, the Wald estimator of the returns to education is:

$$\hat{\rho}_{\text{Wald}} = \frac{\overline{w}_{2,3,4} - \overline{w}_1}{S_{2,3,4} - S_1}$$

where $\overline{w}_j$ and S_j denote respectively the average logarithm of wages and the average years of education of individuals born in quarter j; hence $\overline{w}_{2,3,4}$ is the average for the three last quarters. In this case, the Wald estimator is equivalent to an instrumental variables estimator where a dummy variable indicating whether an individual is born in the first quarter of the year is used as an instrument for education and there are no covariates.

TABLE 4.3
Wald and OLS estimates of the returns to education (standard errors in parentheses).

	(1) Born in first quarter	(2) Born in 2nd, 3rd, or 4th quarter	Difference (2)−(1)
ln(weekly wage)	5.8916	5.9027	.01110 (.00274)
Education	12.6881	12.7869	.1088 (.0132)
Wald estimator			.1020 (.0239)
OLS estimator			.0709 (.0003)

Source: Angrist and Krueger (1991, table 3).

Table 4.3 compares the Wald estimate and the OLS estimate of the returns to education. The OLS estimate is obtained from a regression of log weekly earnings on years of education without any other control variable. Table 4.3 indicates that the Wald estimate is higher than the OLS estimate. But the difference between the Wald and OLS estimates is not statistically significant at the 5% level of confidence. This result suggests that the OLS estimate is little biased. However, the analysis needs to be deepened to the extent that the Wald estimate does not allow us to control for age-related trends in earnings.

To account for age-related trends in earnings, Angrist and Krueger (1991) estimate the following two-stage least squares model (2SLS):

$$S_i = \mathbf{X}_i \boldsymbol{\pi} + \sum_c \delta_c Y_{ic} + \sum_c \sum_j \theta_{jc} Y_{ic} Q_{ij} + \eta_i \tag{4.16}$$

$$\mathrm{W}_i = \mathbf{X}_i \boldsymbol{\beta} + \sum_c \zeta_c Y_{ic} + \rho S_i + \varepsilon_i \tag{4.17}$$

where S_i stands for the years of education of individual i, $\mathbf{X}_i$ is a vector of covariates including age, age-squared,[6] race, whether individuals live in city centers, marital status, and region of residence dummies; Q_{ij} is a dummy variable equal to 1 when the individual was born in quarter j ($j = 1$ or $j = 2$ or $j = 3$) and to zero otherwise; Y_{ic} is a dummy variable equal to 1 if individual i was born in year c and zero otherwise; W_i is the logarithm of the weekly wage. The coefficient ρ is the returns to education. The results are presented in table 4.4. The two first columns present estimates of the returns to education when there are no controls except for the year of birth dummies, which are included in all regressions of this table. The OLS and 2SLS estimates of the returns to education are very close. It is worth remarking that the 2SLS estimate computed in table 4.4 differs from the Wald estimate computed in table 4.3 because the 2SLS identifies the returns to education with the variation in education across each quarter of birth in each year, whereas the Wald estimate is identified by the difference in years of

[6] Angrist and Krueger (1991) include age instead of experience because experience depends on the duration of schooling.

TABLE 4.4
OLS and 2SLS estimates of the returns to education for men born 1930–1939: Census 1980. Year of birth dummies are included in all regressions (standard errors in parentheses).

	(1) OLS	(2) 2SLS	(3) OLS	(4) 2SLS
Years of education	.0711 (.0003)	.0891 (.0161)	.0632 (.0003)	.0600 (.0299)
Race (1=black)	_	_	−.2575 (.0040)	−.2676 (.0458)
SMSA (1=city center)	_	_	.1763 (.0029)	.1797 (.0305)
Married (1=married)	_	_	.2479 (.0032)	.2486 (.0073)
8 region of residence dummies	_	_	yes	yes
Age	_	_	−.0760 (.0604)	−.0741 (.0626)
Age-squared	_	_	.0008 (.0007)	.0007 (.0007)
Sargan overidentification test: *p*-value	_	.6553		.8798
Number of observations	329,509	329,509	329,509	329,509

Source: Angrist and Krueger (1991) data set.

education between the first quarter and the rest of the year. The 2SLS model of Angrist and Krueger has three instruments: being born in the first quarter, second quarter, or third quarter relative to being born in the fourth quarter. The 2SLS estimator is a weighted average of the underlying Wald estimators for the three instruments. Columns 3 and 4 of table 4.4 show that the 2SLS and the OLS estimates of the returns to education are not statistically different when control variables, such as race, a dummy for residence in city centers, region of residence dummies, marital status dummy, and experience are included. The coefficient associated with age is small and negative because the estimates focus on 40- to 49-year-old men who are in the later part of their career. The second-to-last line of table 4.4 provides the overidentification test, which tells us whether the assumption that the vector of instruments is independent from the error terms ε of the causal relation is plausible (see Angrist and Pischke, 2008, for further details). Column 4 indicates that the probability to be wrong if one rejects the null hypothesis that the vector of instruments is independent from the error term ε is close to 88%. Overall, table 4.4 indicates that one year of education has a rate of return of about 6%.

4.2.4 THE LIMITATIONS OF THE INSTRUMENTAL VARIABLES METHOD

To this point, we have assumed that the returns to education, represented by the parameter ρ, are identical for all individuals. This is the hypothesis of the homogeneity of the treatment effect. In actuality, it is likely that a supplementary spell of schooling does not yield the same return to all individuals. In particular, persons who modify their duration of study in response to variations in the instrumental variable, here the birth quarter, likely possess distinctive capacities and may therefore have a return to education different from that of the population as a whole. Let us examine the consequences of this phenomenon.

Compliers, Never Takers, and Always Takers

A priori, persons for whom the returns to education are very high and for whom the pursuit of education bears little cost should have longer durations of study. So their duration of study is probably not affected by their birth quarter. Conversely, persons for whom the returns to education are very weak and for whom the pursuit of education bears a high cost choose to leave school as soon as they can and have no incentive to continue their education past the 10th year. Only persons whose interest in pursuing their studies is neither too weak nor too strong may have their duration of study altered beyond the 10th year by a late-quarter birth. This suggests that evaluating the returns to education by choosing the birth quarter as the instrumental variable amounts to evaluating the returns to education for a very special population. To bring this line of reasoning into sharper focus, let us suppose that the return to education is specific to every individual i. In that setting, the Mincer equation (4.12) becomes:

$$\mathrm{W}_i = a + \rho_i s_i + \varepsilon_i \tag{4.18}$$

where ρ_i designates the return to education for individual i and $\mathrm{W}_i = \ln w_i$.

To understand the consequences of heterogeneity in the returns to education, it is helpful to represent the choice of how long to remain in school on the basis of a model of potential outcomes comprising two lengths of study, for example 10 years and 12 years, which we will denote respectively $s_i = 0$ and $s_i = 1$, and two values of the instrument, $z_i = 0$ if the date of birth lies in the first three quarters of the year, and $z_i = 1$ if not. This allows us to pursue our analysis within a setting in which we can distinguish between a "treated" group ($s_i = 1$), who stay in school for a duration of 12 years, and an "untreated" group ($s_i = 0$), who leave school earlier. Let us also suppose that the propensity to pursue one's studies does depend on date of birth, and let us represent this propensity by the linear relation $\gamma + \beta z_i$. Let us denote η_i the cost, including the psychological cost, of pursuing one's studies until the 12th year. This last is a random variable linked to unobservable characteristics of individual i and so potentially correlated to ρ_i. By an adequate normalization, we may always assume that η_i has a zero mean. For each individual i, the decision to pursue her studies is then defined by the equation:

$$s_i = \begin{cases} 1 & \text{if } \gamma + \beta z_i \geq \eta_i \\ 0 & \text{otherwise} \end{cases} \tag{4.19}$$

Equation (4.19) translates into formal terms the hypothesis that the decision to prolong one's education until the 12th year does depend on one's birth quarter. Thus, on average, if the date of birth of individual i falls in the last quarter rather than in the first three, her probability of completing 12 years of study is increased by $\mathbb{E}(s_i|z_i = 1) - \mathbb{E}(s_i|z_i = 0) = \beta$. This equation also shows that persons born after the month of September decide to pursue their studies until the end of the 12th year, corresponding to the end of high school, only if the cost of doing so is sufficiently low (that is, $\eta_i \leq \gamma + \beta$). Persons for whom this cost is too high terminate their studies when they reach the legal age, in other words, before completing their 11th year. The individuals whose behavior is modified by the instrument are those who prolong their education until the 12th year if they were born after the month of September (that is, $\eta_i \leq \gamma + \beta$), and who would have halted it before the 12th year if they had been born before the month of September

(that is, $\eta_i > \gamma$). These persons, known as *compliers*, have a value of η_i lying between γ and $\gamma + \beta$. Other persons choose a duration of study less than 12 years whatever their birth quarter (that is, $\eta_i > \gamma + \beta$) and are called *never takers*. Finally the *always takers* decide to continue their studies until the 12th year (that is, $\eta_i \leq \gamma$) whatever their birth quarter.

LATE and ATE Estimates

The upshot is that, in this context, the most verisimilar empirically, in which the returns to education are heterogeneous, the method of instrumental variables may make it possible to estimate the average impact of an increase in the duration of study *only* for the compliers, whose duration of study is actually altered by the instrument. It does not measure the returns to education for the always takers who achieve 12 years of study independently of their birth quarter, or for the never takers who quit before the 12th year whatever their birth quarter.

For the instrumental variables method to supply an estimate of the average impact of an increase in the duration of study of the compliers, it is necessary that (1) the instrument z be correlated with the duration of study (the "first stage condition": $Cov(z, s) \neq 0$); (2) that the instrument be orthogonal to the error term ε (the "exclusion restriction condition": $Cov(z, \varepsilon) = 0$), as we saw when the returns to education were assumed to be identical for all. Imbens and Angrist (1994) have shown that it is also necessary that two supplementary conditions are verified. First, (3) the instrument must be orthogonal to the return to education (the "independence condition": $Cov(z, \rho) = 0$). If this is not the case, the instrument yields a selection that no longer permits us to evaluate the causal effect of duration of study. Finally, (4) the instrument must affect all the persons whose behavior it modifies in the same way. So the fact of being born in the first quarter rather than in the last three may reduce the duration of study for certain persons, or have no impact, but must not increase it for anyone. This is the hypothesis of monotonicity, implicitly assumed when parameter β in equation (4.19) is identical, and so of the same sign, for everyone. If the monotonicity hypothesis is no longer verified, the instrument does not permit us to evaluate the average impact of an increase in the duration of study, since it increases it for some people and reduces it for others. If all four of these conditions are verified, the instrumental variables method does make it possible to obtain the average effect of the treatment on the compliers. We thus get the LATE (Local Average Treatment Effect) estimator of the returns to education for the compliers:

$$\rho_{\text{LATE}} = \mathbb{E}\left(\rho_i | \gamma < \eta_i \leq \gamma + \beta\right) = \frac{\mathbb{E}(\mathrm{W}_i | z_i = 1) - \mathbb{E}(\mathrm{W}_i | z_i = 0)}{\mathbb{E}(s_i | z_i = 1) - \mathbb{E}(s_i | z_i = 0)} \tag{4.20}$$

This result is known as the LATE theorem, indicating that an instrument that yields a random allocation of persons between the treatment group and the untreated group, which is independent of the returns to education and which affects the assignment to treatment only in one direction, can be used to estimate the average causal effect on the compliers. When the potential returns to education are heterogeneous and all individuals are not compliers, this effect differs from the average effect of treatment on the whole population, usually denoted ATE (Average Treatment Effect), equal to:

$$\rho_{\text{ATE}} = \mathbb{E}\left(\rho_i\right)$$

The local average treatment effect also differs from the average effect of treatment on the whole group treated, which comprises all those who completed 12 years of study (that is, the compliers and the always takers). The estimator of the average effect of treatment on the treated group, usually denoted ATT (Average Effect of the Treatment on the Treated) is given by:

$$\rho_{\text{ATT}} = \mathbb{E}\left(\rho_i | \eta_i \leq \gamma + \beta z_i\right) \tag{4.21}$$

In our example, the local average treatment effect is different from the effect of treatment on the treated when the decision to pursue further studies is influenced by the returns to education or, formally, when η_i is correlated to ρ_i, which is the most probable situation according to the predictions of the theoretical model of human capital accumulation. We may remark nonetheless that the local average treatment effect and the average effect of treatment on the treated are identical if there are no always takers. In that case, there is no individual with $\eta_i < \gamma$, and we observe that $\rho_{\text{ATT}} = \rho_{\text{LATE}}$. This situation is not, however, a likely one if the instrument is the birth quarter. Finally, in the highly unlikely situation where there is no heterogeneity of the treatment effect, we have: $\rho_{\text{LATE}} = \rho_{\text{ATE}} = \rho_{\text{ATT}} = \rho_{\text{Wald}}$.

Thus, taking into account the possibility of heterogeneous returns to education shows that the choice of instrument may affect the estimation of the returns to education. This problem may be even more thorny to the extent that the compliers represent a small share of the population. This share is simple to calculate. Effectively, the group of compliers comprises persons who stay in school until their 12th year if $z = 1$, so their share is equal to $\mathbb{E}(s_i | z_i = 1)$, from which must be subtracted those who stay in school until the 12th year if $z = 0$, whose share is equal to $\mathbb{E}(s_i | z_i = 0)$. The share of the population who are compliers is thus simply equal to the denominator of equation (4.20). This share represents 0.46% of the population of persons who completed 10 or 12 years of schooling. So an evaluation of the returns to education that takes the birth quarter as an instrumental variable is based on the behavior of a minuscule portion of the population that is highly unlikely to represent the whole.

Oreopoulos (2006) and Carneiro et al. (2011) do in fact demonstrate that because of such selection problems, the estimation of the returns to education by the instrumental variables method may be illegitimately dependent on the instrument chosen. So in assessing the validity of an estimation made with the instrumental variables method, it is essential to take into account the characteristics of the population affected by the instrument. In the case of birth quarter, this population is made up of persons whose returns to education are probably weak and whose duration of study is probably short. Hence the results obtained by taking birth quarter as the instrumental variable will certainly shed no more than a feeble light on the impact of policies that might affect the duration of study of other populations, for example, policies that offer financial support to those pursuing advanced studies.

4.2.5 Lessons from Studies of Siblings and Twins

Another method used to evaluate the returns to education consists of using data about individuals whose abilities are as alike as possible. From this perspective, several contributions estimate the returns to education for siblings, and some studies have even used

populations made up of homozygotic twins (Ashenfelter and Rouse, 1998; Oreopoulos and Salvanes, 2011). To grasp how this method works, let us assume we have available a sample population of homozygotic twins. The returns to education are estimated on the basis of a Mincer equation (simplified here for ease of exposition):

$$\ln w_{ij} = a + \rho s_{ij} + A_j + \varepsilon_{ij}, i = 1, 2$$

where w_{ij} represents the wage of twin i of family j, s_{ij} measures the duration of study of twin i of family j, A_j is a family fixed effect that represents the ensemble of unobserved factors that may affect the capacities of the two twins in family j, and ε_{ij} is a random term of null average proper to twin i of family j. It is possible to estimate the returns to education by eliminating the factor of the unobserved capacities A_j of each family on the basis of the wage differences within each pair of twins. We then obtain the equation:

$$\ln w_{1j}/w_{2j} = \rho(s_{1j} - s_{2j}) + (\varepsilon_{1j} - \varepsilon_{2j}) \quad \textbf{(4.22)}$$

We observe that this makes it possible to estimate without bias the returns to education using the method of ordinary least squares, on condition that the differences in duration of study between twin members of the same family are not correlated to differences in aptitude that may influence their gains (so formally the hypothesis is $Cov(s_{1j} - s_{2j}, \varepsilon_{1j} - \varepsilon_{2j}) = 0$). But there is a major objection to adopting this procedure: perfectly identical twins ought, by definition, to study for the same length of time. For their lengths of study to vary, there must be events that differentiate these twins in some manner. For the estimation by OLS of equation (4.22) to yield an unbiased estimate of the returns to education, events are required that affect their taste for study, or the cost of staying in school, but not the potential gains w_{ij}. In other words, we must assume that the reasons the twins do not study for the same length of time are not correlated to their future earnings. This hypothesis is open to grave doubt (Borjas, 2010, p. 251), and so it is unlikely that the utilization of data bearing on twins will lead to sound estimations of the causal impact of education on earnings.

With these precautions taken fully into account, we note that Ashenfelter and Rouse (1998) find that the differences in the returns to education between genetically identical individuals are slightly weaker (on the order of 10%) than those obtained by comparing the duration of schooling and incomes of any two individuals at random. If we accept the premise that homozygotic twins have identical abilities and that differences in their length of schooling are due to random events that do not change their abilities, these results show that ability and selection biases have little weight. Estimations carried out on the whole of the population would only overestimate the returns to education very slightly.

Oreopoulos and Salvanes (2011) have used Norwegian administrative records that supply information on the educational and professional trajectories of all persons born in that country since 1920. On the basis of these registers, they constructed a very broad sample made up exclusively of siblings and twins, enabling them to examine the correlations between duration of study and future earnings. All other things being equal, they find that siblings (or twins respectively) with one more year of schooling have, on average, 5.2% (or 4.8% respectively) more annual income than their less educated siblings (or twins). These results confirm the results of Ashenfelter and Rouse (1998).

5 THE RETURNS TO EDUCATION

Numerous empirical contributions have been dedicated to the estimation of the private pecuniary returns to education, following the lead of Mincer. Mincer's method has the virtue of simplicity: it uses a simple relation between wage logarithms and duration of study to estimate the rate of internal returns to education. We will now see that subsequent work has loosened certain restrictive hypotheses of the Mincer model in order to sharpen the estimation of this rate of return. Also, more recently, empirical work has tried to go beyond the assessment of private pecuniary returns to education, in an attempt to estimate nonpecuniary returns and the returns of education to society as a whole.

5.1 Private Returns to Education

The Mincer model assumes that every year of schooling has the same return. This hypothesis is open to grave doubt, and it is possible to relax it and estimate models in which every year of schooling may have a different return. It then becomes apparent that estimations grounded on the canonical Mincer model are seriously biased. In this section, we set aside the problems of selection bias previously studied in order to concentrate on the specification of the model estimated. As we will see, the specification of the model can have considerable influence on the estimation of the returns to education.

5.1.1 Private Pecuniary Returns to Education: Beyond the Mincer Model

The Mincer model rests on a set of very restrictive hypotheses (see section 4.1), which leads to estimating the returns to education on the basis of a regression of log earnings on years of schooling, to which seniority is sometimes added.

It will be helpful to briefly recall the hypotheses necessary to derive the Mincer equation, so as to gauge their extent. (1) The rate of return to an added year of schooling is independent of duration of study; (2) the cost of an added year of schooling is proportional to the wage (which allows us to obtain the term $\dot{w}(s)/w(s)$ in equation (4.10)); (3) career duration is sufficiently long (which corresponds to the hypothesis $T \to \infty$ in equation (4.10); (4) career duration is independent of duration of schooling. Additionally, when experience is taken into account, the Mincer gains equation rests on the hypothesis that earnings functions are multiplicatively separable in experience and schooling [equation (4.13)].

These hypotheses are quite likely not valid. The costs of education are not limited to earnings forgone during time spent studying. There are direct costs as well, like tuition fees, and there may also exist psychological costs (see Heckman et al., 2006).

It is possible to calculate returns to education within a less restrictive framework of hypotheses than that chosen by Mincer. Hence Heckman et al. (2008) assume that the age of retirement, denoted $T(s)$, may depend on duration of schooling. They also take into account the existence of tuition fees. A person who studied for duration s and gained experience for duration x earns a wage denoted $w(s, x)$. To calculate the internal rate of return in this framework, it is enough to point out that the internal rate of return of a supplementary duration Δ of education when the duration attained is s, is the rate

of actualization that equalizes the net earnings of durations of schooling s and $s + \Delta$. This is the approach adopted by Becker (1964, chapter 3) in his classic work on human capital. Let us denote $c(t)$ the instantaneous direct cost of schooling. The internal rate of return ρ must then verify the equality:

$$\int_s^{T(s)} w(s, \tau - s) e^{-\rho(\tau - s)} \mathrm{d}\tau - \int_0^s c(\tau) e^{-\rho\tau} \mathrm{d}\tau = \int_{s+\Delta}^{T(s+\Delta)} w(s + \Delta, \tau - s) e^{-\rho(\tau - s - \Delta)} \mathrm{d}\tau - \int_0^{s+\Delta} c(\tau) e^{-\rho\tau} \mathrm{d}\tau \quad \textbf{(4.23)}$$

This equation makes it possible to estimate a marginal internal rate of return to education that varies a priori for each pair (s, Δ). Heckman et al. (2008) have done so on the basis of American census data. In their benchmark case, they assume that workers spend 47 years working irrespective of their educational choice (i.e., a high school graduate works until age 65 and a college graduate until 69). Then, they consider more complex situations where they account for tuition fees and taxes. We confine ourselves here to presenting the benchmark case. To estimate the internal marginal rate of return, they proceed in two steps.

First, they estimate a log earning equation to compute the change in wage profile associated with increases in schooling duration. More precisely, for individuals with two different durations of schooling. they estimate the equation:

$$\ln w = \alpha + \beta s + \delta x - \gamma x^2 + \varepsilon$$

For instance, we can substitute for s a dummy variable equal to 1 for individuals with 8 years of schooling and equal to zero for those with 6 years of schooling. The estimation of this equation with OLS (neglecting selection issues) allows us to compute the expected log earnings for individuals with 6 years of schooling and 8 years of schooling.

Then, the estimated internal return to education is the value of ρ that satisfies equation (4.23) with the predicted earning profiles for the first 47 years of experience.

Table 4.5 displays the rates of return for white men using the public census data for 1980, assuming that the direct cost of education equals zero. In the first row of table 4.5, returns are computed when the log earning is estimated with Mincer's specification, as in the second row of table 4.1. Row 2 of table 4.5 relaxes the assumption of linearity in schooling by including dummy variables for the categories of schooling shown in the table. This modification leads to substantial differences in the estimated rate of return to schooling, showing that the marginal return to education is not constant. The differences are especially important for year 12 (high school) and 16 (college), corresponding to schooling levels associated with degree completion years, which show much larger returns than other schooling durations. These results show that imposing linearity in schooling leads to upward biased estimates of the rate of return to completed years that do not produce a degree, while it leads to downward biased estimates of the degree completion years.

Row 3 relaxes both linearity in schooling and the quadratic specification for experience. To calculate the earnings profiles with a nonquadratic specification for experience, Heckman et al. (2008) use a nonparametric method. Let $\mathrm{w}_i = \ln[w(s_i, x_i)]$ denote

TABLE 4.5
Internal rates of return for white men: Earnings function assumptions.

	Schooling comparisons				
	6–8	8–10	10–12	12–14	14–16
Mincer specification	11	11	11	11	11
Relax linearity in s	3	−11	36	5	18
Relax linearity in s & quad in x	4	−4	28	6	16
Relax linearity in s & quad in x	16	66	45	5	21

Source: Heckman et al. (2008, table 3a).

log earnings, which depend on the schooling level s_i and experience x_i of individual i. They estimate the relation between earnings and experience with a local linear estimator for the conditional expectation $\mathbb{E}[\mathrm{w}_i|s_i = s, x_i = x]$, which is computed from the minimization of an error term, for each s and each year of experience x, between log earning w and what experience can yield:

$$\min_{a,b} \sum_{i=1}^{n} [\mathrm{w}_i - a - b(x_i - x)]^2 \frac{1}{nh} K\left(\frac{x_i - x}{h}\right) \mathbf{1}(s_i = s) \tag{4.24}$$

where

$$K(y) = \begin{cases} (15/16)(y^2 - 1)^2 & \text{if } |y| < 1 \\ 0 & \text{otherwise} \end{cases}$$

is a kernel function and h is a smoothing parameter called the bandwidth, which is set to 5. The indicator function $\mathbf{1}(s_i = s)$ denotes that only observations with schooling equal to s are used in the computation. The estimator of parameter a is equal to the estimator of the conditional mean $\mathbb{E}[\mathrm{w}_i|s_i = s, x_i = x]$. This procedure provides nonparametric estimates of the earnings–experience relationship separately for each education level. Let us denote by $\widehat{\mathrm{w}}_i(s, x)$ the estimator of $\mathbb{E}[\mathrm{w}_i|s_i = s, x_i = x]$. These predicted earning profiles are used to compute the returns to education in equation (4.23). Row 3 displays the results of this procedure when it is assumed that there is a common experience profile for all levels of schooling. In that case, one computes the estimator of $\mathbb{E}[\mathrm{w}_i|, x_i = x]$ for all levels of schooling without using the indicator function in equation (4.24). The assumption of quadratic specification for experience is relaxed, but not the assumption that earnings functions are multiplicatively separable in experience and schooling, as assumed by Mincer [equation (4.13)]. The results of row 3, which are not very different from those of row 2, lead Heckman et al. to conclude that the assumption that earnings are quadratic in experience is empirically innocuous for estimating returns to schooling once linearity in years of schooling is relaxed. In row 4, where earnings experience profiles are computed separately for each duration of schooling, using the indicator function in equation (4.24), the assumption that earnings functions are multiplicatively separable in experience and schooling is relaxed. This specification leads to quite different results, suggesting that this assumption is not innocuous.

All in all, these results indicate that the functional form assumptions implied by Mincer-based estimations of the rates of return lead to very large biases. Table 4.5 shows especially that the estimation of the returns to education within a framework of hypotheses less restrictive than those imposed by Mincer leads to markedly higher returns for the years of education that fulfill the requirements for a diploma. This difference is particularly striking for the completion of high school. It is less so for the completion of a college degree. As well, the estimated return to a college degree diminishes significantly when taxes and tuition fees are taken into account. In that case, the results obtained by Heckman et al. indicate that taxes and tuition fees substantially reduce the returns to education for college graduates but have lower impact on the returns to education for individuals with low levels of schooling. This result is consistent with progressive taxation and tuition fees increasing with the level of education.

The size of these estimated returns to education at the end of the 12th year raises important questions, to the extent that more than 15% of new cohorts of American youth do not attain a high school degree, despite its high estimated return. The estimation of an average return to education likely conceals very strong heterogeneity, the exact nature of which has yet to be specified. This heterogeneity may have to do with aspects as diverse as earnings, direct costs (including ones incurred in the psychological dimension), the cost of access to credit, risk aversion, preference for the present, problems of temporal incoherence, or indeed the capacity to ancipate future gains. It is necessary to better understand the nature of this heterogeneity in order to work out suitable public policies (see the discussion in Heckman et al., 2008).

5.1.2 Private Pecuniary Returns to Education: Some Order of Magnitude

While remaining aware that evaluations of returns to schooling have their limitations, it is worthwhile to give orders of magnitude for some OECD countries. Figures 4.15 and 4.16 show estimates of the private internal rates of return to education in 2005–2008 in 21 OECD countries (OECD, 2012). The approach is similar to that of equation (4.23). The internal rates of return to education are computed as the net benefits from getting further education (e.g., going from no degree to an upper secondary degree, or from an upper secondary to a tertiary degree). These figures give an idea of the incentives to get educated at different levels. The direct costs of education are based on the private expenditure per year and the length of education in each country. Implicitly, in equation (4.23) the indirect (opportunity) costs of education are also taken into account by the portion of forgone wages, which would have been earned at the lower level of education, for the duration of the higher education program. The benefits of education are based on differences in earning between those who have a given degree and those who do not have such a degree but do have the degree just below. They are calculated each year as the average of earnings in constant dollars (at purchasing parity power) for each age–degree group, so as to obtain age–earnings profiles by country. Hence, the impact of age and diploma on earnings is not estimated using a Mincer equation as in Heckman et al. (2008). The approach here is closer to the financial analysis of an investment. Average earnings are adjusted to take into account taxes and social transfers, as well as the risk of unemployment for each level of education. However, grants and subsidized loans sometimes received early during education are ignored in the calculations, as well as pension differences after retirement.

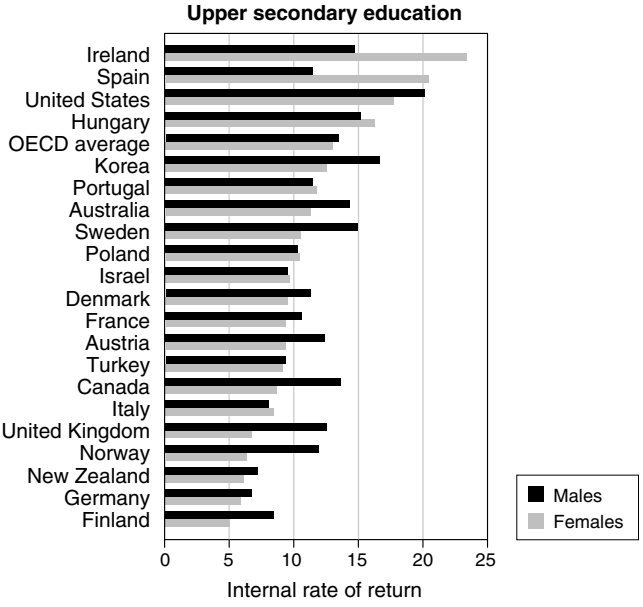


FIGURE 4.15
Internal rates of return to education from no degree to an upper secondary degree for 21 OECD countries, 2008 or latest year available. The OECD average is the nonweighted average of percentages among 25 OECD countries for which estimates are available, including those not represented in this figure. Data not available for non-OECD countries.

Source: OECD (2012, tables A9.1 and A9.3, pp. 174 and 178).

On average, in the OECD the rate of return to upper secondary education is quite substantial at 13%, a little bit higher than that to tertiary education (12%). The situation of the labor market can of course influence incentives: as unemployment increases, the opportunity cost of getting further education decreases, while the relative gains from tertiary education increase (since the more highly educated typically experience less increase in unemployment than the less educated). More structurally, the differences across countries can stem from a discrepancy between demand and supply of skilled labor. For instance in countries where there is a strong demand for highly skilled labor (due to catch-up, e.g. Poland and Hungary) but still a large fraction of the population that lacks tertiary education, the latter can yield substantial returns. Of course this situation will tend to reverse as the share of the population with skills progressively increases. Another factor is the overall wage dispersion: a compressed wage structure, typical of the Nordic countries for instance, will typically generate lower returns to higher education. This is one of the reasons that higher education is basically free of charge in these countries, with many subsidies and grants for students. Conversely, in countries with substantially larger overall earnings inequality, such as the United States, rates of return will tend to be higher.

Against this backdrop, certain countries have faced increasing wage inequality over the last four decades of the 20th century (a problem to which we will return in chapter 10). This increase in wage inequality, which is particularly sharp in the United States, goes in tandem with a rise in the returns to education, a phenomenon illustrated in figure 4.17. It shows that the returns to education for men fluctuated widely during the

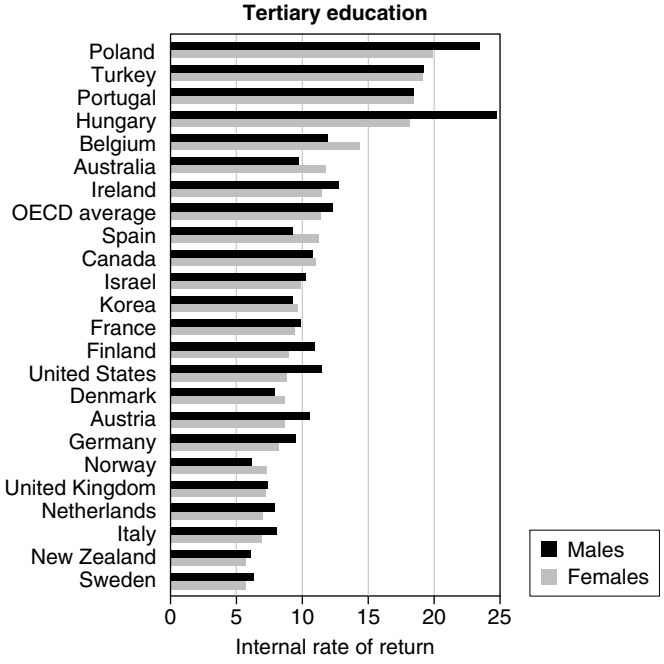


Figure 4.16
Internal rates of return to education from an upper secondary degree to a tertiary degree for 21 OECD countries, 2008 or latest year available. The OECD average is the nonweighted average of percentages among 29 OECD countries for which estimates are available, including those not represented in this figure. Data not available for non-OECD countries.

Source: OECD (2012, tables A9.1 and A9.3, pp. 174 and 178).

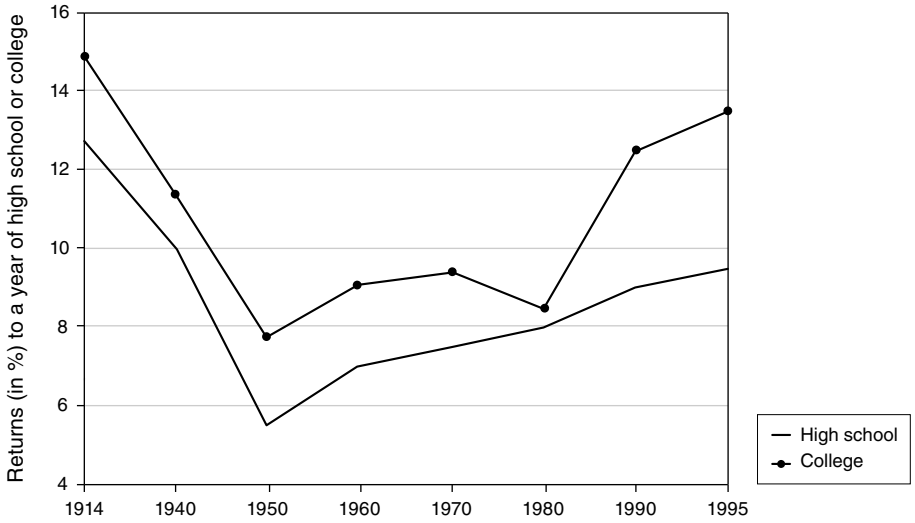


Figure 4.17
Returns to a year of schooling for men, 18–34 years old in the United States from 1914 to 1995.

Source: Goldin and Katz (2001, table 4, figures 1 and 6).

20th century in the United States, decreasing between 1914 and 1940 and then increasing between 1950 and 1995. Goldin (2001) and Goldin and Katz (2001) maintain that the phase of decrease resulted from a considerable expansion of secondary schooling at the end of the 1910s and in the 1940s. The slowdown in the expansion of schooling in the United States after the World War II helps, from this point of view, to explain the increase in the returns to education between 1950 and 1995.[7]

5.2 Private Nonpecuniary Returns to Education

The private gain from years of study does not boil down to the chance of a better wage. Schooling actually exerts effects in a range of dimensions. It may promote better decision making in the areas of health, choice of partner, and the schooling of one's children. It may also modify styles of consumption and one's interest in work. Oreopoulos and Salvanes (2011) stress that schooling may affect preferences in a way that makes individuals more patient, more goal-oriented, and less likely to engage in risky behavior. Many studies also detect a positive relation between duration of schooling and health status. A longer duration of schooling is also associated with increased satisfaction in life and decreased probability of experiencing divorce or parenthood in adolescence (Grossman, 2006). Yet the observation that such correlations exist does not warrant the conclusion that there is a causal effect of education on nonpecuniary gains. Basing themselves on methods used to evaluate the impact of education on pecuniary gains, Oreopoulos and Salvanes (2011) attempt to identify such a causal effect by comparing twins and siblings and then analyzing the consequences of increases in the period during which school attendance is compulsory. Their results, obtained on Norwegian and North American data, confirm the impact of education on nonpecuniary gains.

5.3 Social Returns to Education

Estimates of the private returns to education no doubt fail to render a full account of the benefits that flow to society as a whole from investments in schooling. It is possible that education exerts positive externalities and that the social returns to education are superior to the private ones. Nevertheless, it is also possible that the social returns to education are inferior to the private ones, if the role of the educational system does essentially consist of selecting individuals as a function of personal characteristics which employers cannot observe, as hypothesized in section 3.

Empirical studies dedicated to the social returns to education tend to show that education does indeed exert positive externalities and that the social returns to education are superior to the private ones. In essence, these positive externalities result from the fact that education improves the capacities of individuals for socialization. Education teaches how to communicate with and understand, how to learn from and instruct, others. Hence education reduces criminality, improves involvement in civic activities like voting and participation in associations, and promotes the diffusion of knowledge.

5.3.1 Social Engagement

Duration of study and various forms of civic involvement are systematically correlated, so much so that Helliwell and Putnam (2007, p. 1) assert that "education is one of

[7] These problems are brought into sharper focus in chapter 10, on technical progress and inequality.

the most important predictors—usually, in fact, the most important predictor—of many forms of social engagement, from voting to chairing a local committee to hosting a dinner party to trusting others." The available research suggests that these correlations do reveal, at least in part, a causal impact of schooling. For example, Milligan et al. (2004) exploit variations in the compulsory legal minimum of schooling in the United States and the United Kingdom to detect such a causal relation. They find that education increases the information voters can access about candidates and political parties and it increases their turnout at elections, their participation in political meetings, and their voluntary engagement in civic activities. Oreopoulos and Salvanes (2011) confirm these results for data bearing on the United States and Norway, with an identification strategy that relies on variations in the compulsory duration of schooling and a comparison of the behavior of twins and of sets of male and female siblings. Relying on this empirical evidence, Glaeser et al. (2007) argue that the positive cross-country relation between democracy and education comes from the fact that education increases the societywide support for democracy because democracy relies on people with high participation benefits for its support. They find that better-educated nations are more likely both to preserve democracy and to protect it from coups.

5.3.2 Criminality and Violence

By promoting better socialization, by augmenting potential earnings, and by giving focus and structure to the activity of young people who stay in school, education may play a role in reducing criminality. Empirical research finds that education does indeed have a negative impact on criminality. Lochner and Moretti (2004) used data on men in the United States for the period 1960–1980. Using changes in the compulsory attendance laws of states over time to account for the endogeneity of schooling decisions, they estimate that education to high school level reduces criminality. They find that the externality connected to the reduction of criminality represents between 14% and 26% of the private returns to education. Using the same identification strategy for men and women in the United Kingdom over the period 1984 to 2002, Machin, Marie, and Vujić (2011) find a strong and significant negative impact of education especially on property crimes.

5.3.3 Labor Mobility

Geographical labor mobility contributes substantially to the effective functioning of labor markets. A more mobile labor force facilitates the reorganization of the apparatus of production and reduces the socially costly phenomenon of unemployment. Education, by enhancing communicative capacity and adaptability to new environments, can have a positive impact on labor mobility. Machin, Pelkonen, and Salvanes (2011) identify the impact of schooling on labor mobility by using as their instrument an educational reform which increased the years of compulsory schooling in Norway by two years. The timing of the reform was geographically dispersed in a quasi-random fashion. They find that the length of compulsory education has a causal impact on mobility of individuals at the lowest levels of educational attainment: one additional year of education increases the annual mobility rates by 15%. This result leads them to argue that a significant part of the United States–Europe difference, as well as the European North-South difference in labor mobility, is likely due to differences in levels of education in the respective regions.

5.3.4 Spillover on Children

Currie and Moretti (2003) estimate, while controlling for selection biases, that better education of mothers exerts a positive impact on the health of their offspring. An original confirmation of this result was supplied by Chevalier and O'Sullivan (2007). They start with the well-documented observation that low birth weight has negative effects both large and long-term on child development, even in rich countries. Hence they use the 1947 reform of the minimum age for school leaving in the United Kingdom to identify a causal effect of the schooling of mothers on the birth weight of their offspring. They find a modest but significant effect of the educational level of future mothers on the birth weights of their children. The reason is that low birth weight is very largely the result of lifestyle choices (nourishment, hygiene, exposure to stress, consumption of tobacco, alcohol, or other drugs, etc.) made by future mothers before and during pregnancy. Schooling acts on lifestyle choices and so indirectly on birth weight.

5.3.5 Knowledge Externalities

Rauch (1993), Acemoglu and Angrist (2000), and Moretti (2004) have attempted to assess the difference between the private returns to education and the social ones arising from externalities by comparing the impact of education on earnings of individuals situated in environments in which the level of general education differs. Rauch (1993) estimates, through a comparison of the incomes of individuals situated in different cities, that knowledge externalities increase the returns to education by 3 to 5 percentage points. The studies of Acemoglu and Angrist (2000) and Moretti (2004) focus especially on the problem of the endogeneity of educational choices, which may bias estimates. Acemoglu and Angrist exploit the heterogeneity of compulsory attendance laws and child labor laws in U.S. states between 1920 and 1960 to pinpoint exogenous variations in the environment, which may influence educational choices. In this context, they do find positive, though slight, knowledge externalities that improve the returns to education on the order of 1 percentage point and do not differ significantly from zero. Moretti (2004), using a different methodology, also finds positive externalities corresponding to an improvement of the returns to education lying between 0.6 and 1.2 percentage points.

More advanced education can also favor the discovery and adoption of new technologies (Foster and Rosenzweig, 1996), which themselves exert macroeconomic externalities that are a source of growth (Nelson and Phelps, 1966; Aghion and Howitt, 1998). Research on growth places a great deal of emphasis on the central role of education in this domain. Empirical work on international macroeconomic data generally highlights a positive impact of education on growth (see Topel, 1999; Krueger and Lindahl, 2001). Using data from 31 countries for the years 1960–1990, Hanushek and Kimko (2000) tighten the focus, showing that variations in the quality of education as measured by achievement in mathematics and the sciences have a highly significant impact on economic growth. Hanushek and Woessmann (2007) extended this study to cover 50 countries and the 10 extra years 1990–2000. They confirm the earlier results, including for developing countries, and offer an interesting order of magnitude. It is their calculation that on average a reform of the education system that produced a (modest) rise of 0.5 standard deviation point in test score achievement would, after 30 years, raise GDP at that horizon by 4 points. Such a gain is considerable: 4 points of GDP covers the bulk of expenditure on primary and secondary education in a wide range of countries (see figure 4.1).

On the whole, empirical work does suggest that the social returns to education do exceed the private ones. This observation justifies to some extent the preponderant role of the state in expenditure on education—a problem to be investigated more fully in chapter 14.

5.4 What Is Really Important in Education?

In terms of education policy, duration of study is one of the ingredients allowing individuals to accrue human capital. But it is not the only ingredient and is perhaps not even the most important one. In what follows, we review the factors other than duration of schooling that most influence the accumulation of human capital.

5.4.1 Test Scores

Empirical studies generally find that persons who achieve the highest scores on tests measuring knowledge obtain higher earnings in the labor market (Murnane et al., 1995, 2001; Currie and Thomas, 2001). Hence quantity of schooling as incorporated into equations of the Mincer type very imperfectly reflects the skills of an individual. These skills are measured more finely by tests that assess cognitive capacity, such as the IQ (intelligence quotient) test, and general knowledge or by tests that assess noncognitive capacities like perseverance, dependability, and consistency. A number of studies have introduced an indicator for tests of this kind into Mincer equations as a supplementary explanatory variable. Hanushek and Zhang (2009) carry out this exercise for 13 OECD countries. These studies all conclude that substantial explanatory power attaches to cognitive and noncognitive capacities. To give an order of magnitude, a one standard deviation increase in the coefficient of a variable that measures levels of cognitive capacity (by means of tests, for example) increases the present value of life-cycle earnings in a range between 10% and 20% (Hanushek and Rivkin, 2012). These results suggest that it is important to take into account the quality of schooling, not just its duration, in evaluating the returns to education.

5.4.2 Teacher/Pupil Ratio

Some studies find that the teacher/pupil ratio, the expenditure per pupil, and the wages of teachers appear to have a positive impact on income obtained by students when they leave school (Card and Krueger, 1992; Altonji and Dunn, 1996). For example, Card and Krueger (1992), using data for the United States, show that the rate of return to schooling is higher in states where the pupil/teacher ratio is lower. They estimate that bringing the pupil/teacher ratio down by 10 increases the rate of return to education by around 9 percentage points. These results have occasioned much debate. For instance, Hanushek et al. (1996) conclude that these results stem from an aggregation bias due to the fact that Card and Krueger (1992) consider only the average characteristics of schools by state and not the characteristics of each individual's school. Hanushek (2002) reviewed the results of 376 published studies focusing on the impact of expenditure on education on the performance of students, which show that it is difficult to detect a systematic influence of expenditure on education on the performance of students. Studies making use of randomized or natural experiments nevertheless find a positive impact of reductions in class size on returns to education. In particular, the use of data

issuing from the Tennessee Student/Teacher Achievement Ratio, which consisted of randomly assigning primary school students and teachers to classes of varying size, has shed an interesting light. These data cover more than 6,000 pupils followed since 1985. They show that a reduction in class size during primary school raises the level of success on achievement tests and the probability of staying in school longer (Krueger, 1999). These effects were, by the way, more marked in pupils from disadvantaged minorities. The contribution of Angrist and Lavy (1999) confirms this conclusion using a natural experiment grounded in the Maimonides rules of the Talmud, which dictate a maximum class size of 40 pupils. These rules, in force in public schools in Israel, entail that a new class systematically springs up whenever an existing class exceeds its quota of 40. The arrival of a new pupil in a class of 40 pupils thus brings the size of that class down from 40 to 20.5 pupils. Analysis of the consequences of this natural experiment (assuming that the arrival of a new pupil is an entirely random event) shows that reductions in class size improve student performance.

5.4.3 Degrees and School Quality

Knowledge acquisition, when successful, is generally rewarded with a degree capable of influencing the benefits derived from an extra year of study. In France the work of Goux and Maurin shows that years of study not recognized by a degree entail significant variations in remuneration (+3.2% per year) but nevertheless have an impact two to three times weaker than years that are so recognized. Goux and Maurin also estimate that the type of degree significantly influences earnings in France, where, as in Germany, different educational systems coexist and compete. An engineering degree from a "grande école" (an elite post-secondary institution) leads to wages 25% higher than the degree awarded upon completion of the "deuxième cycle" in university (the equivalent of a master's degree), though the periods of study are of comparable length.

In the United States, Jaeger and Page (1996) estimate that the acquisition of a degree has a significant impact on hourly wages. Comparing the performances of individuals who obtained a degree with those of individuals who failed, these authors find that the degree contributes to around one quarter of the return to education of 16 years of study and to more than half of the return to the four years from the 12th (the last year of high school) to the 16th (the last year of undergraduate study in college). But their estimates are based on regressions and, the presence of numerous control variables notwithstanding, it is possible that their results derive from characteristics not observed by the econometrician.

Martorell and Clark (2010) employ a clearly more convincing identification strategy to estimate the signaling value of a high school diploma in the United States. To that end, they exploit the results of high school exit examinations and other tests required to obtain a high school diploma. More precisely, Martorell and Clark estimate the signaling value of a diploma via a regression discontinuity strategy that compares the earnings of workers who narrowly passed these exams (and so earned a high school diploma) with the earnings of students who narrowly failed these exams (and so did not). The central hypothesis of this strategy is that these two groups of students have, on average, the same productive characteristics observed by firms (but not by the econometrician). Hence any income disparity is interpretable as a causal impact of the value of

the diploma. Using data from Florida and Texas—two of the states that use statewide high school exit exams—Martorell and Clark show that the diploma has little effect on earnings.

5.4.4 Teacher Quality

Teacher quality appears to play a central role in pupil performance. If one equates the effectiveness of a teacher with the results her students obtain on tests and exams, which is a standard assumption, one must be aware of a potential selection effect: it is possible that the best teachers choose schools where they are likely to encounter the best pupils, and the best pupils (or rather their parents) do the same, choosing schools with a reputation for high-quality teaching staff. By compiling various studies appearing after 2004 and bearing on the effectiveness of teachers in the United States, Hanushek and Rivkin (2012) are able to approximate a distribution of the effectiveness of teachers that takes into account an eventual selection effect. They then estimate that if he goes from a teacher situated in the bottom quartile to one situated in the top quartile, a pupil from the 50th percentile (the median) in the distribution of results achieved in mathematics will find himself in the 58th percentile the following year. Hanushek and Rivkin note that this improved achievement is greater than that yielded by estimations of a 10-pupil reduction (a hefty one, in other words) in class size.

Knowing the distribution of the effectiveness of teachers, and inferring from a range of studies the gain that accrues to the future earnings of a pupil from an increase in teacher effectiveness, Hanushek and Rivkin are able to calculate to an order of magnitude the added value per pupil of an increase (or a reduction) in the effectiveness of a teacher. To obtain the collective surplus, in other words the added value per teacher for a whole class, it is necessary to multiply that by the class size. The results of these calculations are presented in figure 4.18, which sums up well the present state of knowledge of the effect of teacher quality on earnings over the life cycle.

Note for example that a teacher situated in the 60th percentile increases individual earnings by $5,292 with respect to a teacher placed in the median of the distribution (the 50th percentile). This shift from the median to the 60th percentile leads to an additional present value of $105,830 for a class of 20 pupils. For a teacher situated in the 90th percentile, the added value is close to $500,000 for a class of 20 pupils. It is important to note that the inverse relation holds true for a "bad" teacher, who occasions a loss to society. Hanushek and Rivkin likewise point out that continuity of teacher quality is an important parameter of the education system: the gain for the pupil of a year with a good teacher is canceled if she finds herself confronted with a bad one the following year.

5.4.5 Noncognitive Abilities

Noncognitive factors is a collective term for motivation, personality, temporal preferences, the ability to cooperate, conscientiousness, extraversion, emotional stability, and sociability. "Conscientiousness," a measurement of the capacity to control, regulate, and direct one's impulses, constitutes the dimension of personality most strongly associated with success in school and in professional life. The association between conscientiousness and success in school is even stronger than that between "intelligence" as measured

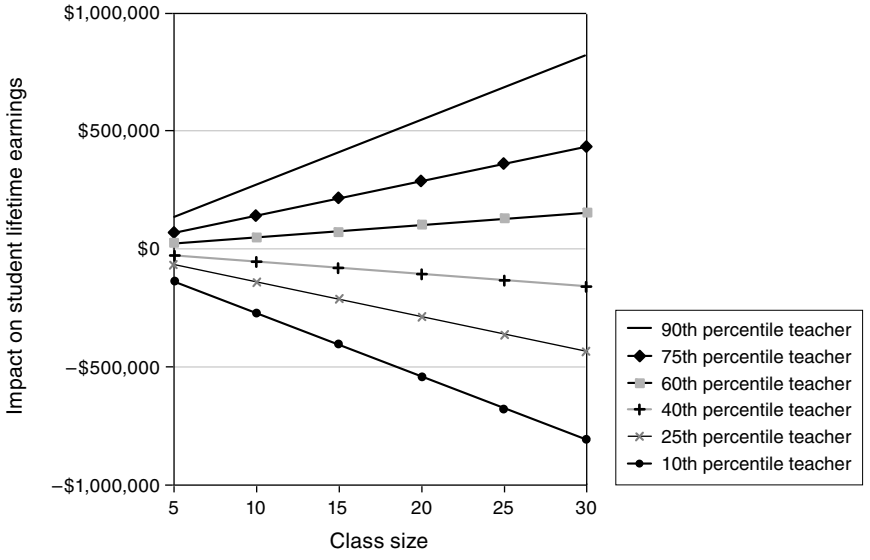


FIGURE 4.18
Impact of teacher effectiveness on student lifetime income by class size (compared to average teacher).

Source: Hanushek and Rivkin (2012).

by IQ tests and success in school. Conscientiousness of character is as much a predictor of the marks a student will get while an undergraduate in university as success in university admission tests (Cunha et al., 2006).

Numerous empirical studies find that noncognitive factors as a whole have at least as much influence as cognitive ones on the behavior of individuals. A good illustration of the weight these factors bear when it comes to education is given in figure 4.19 (Cunha et al., 2006), which shows that high levels of cognitive and noncognitive capacity are associated with lower rates of attrition from high school. Cunha et al. obtain analogous results bearing on the probability of finding oneself in prison at age 30 and of nonmarital pregnancy.

5.4.6 LIFE-CYCLE SKILL FORMATION

At what stages in the life of an individual should investment in training be a priority? Figure 4.20 well summarizes what empirical research has to teach us. It presents the rate of return to human capital at different stages of the life cycle for a person of given abilities (assuming that the same amount is invested at each stage) as a function of the person's age.

We see that, all else being equal, the rate of return on a euro invested in the education of a young person exceeds the rate of return on a euro invested in the education of someone older. For a given (and constant) opportunity cost of the sums invested equal to r, the optimal curve of investment in education dictates the biggest investment when an individual is young, or even very young, since the rate of return is visibly highest to preschool programs.

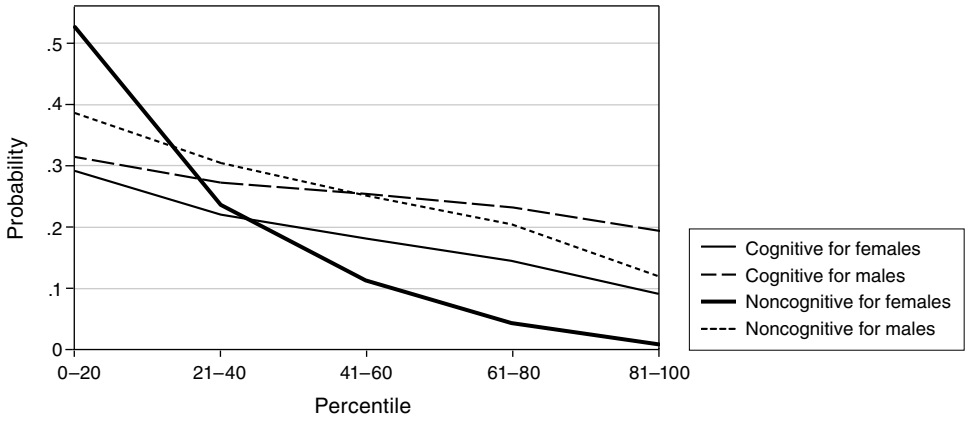


Figure 4.19
Probability of being a high school dropout and increased ability.

Source: Cunha et al. (2006).

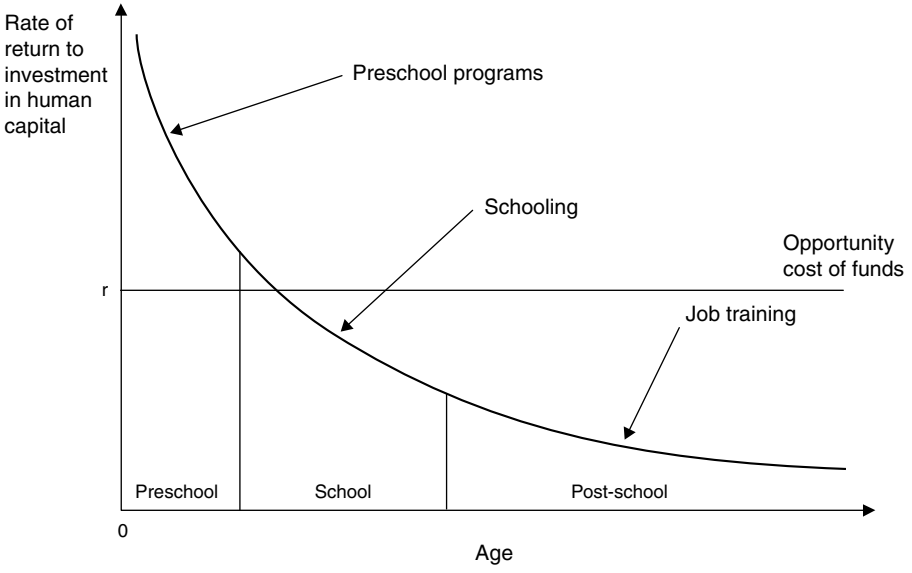


Figure 4.20
Rate of return to investment in human capital by age.

Source: Cunha et al. (2006).

Heckman (2000) and Carneiro and Heckman (2003) have brought together the results of a number of studies on the effectiveness of primary and secondary schooling in the United States; they find that expenditure per student and class size have a weakly significant impact on the probability that students will stay in school longer

and on future earnings. The return to assistance programs proves to be higher when they are aimed at young children. Heckman estimates, however, that the net return to this type of imprecisely targeted investment is negative at all levels of primary and secondary schooling in the United States, even though the quality of the teaching there is often criticized. These results do not mean that the quality of teaching has no influence on individual performance. Rather, they indicate that assistance spread thinly over the whole of primary and secondary education is not socially efficient (not in the United States, at any rate). Conversely, a number of studies have emphasized that public assistance in the training of children from disadvantaged backgrounds is highly effective (Carneiro and Heckman, 2003).

There is much evidence to prove that targeted intervention from an early age has strong and long-lasting effects (Cunha and Heckman, 2010). The most emblematic, and perhaps also the most studied, is the High/Scope Perry Preschool Program, which started in 1962 in the state of Michigan. It consists of a controlled experiment on an initial population of 123 African American children aged 3 and 4 from disadvantaged backgrounds and with low IQs (between 70 and 85). Out of these 123 children, 58 had the benefit of special classes with low teacher/pupil ratios (1/6) for 2.5 hours per day, Monday to Friday, over 2 years. During this period the teachers also had weekly interviews, lasting 1.5 hours, with the parents. The performance of the children from the test group until their 40th year is compared with that of the control group (the ones who did not attend the special classes).

Another much-studied program is that of the Chicago Child-Parent Centers, which were launched in 1967 in 11 public schools in poor neighborhoods of Chicago. Each center offered a preschool program for 3 hours per day, over a 9-month period, to children aged 3 and 4. The program also offered medical and social services, and meals were free during the sessions. As with the Perry Preschool Program, the parents were not left out of the experiment. They received frequent visits and even had the chance to finish their own education. The program was extended in 1978 to 24 centers and broadened to include after-school activities (kindergarten).

Table 4.6 presents a cost-benefit analysis of these experiments, with benefits discounted at a 3% rate. It shows that the Perry Preschool Program enabled parents to reduce their expenditure on child care. The earnings gains for the participants are large, but it is the reductions in expenditure related to criminality (incarceration, criminal justice system, damages awarded to victims) that are the most spectacular. K–12 represents the gains due to improvements in pupil quality and reduced expenditure on special education. College/adult designates the extra tuition costs paid by those who "go to college" (enroll in post-secondary education). These costs are not huge, nor are the gains realized by fewer entries into welfare. Future Generation (FG) Earnings is an evaluation of the improvement in earnings of the descendants of program participants. In the end, the cost-benefit ratio comes to 9.11. To put it another way, every dollar invested in the program generates a total collective return of $9.11. So in addition to its positive impact on the well-being of the beneficiaries and the reduction in social inequality, the money expended on the High-Quality Preschool Program makes a substantial positive contribution to the state's budget. The conclusions to be drawn for the Chicago Child-Parent Center are analogous.

TABLE 4.6
Economic benefits and costs of the High/Scope Perry Preschool Program and the Chicago Child-Parent Center Program. All values are discounted at 3% and are in 2004 dollars.

	Perry	Chicago CPC
Child care	986	1916
Earnings	40537	32099
K–12	9184	5634
College/adult	−782	−644
Crime	94065	15329
Welfare	355	546
FG earnings	6181	4894
Abuse/neglect	0	344
Total benefits	150525	60117
Total costs	16514	7738
Net present value	134011	52380
Benefits-to-cost ratio	9.11	7.77

Source: Cunha et al. (2006, table 4).

6 SUMMARY AND CONCLUSION

- Expenditure on education represents an important and growing percentage of GDP in the OECD countries. For example, in 1999 the United States and Sweden devoted 7.3% and 6.7% of GDP respectively to spending on education.
- The theory of human capital justifies educational choices by assuming that education favors the accumulation of competencies and increases wage earnings. It predicts that individuals have an interest, after completing their schooling, in gradually trimming back the amount of time they devote to training over the course of the life cycle. The profile of wages ought thus to be concave with respect to age, something solidly verified in practice.
- If the productive characteristics of individuals are unobservable, education may be looked on as a signaling activity, allowing the most productive workers to bring themselves to the notice of firms. Signaling activity may lead to overeducation, which can be reduced by cross-subsidies aimed at limiting spending on education. In practice, the significance of overeducation remains to be proved.
- Estimation of the returns to education must deal with the existence of selection bias.
- Empirical studies have a great deal of difficulty in detecting any systematic influence of expenditure on education on the performance of students.
- The estimation of earnings functions linking earnings to, among other things, the duration of schooling and professional experience, allows us to assess the

return to a year of extra education. Overall, research in this field finds that this return lies on average in the 6–15% range. On this point, it should also be noted that the returns to education are heterogeneous across individuals and across years of education.

- The empirical studies available indicate that education improves social involvement, reduces criminality, improves labor mobility, and favors the transmission of knowledge. This implies that the social returns to education are larger than the private returns to education.
- Class size and teacher quality significantly influence students' achievement and their likelihood of staying in school longer.
- The private and social returns to investment in education are higher when the investments are made in young people and ones from underprivileged backgrounds.

7 RELATED TOPICS IN THE BOOK

- Chapter 1, section 2.3: Labor cycle and retirement
- Chapter 8, section 3: Measuring discrimination
- Chapter 8, section 5.2: The importance of premarket factors
- Chapter 10, section 2: Technological progress and inequality
- Chapter 11, section 1.2: The Stolper and Samuelson theorem
- Chapter 11, section 3: Migrations
- Chapter 14, section 2.2: Why promote training?

8 FURTHER READINGS

Becker, G. (1964). *Human capital*. New York, NY: National Bureau of Economic Research.

Card, D. (1999). The causal effect of education on earnings. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 3A, chap. 30). Amsterdam: Elsevier Science.

Heckman, J., Lochner, L., & Todd, P. (2006). Earnings equations and rates of return: The Mincer equation and beyond. In E. A. Hanushek & F. Welch (Eds.), *Handbook of the economics of education* (chap. 7, pp. 307–458). Amsterdam: Elsevier.

Hanushek E. (2009). The economic value of education and cognitive skills. In *Handbook of education policy research* (pp. 39–56). New York, NY: Routledge.

Oreopoulos, P., & Salvanes, K. (2011). Priceless: The nonpecuniary benefits of schooling. *Journal of Economic Perspectives*, *25*(1), 159–184.

Spence, M. (1974). *Market signaling*. Cambridge, MA: Harvard University Press.

REFERENCES

Acemoglu, D., & Angrist, J. (2000). How large are human capital externalities? Evidence from compulsory schooling laws. *NBER Macroeconomics Annual*, *15*, 9–59.

Acemoglu, D., Autor, D. H., & Lyle, D. (2004). Women, war, and wages: The effect of female labor supply on the wage structure at midcentury. *Journal of Political Economy*, *112*(3), 497–551.

Aghion, P., & Howitt, P. (1998). *Endogenous growth theory*. Cambridge, MA: MIT Press.

Altonji, J., & Dunn, T. (1996). Using siblings to estimate the effect of school quality on wages. *Review of Economics and Statistics*, *78*, 665–671.

Angrist, J., & Krueger, D. (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, *106*, 976–1014.

Angrist, J., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics*, *114*, 553–575.

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Ashenfelter, O., Harmon, C., & Oosterbeek, H. (2000). A review of the schooling/earnings relationship, with tests for publication bias (Working Paper No. 7457). National Bureau of Economic Research, Cambridge, MA.

Ashenfelter, O., & Rouse, A. (1998). Income, schooling and ability: Evidence from a new sample of identical twins. *Quarterly Journal of Economics*, *113*, 253–284.

Ashenfelter, O., & Rouse, A. (1999). The payoff to education (Mimeo). Princeton University, Princeton, NJ.

Becker, G. (1964). *Human capital*. New York, NY: National Bureau of Economic Research.

Ben-Porath, Y. (1967). The production of human capital and the life cycle of earnings. *Journal of Political Economy*, *75*, 352–365.

Blundell, R., & Costa Dias, M. (2009). Alternative approaches to evaluation in empirical microeconomics. *Journal of Human Resources*, *44*(3), 565–640.

Blundell, R., Dearden, L., & Sianesi, B. (2005). Evaluating the effect of education on earnings: Models, methods and results from the National Child Development Survey. *Journal of the Royal Statistical Society Series A*, *168*(3), 473–512.

Borjas, G. (2010). *Labor economics* (5th ed.). New York, NY: McGraw-Hill Irwin.

Card, D. (1999). The causal effect of education on earnings. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 3A, chap. 30). Amsterdam: Elsevier Science.

Card, D., & Krueger, A. (1992). Does school quality matter? Returns to education and the characteristics of public schools in the United States. *Journal of Political Economy*, *100*, 1–40.

Carneiro, P., & Heckman, J. (2003). Human capital policy (Working Paper 9495). National Bureau of Economic Research.

Carneiro, P., Heckman, J., & Vitacyl, E. (2011). Estimating marginal returns to education. *American Economic Review*, *101*(6), 2754–2781.

Chevalier, A., & O'Sullivan, V. (2007). Mother's education and birth weight (IZA Discussion Paper No. 2640).

Cho, K., & Kreps, K. (1987). Signaling games and stable equilibria. *Quarterly Journal of Economics*, *102*, 179–221.

Cunha, F., & Heckman, J. (2010). Investing in our young people (IZA Discussion Paper No. 5050).

Cunha, F., Heckman, J., Lochner, L., & Masterov, D. (2006). Interpreting the evidence on life cycle skill formation. In E. Hanushek & F. Welch (Eds.), *Handbook of the economics of education* (chap. 12, pp. 697–812). Amsterdam: Elsevier.

Currie, J., & Moretti, E. (2003). Mother's education and the intergenerational transmission of human capital: Evidence from college openings and longitudinal data. *Quarterly Journal of Economics*, *118*(4), 1495–1532.

Currie, J., & Thomas, D. (2001). Early test scores, socioeconomic status, school quality, and future outcomes. *Research in Labor Economics*, *20*, 103–132.

Foster, A., & Rosenzweig, M. (1996). Technical change in human capital return and investments: Evidence from the green revolution. *American Economic Review*, *86*, 931–953.

Glaeser, E., Ponzetto, G., & Shleifer, A. (2007). Why does democracy need education? *Journal of Economic Growth*, *12*, 77–99.

Goldin, C. (2001). The human capital century and American leadership: Virtues of the past. *Journal of Economic History*, *61*, 263–292.

Goldin, C., & Katz, L. (2001). Decreasing (and then increasing) inequality in America: A tale of two half centuries. In F. Welch (Ed.), *The causes and consequences of increasing income inequality*. Chicago, IL: University of Chicago Press.

Goux, D., & Maurin, E. (1994). Education, expérience et salaires. *Economie et Prévision*, *116*, 155–179.

Grossman, M. (2006). Education and nonmarket outcomes. In *Handbook of the economics of education* (vol. I, chap. 10, pp. 578–635). Amsterdam: Elsevier.

Hanushek, E. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature*, *24*, 1141–1177.

Hanushek, E. (2002). Public provided education. In A. Auerbach & M. Feldstein (Eds.), *Handbook of public economics* (vol. 4, chap. 30, pp. 2045–2141). Amsterdam: Elsevier.

Hanushek, E., & Kimko, D. (2000). Schooling, labor force quality, and the growth of nations. *American Economic Review*, *90*, 1184–1208.

Hanushek, E., & Rivkin, S. (2012). The distribution of teacher quality and implications for policy. *Annual Review of Economics*, *4*(1), 131–157.

Hanushek, E., Rivkin, S., & Taylor, L. (1996). Aggregation and the estimated effects of school resources. *Review of Economics and Statistics*, *78*, 611–627.

Hanushek, E., & Woessmann, L. (2007). The role of school improvement in economic development (Working Paper No. 12832). NBER, Cambridge, MA.

Hanushek, E., & Zhang, L. (2009). Quality-consistent estimates of international schooling and skill gradients. *Journal of Human Capital*, *3*, 107–143.

Heckman, J. (1976). A life cycle model of earnings, learning and consumption. *Journal of Political Economy*, *84*, 11–44.

Heckman, J. (2000). Policies to foster human capital. *Research in Economics*, *54*(1), 3–56.

Heckman, J., Lochner, L., & Todd, P. (2006). Earnings equations and rates of return: The Mincer equation and beyond. In E. A. Hanushek & F. Welch (Eds.), *Handbook of the economics of education* (chap. 7, pp. 307–458). Amsterdam: Elsevier.

Heckman, J., Lochner, L., & Todd, P. (2008). Earnings functions and rates of return. *Journal of Human Capital*, *2*(1), 1–31.

Helliwell, J., & Putnam, R. (2007). Education and social capital. *Eastern Economic Journal*, *33*(1), 1–19.

Imbens, G., & Angrist, J. (1994). Identification and estimation of local average treatment effects. *Econometrica*, *62*(2), 467–475.

Jaeger, D., & Page, M. (1996). Degrees matter: New evidence on sheep skin effects in the returns to education. *Review of Economics and Statistics*, *78*, 733–740.

Krueger, A. (1999). Experimental estimates of education production functions. *Quarterly Journal of Economics*, *114*, 497–532.

Krueger, A., & Lindahl, M. (2001). Education for growth: Why and for whom? *Journal of Economic Literature*, *39*(4), 1101–1136.

Lochner, L., & Moretti, E. (2004). The effect of education on criminal activity: Evidence from prison inmates, arrests and self-reports. *American Economic Review*, *94*(1), 155–189.

Machin, S., Marie, O., & Vujić, S. (2011). The crime reducing effect of education. *Economic Journal*, *121*, 463–484.

Machin, S., Pelkonen, P., & Salvanes, K. (2011). Education and mobility. *Journal of the European Economic Association*, *10*(2), 417–450.

Martorell, P., & Clark, D. (2010). The signaling value of a high school diploma (Working Paper No. 557). Princeton University Industrial Relations Section. http://irs.princeton.edu/pubs/pdfs/557.pdf.

Mas-Colell, A., Whinston, M., & Green, J. (1995). *Microeconomic theory*. New York, NY: Oxford University Press.

Milligan, K., Moretti, E., & Oreopoulos, P. (2004). Does education improve citizenship? Evidence from the United States and the United Kingdom. *Journal of Public Economics*, *88*(9–10), 1667–1695.

Mincer, J. (1974). *Schooling, experience and earnings*. New York, NY: National Bureau of Economic Research.

Moretti, E. (2004). Estimating the social return to higher education: Evidence from longitudinal and repeated cross-sectional data. *Journal of Econometrics*, *121*(1–2), 175–212.

Murnane, R., Willett, J., Braatz, M., & Duhaldeborde, Y. (2001). Do different dimensions of male high school students' skills predict labor market success a decade later? Evidence from the NLSY. *Economics of Education Review*, *20*, 311–320.

Murnane, R., Willett, J., & Levy, F. (1995). The growing importance of cognitive skills in wage determination. *Review of Economics and Statistics*, *77*(2), 251–266.

Nelson, R., & Phelps, E. (1966). Investment in humans, technological diffusion, and economic growth. *American Economic Review*, *56*, 69–75.

OECD. (2012). *Education at a glance*. Paris: OECD Publishing.

Oreopoulos, P. (2006). Estimating average and local average treatment effects of education when compulsory schooling laws really matter. *American Economic Review*, *96*(1), 152–175.

Oreopoulos, P., & Salvanes, K. (2011). Priceless: The nonpecuniary benefits of schooling. *Journal of Economic Perspectives*, *25*(1), 159–184.

Psacharopoulos, G. (1985). Returns to education: A further international update and implications. *Journal of Human Resources*, *20*, 583–604.

Rauch, J. (1993). Productivity gains from geographic concentration of human capital: Evidence from the cities. *Journal of Urban Economics*, *34*, 380–400.

Solon, G. (1999). Intergenerational mobility in the labor market. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 3A, chap. 29). Amsterdam: Elsevier.

Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics*, *87*, 355–374.

Swinkels, J. (1999). Education signaling with preemptive offers. *Review of Economic Studies*, *66*, 949–970.

Topel, R. (1999). Labor markets and economic growth. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 3C, chap. 30). Amsterdam: Elsevier Science.

Weiss, A. (1983). A sorting-cum-learning model of education. *Journal of Political Economy*, *91*, 420–442.

Weiss, Y. (1986). The determination of life-cycle earnings: A survey. In O. Ashenfelter & R. Layard (Eds.), *Handbook of labor economics* (vol. 1, chap. 11, pp. 603–640). Amsterdam: Elsevier Science.

PART TWO

Imperfectly Competitive Labor Markets

CHAPTER 5

Job Search

In this chapter we will:

- See what unemployed workers do to find a job
- Study how the duration of unemployment depends on the characteristics of unemployed workers
- Study how unemployment insurance influences the duration of unemployment
- Learn how economists empirically evaluate this influence
- Examine how the contribution of Lalive, van Ours, and Zweimüller (2006) assesses the impact of modifications in the Austrian unemployment insurance system on the duration of unemployment (The main results of this contribution can be replicated with data and programs available at www.labor-economics.org.)
- Review the effects of job search help and checking on job search effort
- See how the equilibrium search model explains why identical workers can be paid differently
- Learn why wages rise, on average, as workers gain experience and why large firms pay higher wages than small firms to identical workers

INTRODUCTION

The economic theory of labor supply pays no attention to the time and cost of looking for work. The consumption of "leisure"—even when this term is extended to cover home production—remains the sole alternative to waged work, and by definition an agent who utilizes the total amount of time at his disposal in the form of leisure is described as a *nonparticipant*. So from this perspective there is no place for the unemployed person, even though her principal activity amounts to looking for work. Such a description of the labor market implicitly assumes a structure of *perfect information*. It supposes that

each agent knows all the particulars about all the jobs on offer and that he merely has to decide the number of hours—potentially as low as zero—that he wants to devote to work, given the (supposedly) single and universally known wage prevailing in the labor market. There is no need to *look* for a job that would suit him. Such a hypothesis is no doubt too simplistic, to the extent that, as we document below, many unemployed workers do devote considerable effort to looking for work without getting satisfactory job offers. So, we must examine the consequences of *imperfect information*. This is precisely the purpose of job search theory: to study the behavior of an individual who has imperfect information about jobs and wages.

In the job market, the imperfection of the available information occurs in the form of a number of possible wages that an agent might be able to command. Hence the job seeker surveys the labor market so as to find the highest wage being paid for the services she can supply. This procedure is no different from that adopted by a person looking for an apartment (at the best possible rent) or a loan (at the best possible rate of interest). It was Stigler (1961, 1962) who first highlighted this common process in all markets where information is imperfect. The modern theory of the job search arose in the 1970s with the formalizations of McCall (1970) and Mortensen (1970).

The first section of this chapter portrays the activity of job search, using surveys that describe the amount of time that persons both employed and unemployed spend looking for work each day. It lays particular stress on the influence of economic incentives on the amount of time devoted to job search.

Section 2 of this chapter lays out the basic job search model, in which an agent keeps looking as long as she entertains the hope of improving her welfare by continuing to search. This model is useful to understand how the duration of the search depends on individual preferences and the overall characteristics of the environment in which it takes place. The theory of job search is not in conflict with the theory of labor supply. By giving a prominent role to imperfect information, job search theory adds the category "unemployed" to those of "employed" and "nonparticipant." In this way it sheds supplementary light on the decision to participate in the labor market, which no longer takes the form of a choice between work and nonparticipation; rather, the choice now lies in knowing whether it is *worthwhile* to *look* for work. In other words, to hold a paid job you must first have decided to look for one. A good synthesis of this theory can be found in Mortensen (1986) and Mortensen and Pissarides (1999).

Section 3 presents the empirical analysis of a reform of the Austrian unemployment insurance system carried out by Lalive et al. (2006). It details the econometric strategies utilized to identify and evaluate the effects of reforms to a system of unemployment insurance on the duration of unemployment and the rate of return to employment. This section also presents the econometric techniques adopted to estimate the models of duration and offers a synthesis of the principal empirical results in the domain of job search.

In the first three sections of this chapter, the distribution of wages is a given parameter, which is not explained. We will see in section 4 that, *when the distribution of wages is rendered endogenous*, the search model allows us to go further and to explain why identical workers can be paid differently, contrary to the model of perfect competition studied in chapter 3, which assumes that wage differences reflect only differences in talent and the hardship of tasks. We start to see in section 4 that many empirical studies

do suggest that wage differences do not solely reflect differences of talent and the hardship of tasks. It is apparent that persons who have the luck to be hired by larger or more productive firms have higher wages than persons with identical characteristics who have not had the same luck. These differences in remuneration are enduring and represent an important portion, on the order of one quarter, of wage differences as a whole. The existence of job search costs or "frictions" can explain this phenomenon. In the presence of job search costs, the competition that would allow wage earners to be remunerated at their marginal productivity cannot fully play out. Search costs cause "rents" to materialize: the wage earner incurs a loss if she loses her job when it is costly to find another; the employer also incurs a loss when a worker quits if it is costly to hire another. To avoid this loss, the employer is ready to give wage earners a remuneration higher than what they could get by looking for another job. For her part, the worker is ready to accept a wage lower than what she could extract from another employer if job search is costly. Hence the costs of job hunting can exert influence on wage formation. From that standpoint, we will analyze the behavior of employers in the context of the job search model. For a long time the theory of job search developed within the framework of partial equilibrium, which left it unable to explain the formation of the wage distribution that confronts job seekers. To make it complete, the behavior of employers has been introduced so as to arrive at a description of labor market equilibrium. By attributing well-defined strategic behavior to firms, these "equilibrium search models" are able to portray the process of wage formation as endogenous and to explain why identical workers can be paid differently. We see that in reality it is essential to take into account not just job search costs and the search activity associated with them but also on-the-job search in order to explain the empirical properties of wage distribution.

1 WHAT DO JOB SEEKERS DO?

Discussion of the situation of unemployed persons often devolves into mere caricature. Some take the view that the unemployed can always find work if they really want to, so there is no point in supporting them financially while they do so. At the opposite extreme, the unemployed are viewed as victims who deserve the most generous indemnification possible. It is possible to rise above these caricatural stances thanks to surveys that tell us precisely how the unemployed react to economic incentives. Such information is very useful in making unemployment insurance function more effectively.

Job search is linked to the work available: it is an activity aimed at the goal of earning remuneration. But the returns to job search are generally different from the returns to wage-earning activity. If you are unemployed, the income you can derive from an hour of job search will certainly be less than what you could obtain from an hour of work, if you had a job. From an hour of work you derive a wage, whereas an hour of searching for work gives you the chance of obtaining a job interview, or in the best case, of being hired. This variation in return implies that time devoted to job search is highly likely to be shorter than the time devoted to waged work, which is indeed the case, as we will see. We will show further that time devoted to job search, just like time devoted to work, is sensitive to financial incentives.

1.1 How Job Seekers Spend Their Time

Detailed surveys of how persons spend their time can be mined to shed valuable light on the behaviors of the unemployed and wage earners (Krueger and Mueller, 2010, 2011, 2012). Table 5.1, based on the American Time Use Survey (ATUS), presents comparative information about how wage earners and unemployed persons in the United States spend their time. The ATUS is a nationally representative sample drawn from households that have completed their final interview for the Current Population Survey (CPS). Individuals are queried in detail on how they used their time the day before the day of the survey. Unemployed workers are individuals who declared that they did not work in the previous week, that they actively looked for work in the previous four weeks, and that they were available to start work. Employed workers are all those who declared that they had a waged job. The sample is restricted to people age 20 to 54 years. Job search activities typically include calling or visiting a labor office, reading and replying to job advertisements, and job interviews.

Table 5.1 reveals that the unemployed spend on average 32 minutes per day looking for work, a duration that is far less than the time wage earners spend at work, which amounts to 325 minutes. This difference may flow from differences in observed characteristics, such as age, educational qualification, gender, and differences in unobserved ones, such as psychological state, between the unemployed and wage earners. It may also flow from responses to incentives. The labor supply model suggests that two effects may influence the amount of time spent searching for work. First, given that an hour of job search returns less than an hour of waged work, the *substitution effect* ought to result in less time being devoted to job search by an unemployed person than a waged employee devotes to working. Second, since the income of an unemployed person is less than that of a wage earner, the *income effect* ought to result in more time being spent on job search. Table 5.1 suggests that the substitution effect is largely dominant. Thus the unemployed devote more time to domestic production, shopping, and taking care of other members of their household than wage earners do. Sleep, leisure, sports, and socializing also bulk large in their use of their time.

It should be stressed that the time devoted to job search by the unemployed reported in table 5.1 is an average that comprises a high proportion of persons whose use of their time indicates that they did not look for a job at all the day before the survey. In fact, just 20% of unemployed persons engaged in job search activity the day before

TABLE 5.1
Average minutes per day by activity and employment status in the United States in 2003–2006.

	Employed	Unemployed
Sleep	496	555
Personal care and eating	110	97
Home production, shopping, care of others	158	254
Leisure, travel, sports, and socializing	320	442
Work	325	10
Job search	1	32

Source: Krueger and Mueller (2012, table 3, p. 773) and personal computations.

the survey. The average daily search duration of unemployed persons who did actually hunt for work the day before they were queried is thus 160 minutes.

Table 5.1 also shows that those earning a wage devote on average no more than a minute per day to job search. The fact is, only 0.7% of wage earners search for (another) job, and those who do so spend about 14 minutes a day on the task. From this perspective, the situation of wage earners is much like that of inactive persons, meaning those who declare that they do not have a job and are not searching for one: 0.7% of inactive persons state that they took steps to search for work the day before the survey and did so for an average duration similar to wage earners who searched.

The preceding data apply to the United States, but Krueger and Mueller (2010, 2012) have reported analogous observations for Canada and European countries.

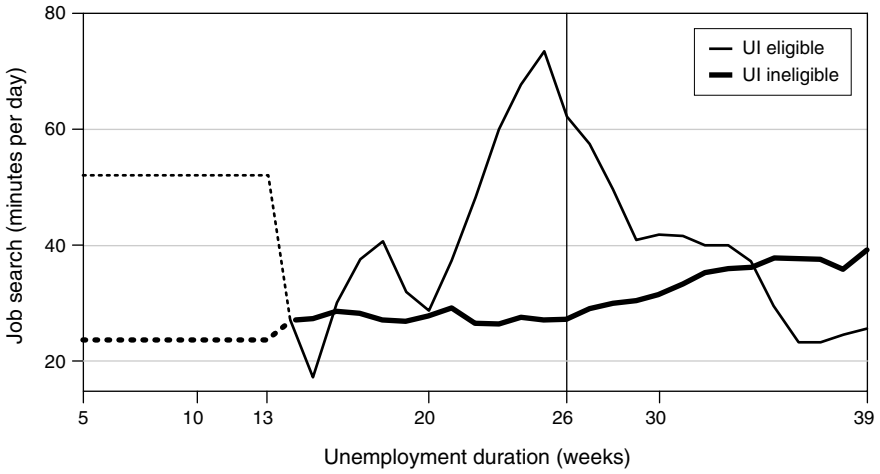
1.2 How Economic Incentives Affect the Time Dedicated to Job Search

Research on job search activity shows that unemployed persons react to economic incentives. Thus Krueger and Mueller (2010) find that workers who expect to be recalled by their previous employer search substantially less than the average unemployed worker, and that across the 50 U.S. states and D.C. the time spent looking for a job is inversely correlated to the level of unemployment benefits, with an elasticity between -1.6 and -2.2. Analysis of a more specific survey of unemployed persons receiving benefits in New Jersey leads, however, to an elasticity of around -0.3, clearly weaker in absolute terms (Krueger and Mueller, 2011). Krueger and Mueller (2010) also find that job seekers who likely have less access to financial resources (e.g., because they do not have a working spouse) tend to respond more to unemployment insurance (UI) benefits than do those with greater financial wherewithal.

As a general rule, more generous unemployment benefits paid for a longer period diminish the amount of time devoted to job search. This phenomenon is illustrated in figure 5.1, which reports the amount of time spent searching by persons eligible for unemployment insurance, and those not eligible, as a function of the duration of their spell of unemployment in the United States over the period 2003–2006.

This figure shows that job search by unemployed persons benefiting from unemployment insurance intensifies as week 26 (the point at which benefits will come to an end) approaches. The increase in the amount of time devoted to the search is considerable, going from less than 20 minutes to more than 70 minutes between week 15 and week 26, and falling back to 20 minutes toward week 34. This observation, which is not replicated in the case of noneligible unemployed persons, strongly suggests that financial support during unemployment influences the amount of time devoted to job search. It is verisimilar to suppose that unemployed persons receiving benefits intensify their job search as their period of eligibility draws to a close in order to avoid the drop in income set to occur at that time if they have not found work.

But figure 5.1 also reveals that the amount of time devoted to job search by those not receiving unemployment benefits is, for the most part, inferior to that of the eligible unemployed, which runs counter to the supposition that the more generous the unemployment benefit, the less time will be spent on job search. Krueger and Mueller (2010) show that this gap persists even for persons with the same *observable* characteristics (age, gender, education, marital status, number of children). It might, however, arise



Note: The dotted lines refer to the average of time spent on job search before week 14.

Figure 5.1
Job search by unemployment duration in the United States over the period 2003–2006.

Source: Krueger and Mueller (2010, figure 3, p. 305).

from *unobservable* characteristics like personal motivation, self-esteem, or any other characteristic capable of influencing both eligibility for unemployment insurance and the intensity of job search. It is in fact highly likely that psychologically fragile persons who experience difficulty in career planning will have had shorter spells of past employment, which bar them from access to unemployment benefits, and will likewise be less motivated to hunt for work, irrespective of the level of unemployment benefit. In this context, the difference between the two groups in the intensity of job search flows not from the generosity of unemployment insurance but rather from differences in the unobserved characteristics of the two populations. Hence the observation that the duration of job search differs between the eligible and noneligible unemployed does not warrant the conclusion that financial support during unemployment has a causal impact on the amount of time devoted to job search. In section 3.1 of this chapter we will examine in greater detail the empirical strategies used to try to identify such a causal impact.

1.3 Methods of Job Search: An Internet Revolution?

On the basis of the National Longitudinal Survey of Youth (NLSY97) conducted during 2008–2009, Kuhn and Mansour (2011) have documented with great precision the search methods of American job seekers aged between 24 and 28. Table 5.2 presents their principal results. The data used make it possible to know, for each search method, the proportion of unemployed persons using the Internet and the proportion of unemployed using *offline* methods. Clearly these two channels are not exclusive.

We observe that, on average, an unemployed person makes use of 1.58 of the 9 methods of *offline* research classified as active, the most frequent being "contacted friends or relatives," followed by "contacted employer directly." The proportion is 1.44

Table 5.2
Search methods of unemployed workers.

Method	Share of workers using offline methods	Share of workers using online methods
Active search method		
Contacted employer directly	0.36	0.29
Contacted public employment agency	0.19	0.19
Contacted private employment agency	0.07	0.08
Contacted friends or relatives	0.44	0.11
Contacted school/university employment center	0.05	0.06
Sent out résumés or filled out applications	0.24	0.48
Checked unions or professional registers	0.03	0.03
Placed or answered ads	0.16	0.17
Other active methods	0.04	0.03
Total active search methods	**1.58**	**1.44**
Passive search methods		
Looked at ads	0.30	0.32
Attended job training programs or courses	0.06	0.03
Other passive methods	0.02	0.02
Total passive search method	**0.38**	**0.37**

Source: Kuhn and Mansour (2011, table 2, p. 22).

out of 9 for *online* job search, the most frequent being "sent out resumes or filled out applications," followed by "contacted employer directly." It is interesting to note that job search on the Internet is much more "formal" than offline search. The "contacted friends or relatives" approach is adopted by only 11% of unemployed persons via Internet, whereas it is adopted by 44% of offline job seekers. Combining the online and offline methods, a typical job seeker uses 3.02 of the nine active methods. Table 5.2 also shows that the use of passive search methods is identical online and offline, the passive method most frequently resorted to being "looked at ads" in both cases.

Table 5.2 testifies to the importance the Internet has assumed in job search. It might lead us to suppose that the Internet has increased the efficiency of job search, probably by supplying more information, more rapidly, to job seekers and employers. Kuhn and Skuterud (2004), based on data analogous to those of Kuhn and Mansour (2011), but for the period 1998–2000, arrive at no such conclusion. Against all expectation, they found that for identical observable characteristics, unemployed workers who look for work online have longer unemployment durations than non-Internet searchers. With data relative to the same population but pertaining to 2008–2009, hence markedly more recent, Kuhn and Mansour (2011) find, in contrast, that for identical observable characteristics, unemployed workers who look for work online have an average duration of unemployment 25% shorter than non-Internet searchers.

Kuhn and Mansour (2011) put forward two reasons for this spectacular trend reversal. The first is the improved quality of the majority of Internet job search sites, both public and private. The second is the enormous growth in penetration by the Internet: over the 10 years that separate the two studies, the proportion of young unemployed who looked for work online went from 24.2% to 74.4%. By connecting many more workers with many more employers in very short timespans and at very low cost, the Internet has in all likelihood effected a large reduction in labor market frictions. Research on this subject is still too sparse, however, for us to accept such a conclusion unreservedly.

We now describe job search behavior more precisely, with the help of a model that has proved its utility for understanding and evaluating with precision the effects of the various parameters of unemployment insurance systems.

2 BASIC JOB SEARCH THEORY

Job search theory arises initially out of a basic model—called today the partial model—describing the behavior of a person looking for work in a situation of imperfect information. This model furnishes precise conclusions about the effects of a change in the environment or in economic policy. The basic model is grounded, however, on oversimple hypotheses, and these we must abandon in order to describe the reality of the search process better. For one thing, in this model all the unemployed have access, in exogenous fashion, to unemployment insurance benefits, they are not allowed to select the intensity of their search, and they cannot look for (another) job once they are employed. Finally, the basic model is situated in a stationary environment. We first lay out the basic model, then analyze the changes that emerge as we abandon these four hypotheses.

2.1 The Basic Model

In the job search model, the optimal strategy of a person looking for work consists simply of choosing a *reservation wage* that represents the lowest remuneration he will accept. The amount chosen depends on all the parameters that go to make up the economic environment, in particular the benefits paid to those who are unemployed and the arrival rate of job offers. Hence most often it is enough to know how the reservation wage varies in order to discern the effects of economic policy on the duration of unemployment. As well, when it is linked to the labor supply model presented in chapter 1, the job search model makes it possible to shed light on the choice of nonparticipation, unemployment, or work.

2.1.1 The Search Process and the Reservation Wage

The basic job search model aims to describe the behavior of an unemployed person who dedicates all of his efforts to looking for a job, when the conditions in which this search takes place do not vary over time. The dynamic aspect of the model makes it possible to define the optimal job search strategy. The model is explicitly dynamic but is situated in a stationary environment.

The Discounted Expected Utility of an Employee

The main hypothesis of the job search model is that the job seeker does not know exactly what wage each job pays. So by looking, he can expect to improve his prospects of earning. We epitomize these imperfections in the available information by postulating that the job seeker knows only the cumulative distribution of the possible wages. We further assume that this distribution is the same at each date and that successive wage offers are independent draws from this distribution. This stationarity assumption means that, at any time, a person looking for work faces the same structure of information. We use $H(.)$ to denote the cumulative distribution function of all possible wages.

A job offer comes down to the proposal of a constant real wage w, which the worker will receive on each date as long as he remains with the firm that makes the offer. If we assume that the agent is risk-neutral and if, for the sake of simplicity, we leave out of account the disutility of work, his *instantaneous* utility then simply equals w. This means that over a short interval of time, $\mathrm{d}t$ in length,[1] the agent attains a level of instantaneous satisfaction equal to $w\mathrm{d}t$. Let us further assume that over each short interval of time $\mathrm{d}t$, any job whatsoever can disappear at the rate $q\mathrm{d}t$, where $q > 0$ is a constant exogenous parameter. Over each short interval of time $\mathrm{d}t$, a waged worker thus loses his job at the rate $q\mathrm{d}t$. Let us assume that the real instantaneous rate of interest r is constant and exogenous. A single dollar invested in the financial market on date t brings in $1 + r\mathrm{d}t$ dollars in $t+\mathrm{d}t$. The discounted value of a dollar at date t that will be available at date $t+\mathrm{d}t$ is thus equal to $1/(1 + r\mathrm{d}t)$. The term $1/(1 + r\mathrm{d}t)$ thus represents the discount factor over each short interval of time $\mathrm{d}t$. In a stationary state, the discounted expected utility V_e of an employed person receiving wage w satisfies the following relation:

$$V_e = \frac{1}{1 + r\mathrm{d}t}\left[w\mathrm{d}t + (1 - q\mathrm{d}t)V_e + q\mathrm{d}tV_u\right] \tag{5.1}$$

This relation indicates that the discounted expected utility stemming from being hired is equal to the discounted sum of the flow of income $w\mathrm{d}t$ over the interval of time $\mathrm{d}t$, and the discounted expected future income. With probability $(1 - q\mathrm{d}t)$ this future income does coincide with the expected utility V_e associated with continued employment, and with complementary probability $q\mathrm{d}t$ it conforms instead to V_u, the discounted expected utility of an unemployed person. Multiplying the two sides of relation (5.1) by $1 + r\mathrm{d}t$ and rearranging the terms of this expression, we arrive at:

$$rV_e = w + q(V_u - V_e) \tag{5.2}$$

This equation is easy to interpret.[2] It shows that, at every moment, a job entails discounted expected flow of income rV_e equal to wage w, to which is added average income

[1] The unfolding of time may be described in continuous or discrete manner; we have chosen the former because it is analytically more simple and has been adopted almost universally by all published work in this field.

[2] Mathematical appendix D at the end of the book supplies a rigorous proof of formulas analogous to equation (5.2) and shows that they effectively correspond to the stationary state of a model where a particular event (here, the loss of work) follows a Poisson process.

$q(V_u - V_e)$ deriving from any possible change in the employee's status. This average income is in fact a loss resulting from the wage worker's having quit his job.

Equality (5.2) allows us to express the discounted expected utility of an employee receiving wage w—which we henceforth denote $V_e(w)$—in the following manner:

$$V_e(w) - V_u = \frac{w - rV_u}{r + q} \tag{5.3}$$

It is thus apparent that the difference between the expected utility of an employee and that of an unemployed person expands with the wage accepted and shrinks with the discounted expected utility of the unemployed person.

The Optimal Search Strategy

To simplify the exposition we will assume that a job seeker can only meet a single employer on any date (see Mortensen, 1986, for the possibility of multiple offers). The employer offers the job seeker the constant wage w over the duration of her employment, which she is free to accept or refuse. The optimal job search strategy is then as follows:

1. If the job seeker receives no offer on date t, she continues looking. This behavior results from the stationarity of intertemporal utility V_u.

2. If the job seeker receives a wage offer w, she accepts if $V_e(w) > V_u$. If not, she continues looking.

Since a job seeker's expected utility V_u does not depend on a particular wage offer w, relation (5.3) shows that $V_e(w)$ is an increasing linear function of the wage offered. This relation also shows that phase 2 of the search strategy amounts to the adoption of a "stopping rule" that dictates accepting wage w if and only if it is superior to a threshold-value x defined by:

$$x = rV_u \tag{5.4}$$

The acceptance of an offer exactly equal to x procures for the job seeker the same level of utility that she gets by remaining unemployed; in other words, $V_e(x) = V_u$. As in the theory of labor supply laid out in chapter 1, wage x continues to be called the *reservation wage*, but we will see in section 2.1.3 that it means something tangibly different.

Thus the job search model shows that the optimal strategy consists of continuing to hunt for a job as long as incoming job offers entail wages below the reservation wage. The optimality of this strategy, which is known as sequential search, was demonstrated by McCall (1970). Other strategies are conceivable. Stigler (1962), for example, proposes a nonsequential strategy that consists of deciding, at the moment unemployment commences, to review a set number of job offers. This strategy is not optimal, as it may lead to continuing to search for a job even after receiving an offer greater than the reservation wage.

The Discounted Expected Utility of a Job Seeker

The existence of the stopping rule makes it possible to deduce numerous characteristics of the search process from those of the reservation wage. To make the factors that

determine the reservation wage explicit, we need to specify more precisely the discounted expected utility V_u of a job seeker. Accordingly, we will designate by λ the arrival rate of job offers. This rate encapsulates the difficulties encountered while looking for a job. It reflects the general state of the labor market, but it also depends on the personal characteristics of the job seeker—age and educational qualifications, for example—and the effort he puts into the search. In the basic model, we assume that this rate λ is a constant exogenous quantity. Moreover, the search for a job entails costs at every turn. Some are financial, like the cost of getting about, buying specialized magazines, and sending out applications. But it is equally necessary to include the opportunity cost of the search, in other words, the value of a period of time that could have been devoted to other activities. All these costs will be summed up, at each date, by a single scalar $c > 0$. There are also gains associated with periods of looking for work. These comprise unemployment benefits and also perhaps the consumption of domestic production and leisure. If, for each date, we express the sum of these gains by the scalar $b > 0$, the *net* instantaneous income from looking for work, denoted z, is then equal to $(b - c)$.

At any moment the status of a job seeker may change with rate λ. If he does actually receive an offer, he will not accept unless the wage that goes with it is more than his reservation wage x. The discounted utility V_λ expected upon receiving an offer of employment is thus equal to:

$$V_\lambda = \int_0^x V_u \mathrm{d}H(w) + \int_x^{+\infty} V_e(w) \mathrm{d}H(w)$$

Conversely, if the job seeker receives no offers, he keeps looking, which procures for him a discounted expected utility equal to V_u. Now, during a short interval of time $\mathrm{d}t$ in length, a job seeker gains $z\mathrm{d}t$ and has a probability $\lambda \mathrm{d}t$ of receiving a job offer. In the stationary state, his expected utility thus satisfies:

$$V_u = \frac{1}{1 + r\mathrm{d}t} [z\mathrm{d}t + \lambda \mathrm{d}t V_\lambda + (1 - \lambda \mathrm{d}t) V_u]$$

If we multiply the two sides of this equality by $1 + r\mathrm{d}t$ and rearrange terms, we find that a job seeker's discounted expected utility is defined by the following trade-off equation:

$$rV_u = z + \lambda \int_x^{+\infty} [V_e(w) - V_u] \, \mathrm{d}H(w) \qquad \textbf{(5.5)}$$

Like relation (5.2) defining an employee's discounted expected utility, this equation has to be interpreted by examining the various ways the assets V_u of an unemployed person may be invested. In the "financial" market these assets will bring in rV_u at any moment, while if "invested" in the labor market they will procure income z augmented by the value $\lambda(V_\lambda - V_u)$ of the average gain linked to the change of status of a person who is looking for work.

Reservation Wage, Hazard Rate, and Average Duration of Unemployment

With the help of relations (5.3) and (5.4), which define respectively the intertemporal utility $V_e(w)$ of an employee and the reservation wage x as a function of the discounted expected gain V_u of an unemployed person, we easily arrive at the following

equation, which implicitly characterizes the reservation wage as a function of the parameters of the model:

$$x = z + \frac{\lambda}{r+q}\int_x^{+\infty}(w-x)\mathrm{d}H(w) \tag{5.6}$$

We can show (1) that there is only one optimal value for this reservation wage, and (2) that it maximizes the intertemporal utility of a job seeker. For that, we need merely observe that relation (5.5) defines V_u as a function of x and verify that the derivative of this function is null for the value of x given by (5.6). Equation (5.6) shows that the reservation wage is equal to the net income from the job search plus the discounted expected value of what the job search can yield above the reservation wage. This way of characterizing the reservation wage is instructive, for it brings out clearly the optimality of the search strategy adopted by the job seeker.

The values of two other important variables flow from knowing the reservation wage. These are the "hazard rate," or the exit rate from unemployment, and the average duration of unemployment. Since a job seeker becomes employed when (1) she receives a wage offer—which occurs at rate λ—and (2) the offer is at least equal to her reservation wage—which occurs with probability $[1 - H(x)]$—the exit rate from unemployment takes the value $\lambda[1 - H(x)]$ at any moment. When the number of job seekers is large, this rate merges with the hazard rate. The average duration of unemployment, denoted T_u, is then given by:

$$T_u = \frac{1}{\lambda[1-H(x)]} \tag{5.7}$$

The interpretation of this last relation is very intuitive: it means that if a job seeker has one chance in ten of becoming employed in any week, she will on average remain unemployed for ten weeks.[3] Relation (5.7) also shows that the average duration of unemployment is an increasing function of the reservation wage: when a person who is looking for work raises the level of the wage she is demanding, on average it prolongs the duration of the search.

2.1.2 Comparative Statics of the Basic Model

The comparative statics properties of the job search model are very easily obtained if we write relation (5.6), which defines the reservation wage, in the following form:

$$\Phi(x,z,r,\lambda,q) = 0 \quad \text{with} \quad \Phi(x,z,r,\lambda,q) \equiv x - z - \frac{\lambda}{r+q}\int_x^{+\infty}(w-x)\mathrm{d}H(w) \tag{5.8}$$

We can easily verify that the partial derivatives of the function Φ possess the following properties:

$$\Phi_x > 0, \quad \Phi_z < 0, \quad \Phi_r > 0, \quad \Phi_\lambda < 0 \quad \text{and} \quad \Phi_q > 0$$

[3] Mathematical appendix D at the end of the book shows that if a random variable follows a Poisson process of parameter a, then the mathematical expectation of this variable is equal to $1/a$.

As relation (5.8) implies $\partial x/\partial i = -\Phi_i/\Phi_x$, $i = z, r, \lambda, q$, we immediately obtain the direction of the variations in the reservation wage as a function of the parameters of the model, or:

$$\frac{\partial x}{\partial z} > 0, \quad \frac{\partial x}{\partial \lambda} > 0, \quad \frac{\partial x}{\partial r} < 0 \quad \text{and} \quad \frac{\partial x}{\partial q} < 0 \tag{5.9}$$

With the help of relation (5.7), we deduce from this the main comparative statics properties of the average duration of unemployment. The result is:

$$\frac{\partial T_u}{\partial z} > 0, \quad \frac{\partial T_u}{\partial r} < 0 \quad \text{and} \quad \frac{\partial T_u}{\partial q} < 0$$

The rise in the reservation wage and the average duration of unemployment that follow from a rise in the net income z from looking for work, constitute an important result of this theory. This means, all other things being equal, that an increase in unemployment benefits should have the effect of lengthening the duration of unemployment. This result is highly intuitive: it simply makes sense that a job seeker receiving higher compensation will be more demanding in terms of the wage he hopes to get, and that that on average will lengthen the amount of time he spends looking. This strong prediction of the theory of job search has often been contested (see Atkinson and Micklewright, 1991, for a detailed critical analysis of it). On the theoretical level, however, it is unassailable, since the person looking for work does in fact receive benefit payments from the unemployment insurance system (which is the case in the basic model just presented). Let us suppose, for example, that unemployment benefits rise from 0% to 100% of the current average wage. It is hard to believe that a change of that magnitude in the size of the payments will have no positive influence on the average duration of unemployment. But leaving aside this exaggerated example, the extent of the influence is a priori unknown. Moreover, a very large percentage of those looking for work receive no unemployment benefits. We will see that, for them, an increase in unemployment benefits is highly likely to have an inverse effect on their reservation wage (a point rigorously established in section 2.2.1 of this chapter, which deals with the eligibility effect). Given these circumstances, we have to turn to empirical studies to get an idea of the sign and the order of magnitude of the unemployment benefits elasticity of the average duration of unemployment. We will see below that in general this elasticity is slight when the amount of unemployment benefits takes a "reasonable" magnitude.

The other implications of the model are also easy to grasp. A rise in r is characteristic of a job seeker who places less value on the future than another. A person of this type has a lower reservation wage and on average spends less time looking for work. When the job loss rate $q$increases, the current demands of job seekers diminish, since the gap between the expected utility of an employee and that of a job seeker shrinks, which reduces the average duration of unemployment. Another interpretation of this relation is that when jobs are of shorter duration, workers are less demanding because they know they will have other opportunities in the future. On the other hand, relation (5.7) shows that an increase in λ, the arrival rate of wage offers, has an ambiguous effect on the amount of time devoted to looking for a job. In this case, job seekers revise their reservation wage upward, which entails a lowering of the term $[1 - H(x)]$ representing the probability of accepting an offer. The direction of consequent change

in the rate of exit from unemployment $\lambda[(1 - H(x)]$ and the average duration of unemployment $T_u = 1/\lambda[(1 - H(x)]$ is then unknown. It should be noted, however, that if the frequency with which job offers arrive has little effect on the reservation wage, the average duration of unemployment decreases with this frequency. Empirical studies do seem to indicate as much (see section 3.2.2).

2.1.3 The Choice of Nonparticipation, Job Seeking, or Employment

Decisions to participate in the labor market are envisaged one way under the theory of labor supply and another way under the theory of job search. The theory of labor supply comprises only two possible states: either one is a participant or one is not. The theory of job search just outlined assumes that workers do participate in the labor market and are thus faced only with the choice between unemployment and employment. It is possible, though, to contemplate a hybrid model that takes into account three possible states: nonparticipation, job seeking, and employment.

The Reservation Wage and Alternative Income

In the theory of labor supply, participation in the labor market depends on a comparison between the current wage w and the reservation wage w_A defined by relation (1.3) in chapter 1. In this theory, decisions to participate can be summarized in the following manner:

$$\begin{cases} w > w_A \implies \text{employee} \\ w \leq w_A \implies \text{nonparticipant} \end{cases} \tag{5.10}$$

The theory of job search defines the reservation wage x as the wage at which the job seeker is indifferent between accepting a job and continuing to look. It depends on the overall characteristics of the labor market, which we will designate by Ω. According to equation (5.6) defining x, these characteristics include the distribution $H(.)$ of possible wages, the net income z associated with the job search, the job offers arrival rate λ, the interest rate r, and the job destruction rate q. Thus in symbolic terms we may write $\Omega = \Omega(H, z, q, \lambda, r)$ and $x = x(\Omega)$. The choice between participation and nonparticipation is based on a comparison between the expected utility of a job seeker V_u and that of a nonparticipant V_I. If the latter receives a constant income R_I at each date, her expected utility is defined by the equality $rV_I = R_I$. This can easily be compared to that of a job seeker, which is such that $rV_u = x$. An agent decides to participate in the labor market if and only if $V_I \leq V_u$, which translates into the inequality $x(\Omega) \geq R_I$. It is apparent that the decision to participate in the labor market is made by comparing the reservation wage to the "alternative income" R_I that a nonparticipant is capable of obtaining at any moment. Individual decisions hence take the following form:

$$\begin{cases} x(\Omega) \geq R_I \implies \text{participant} \\ x(\Omega) \leq R_I \implies \end{cases} \tag{5.11}$$

Moreover, when a participant receives a wage offer w, she accepts if it exceeds her reservation wage. In other words, the decisions of a participant come down to:

$$\begin{cases} w > x(\Omega) \implies \text{employee} \\ x(\Omega) \geq w > R_I \implies \text{unemployed} \end{cases} \tag{5.12}$$

The theory of job search suggests that the rate of participation depends on the set Ω of all the factors affecting the labor market. For example, some studies reveal that a rise in unemployment insurance benefits (an increase of z) is often accompanied by a rise in the participation rate, which itself takes the form of a rise in the unemployment rate (see Moorthy, 1989). In the same way, an increase in the unemployment rate, by lessening the probability of exiting from unemployment, tends to diminish the reservation wage and thus the participation rate. This relationship augments the procyclical character of the participation rates deduced from the labor supply model, in which the lowering of wages in bad economic times gives individuals incentive to withdraw from the labor market.

Discouraged Workers

The theory of job search only takes account of the wage prevailing in the marketplace through the distribution of its possible values. Hence, among nonparticipants, it is difficult to distinguish those who don't want to work at the "current" wage from those who would accept a job for that amount of remuneration but who give up looking because of the costs incurred by doing so and the time they would have to wait before being hired. These nonparticipants are called *discouraged workers*. If we assimilate the average of possible wages $\mathbb{E}(w) = \int_0^{+\infty} w \mathrm{d}H(w)$ to the "current" wage, we can conclude that individuals for whom $x(\Omega) \leq R_I \leq \mathbb{E}(w)$ form the category of discouraged workers. More generally, the "discouraged worker effect" is cited whenever change in certain characteristics of the economic environment implies a lowered participation rate. For example, if job offers arrive with reduced frequency, the reservation wage $x(\Omega)$ falls, and consequently the participation rate falls too (since the latter is by definition the percentage of the population for whom the relation $x(\Omega) \geq R_I$ is satisfied).

Numerous studies allow us to obtain an estimate of the number of discouraged workers. It suffices to identify, among the individuals who claim to be looking for work, those who have not made efforts that count as really "significant" (see OECD, 1994, volume 1 for a precise definition of this adjective). Table 5.3 shows that their number is not negligible.

The Frontier Between Nonparticipation and Job Seeking

The existence of discouraged workers suggests that the frontier between nonparticipation and participation in the labor force is difficult to draw. When does the intensity of

TABLE 5.3
Discouraged workers and job seekers in 2011 (as a percentage of the labor force).

Country	Discouraged workers	Job seekers
Denmark	0.15	7.6
Spain	1.33	21.6
France	0.12	9.3
Germany	0.14	5.9
United States	0.65	8.9
Japan	1.04	4.5

Source: OECD Labor Force Statistics.

the effort made by an individual to find a job qualify him as an active job seeker? The varying definitions of unemployment supply different, and perforce arbitrary, answers to this question. Measurements of unemployment derive from investigations in which, to be considered unemployed, you have to have been without work (during the period in question), have taken steps to look for work, and be ready to start work (in principle) immediately. But these three conditions, in particular the second pertaining to the process of looking, can have different meanings. Thus in the United States individuals who employ passive methods (like looking in the want ads) are classed as nonparticipants, while numerous OECD countries consider job seekers employing both passive and active methods as unemployed (see U.S. Bureau of Labor Statistics, www.bls.gov/cps/cps_htgm.htm).

A number of factors point to the conclusion that the distinction between nonparticipation and unemployment often turns out to be arbitrary. Re-interview programs carried out in the United States with individuals already interviewed the week before reveal that, especially for individuals situated close to the frontier of nonparticipation, the answers given (regarding the same period of reference) can be quite different (Abowd and Zellner, 1985). Some people are hard to classify, and their answers are highly sensitive to the way the interviews are conducted. Jones and Riddell (1999) show that individuals classed as nonparticipants by surveys of the labor force in Canada are anything but uniform in their behavior. These authors distinguish four categories of individuals: the employed, the unemployed, individuals marginally attached to labor market participation, and nonparticipants. Individuals marginally attached to labor market participation, traditionally considered nonparticipants, say that they are not looking for a job but would like to work. They represent 25% to 30% of the volume of unemployment over the period studied by Jones and Riddell. The matrix of transition between different states is presented in table 5.4. It is apparent that individuals marginally attached to labor market participation behave differently on average than nonparticipants, since they have a much higher probability of returning to full participation. The rates at which individuals on the margin of participation do return to employment are closer to those of the unemployed than to those of genuine nonparticipants. Jones and Riddell also emphasize that the category of individuals marginally attached to participation is extremely heterogeneous. Consequently, within the overall group of those who say they would like to work but are not looking for a job, Jones and Riddell distinguish persons who are "waiting" for a job—because they are "waiting to be recalled by their former

TABLE 5.4

The transition matrix between different states in the labor market. Monthly rates for the year 1992 in Canada (standard errors are in parentheses).

From To → ↓	Employed	Unemployed	Nonparticipant + marginally attached
Unemployed	0.112 (0.004)	0.708 (0.005)	0.180 (0.005)
Marginally attached	0.098 (0.005)	0.171 (0.007)	0.731 (0.008)
Nonparticipant	0.026 (0.001)	0.030 (0.001)	0.944 (0.002)

Source: Jones and Riddell (1999, table 1).

employer," or "have found a job but haven't been hired yet," or "are waiting for an answer from an employer"—and discouraged persons who "believe there are no jobs matching their qualifications available in their region." It is apparent that those who are waiting for a job have a rate of return to employment higher than that of the unemployed (equal to 0.200), whereas discouraged workers show behavior closer to that of genuine nonparticipants (their rate of return equals 0.044).

These examples show that taking job search behavior into consideration renders the distinction between labor market participation and nonparticipation ambiguous. In consequence, assessments of unemployment and of the labor force are necessarily arbitrary, and it is generally useful to supplement them with other indicators in order to get a clear picture of the state of the labor markets. In this regard, the employment rate—equal to the ratio between the number of jobs and the population of working age, generally taken to be all those between 15 and 64 years of age—is a supplementary indicator frequently used to gauge what is happening in the labor markets.

2.2 Extensions of the Basic Model

The results obtained using the basic model are numerous. They have however been obtained using hypotheses that are sometimes very restrictive. To expand on what the basic model has to tell us, we first examine the consequences of the conditions of eligibility for unemployment insurance benefits. We then look at the changes we must make when an individual is able to look for a job while he or she is already working. After that we make the assumption that agents can decide how much effort to put into their job search. And finally we study the consequences of the fact that unemployment insurance benefits are not stationary.

2.2.1 Eligibility and Unemployment

In most countries, those who work in exchange for wages have to pay premiums into an unemployment insurance system that allows a wage earner to receive compensation if she loses her job. When these conditions are met, we say that the worker is *eligible* for unemployment insurance benefits. But many people, in particular new entrants into the labor market and those who have been unemployed for a long time, are not eligible for such benefits. For them, finding a job also means becoming eligible, or becoming eligible again. This entails that the reservation wage of those who are not eligible *sinks* when the benefits paid to the unemployed *who do meet the eligibility requirements* rise.

Two Types of Job Seeker

To make this intuition perfectly explicit, we will assume in what follows that there are two types of job seekers: those who are eligible for unemployment insurance benefits and those who are not. This circumstance can be formalized quite simply by assuming on one hand that the instantaneous income of the former always amounts to z, while that of the latter has the value $z_n < z$, and on the other that an individual becomes and remains eligible if she has been employed at least once. In this context, z represents the benefits paid by the unemployment insurance system, while z_n is determined by the welfare system, which generally pays out smaller amounts.

The situation of the eligible job seeker is identical to that of the basic model, and her reservation wage, always denoted by x, continues to be defined by equation (5.6). But the behavior of a noneligible job seeker is not so simple because her expected utility, denoted V_{un}, depends on that of an eligible job seeker, which continues to be denoted V_u. When a noneligible job seeker accepts a job offering an instantaneous wage w, her expected utility $V_e(w)$ satisfies the following equation:

$$rV_e(w) = w + q[V_u - V_e(w)] \tag{5.13}$$

It should be noted that it is the expected utility V_u of an *eligible* job seeker that appears in this expression, for it is assumed, for the sake of simplicity, that unemployment insurance benefits are paid whenever an agent has been employed at least once. For given V_u, relation (5.13) indicates that $V_e(w)$ increases with w, and that the reservation wage of a noneligible job seeker, denoted x_n, satisfies the equality $V_e(x_n) = V_{un}$. Since we always have $x = rV_u$, equation (5.13) allows us to express the expected utility of a noneligible job seeker as a function of the two reservation wages, x and x_n. The result is:

$$rV_{un} = \frac{rx_n + qx}{r + q} \tag{5.14}$$

Assuming that the frequency with which a noneligible job seeker receives job offers is always equal to λ, her expected utility is defined by the following equation, which is analogous to relation (5.5) in the basic model:

$$rV_{un} = z_n + \lambda \int_{x_n}^{+\infty} [V_e(w) - V_{un}]\, \mathrm{d}H(w) \tag{5.15}$$

The Reservation Wage of Noneligible Job Seekers

Observing, from (5.13), that $rV_e(w) = (rw + qx)/(r + q)$, and utilizing expression (5.14) of V_{un}, we arrive, thanks to (5.15) and after several simple calculations, at a relation that implicitly defines the reservation wage x_n of a noneligible person as a function of that of an eligible person. It is written:

$$rx_n = (r + q)z_n - qx + \lambda \int_{x_n}^{+\infty} (w - x_n)\mathrm{d}H(w)$$

It is easy to verify that this relation implies a negative linkage between x_n and x. Since x increases with the instantaneous income z of eligible job seekers, the reservation wage x_n of *noneligible* job seekers is a *decreasing* function of z. This outcome is explainable as follows: a noneligible job seeker knows that by accepting an offer of work, he risks becoming unemployed again in the future at rate q. But in that case, he also knows that he will henceforth be eligible for unemployment benefits $z > z_n$. A rise in z therefore increases the loss occasioned by refusing a job offer, which gives him incentive to lower his reservation wage. On the other hand, we may note that an increase in welfare payments z_n exerts upward pressure on the reservation wage of noneligible job seekers. This implies that a rise in unemployment benefits has an ambiguous impact on unemployment because it increases the unemployment spell of eligible job seekers but it decreases the unemployment spell of those who are not eligible.

2.2.2 On-the-Job Search

As a general rule, an individual who has a job is still able to carry out a search for another one. For the sake of simplicity, we will assume that the costs of job search are negligible for a worker who is employed. The advantage of this hypothesis is that we do not have to make a distinction between employees who have a low wage and are looking for another job and those who are receiving a high wage and therefore are not looking, since the cost of doing so would be too high compared to their earnings prospects. If the costs of searching for a job are null for an employed worker, she always has an interest in looking for another job, and accepts the first offer that exceeds her present wage.

The Behavior of Agents

Let us assume that an employed person receives job offers with a frequency of λ_e, and that she risks losing her job, at any time, with an exogenous constant probability of q. The discounted utility $V_e(w)$ expected by a wage earner whose current remuneration comes to w then has three components. The first corresponds to the instantaneous income w deriving from her waged labor, the second is the average discounted expected gain $q[V_u - V_e(w)]$ due to job loss, and the third is the discounted expected earnings $\lambda_e \int_w^{+\infty} [V_e(\xi) - V_e(w)]\mathrm{d}H(\xi)$ consequent upon a change of employer (which occurs for every wage offer that exceeds the present wage w). Finally, $V_e(w)$ is defined by the following equation:[4]

$$rV_e(w) = w + q[V_u - V_e(w)] + \lambda_e \int_w^{+\infty} [V_e(\xi) - V_e(w)]\,\mathrm{d}H(\xi) \tag{5.16}$$

Deriving this relation with respect to w, we get:

$$V_e'(w) = \frac{1}{r + q + \lambda_e[1 - H(w)]} \tag{5.17}$$

In this way we easily verify that the discounted expected utility $V_e(w)$ of an employee increases with wage w; hence the optimal search strategy for a job seeker is characterized by a reservation wage x such that $V_e(x) = V_u$. Assuming that the arrival rate of job offers is equal to λ_u for a job seeker, and again designating her instantaneous gain by z, her discounted expected utility V_u continues to be defined by equation (5.5), so that:

$$rV_u = z + \lambda_u \int_x^{+\infty} [V_e(\xi) - V_u]\,\mathrm{d}H(\xi) \tag{5.18}$$

Making $w = x$ in (5.16) and comparing (5.18), we immediately get:

$$x = z + (\lambda_u - \lambda_e) \int_x^{+\infty} [V_e(\xi) - V_u]\,\mathrm{d}H(\xi) \tag{5.19}$$

[4] The reader who is not yet sufficiently familiar with this type of equation will benefit from working with a small interval of time $[t, t+\mathrm{d}t]$. In the stationary state, we thus have:

$$(1 + r\mathrm{d}t)V_e(w) = w\mathrm{d}t + q\mathrm{d}tV_u + (1 - q\mathrm{d}t)\left[\lambda_e \mathrm{d}t \int_w^{+\infty} V_e(\xi)\mathrm{d}H(\xi) + \lambda_e \mathrm{d}tV_e(w)H(w) + (1 - \lambda_e \mathrm{d}t)V_e(w)\right]$$

By rearranging a few terms and making $\mathrm{d}t \to 0$ in this formula, we come back to equation (5.16).

Compared to the basic model, this equation indicates that a job seeker must henceforth weight the discounted expected utility of the job search $\int_x^{+\infty}[V_e(\xi) - V_u]\mathrm{d}H(\xi)$ by the *difference* $\lambda_e - \lambda_u$ of the rates with which job offers arrive.

Properties of the Reservation Wage

We will see further in this chapter (section 4.2) that the possibility of moving from one job to another plays an essential role in the elaboration of *equilibrium* search models, that is, models in which the cumulative distribution function $H(.)$ is endogenous. In this regard, it is useful to determine precisely the expression of $V_e(\xi) - V_u$ appearing in (5.19) so as to bring out the dependence between the reservation wage x and the function $H(.)$. By applying the formula of integration by parts[5] to the right-hand side of (5.19), we arrive at:

$$x = z + (\lambda_u - \lambda_e)\left[\left[-\bar{H}(\xi)\,[V_e(\xi) - V_u]\right]_x^{\infty} + \int_x^{+\infty}\bar{H}(\xi)\,V_e'(\xi)\mathrm{d}\xi\right] \quad \text{with} \quad \bar{H}(\xi) \equiv 1 - H(\xi)$$

As we still have $V_e(w) - V_u = \int_x^w V_e'(\xi)\mathrm{d}\xi$, utilizing (5.17) and assuming that $\lim_{\xi\to\infty}\bar{H}(\xi)\,[V_e(\xi) - V_u] = 0$, we finally have:

$$x = z + (\lambda_u - \lambda_e)\int_x^{+\infty}\frac{\bar{H}(\xi)}{r + q + \lambda_e\bar{H}(\xi)}\mathrm{d}\xi \qquad \textbf{(5.20)}$$

This equation implicitly defines the reservation wage as a function of the parameters λ_u, λ_e and the cumulative distribution function $H(.)$. When $\lambda_e = 0$, that is, when there is no on-the-job search, we come back to the reservation wage of the basic model. Vice versa, if $\lambda_e > 0$, the job seeker takes account of the possibilities of future income associated with continuing to look for a job while employed. Adopting this stance has the effect of lowering the reservation wage. If $\lambda_e = \lambda_u$, the reservation wage is equal to the net income z of the job seeker, for a worker then has as many chances of receiving an acceptable offer while employed as he does while unemployed. It is also interesting to note that if $\lambda_e > \lambda_u$, the reservation wage falls *below* z. In this configuration of the parameters, an employee has more chances of obtaining an acceptable offer than a job seeker. The latter thus has an incentive to accept "bad" jobs, which nevertheless afford him better prospects than his present situation of being unemployed. The bulk of the estimations show however that the inequality $\lambda_u \geq \lambda_e$ is the most probable. For example, using data from the Netherlands, van den Berg and Ridder (1998) find that λ_u differs very little from λ_e, while Bontemps (1998) and Kiefer and Neumann (1993) estimate, using French and American data, that λ_u is respectively 10 times and 5 times higher than λ_e. This likely comes about because unemployed job seekers devote more effort to looking for work than employed job seekers do. Be that as it may, taking into account on-the-job search ($\lambda_e > 0$) has the effect of diminishing the size of the reservation wage in comparison to the one that emerges from the basic model ($\lambda_e = 0$).

[5] This formula reads $\int u\mathrm{d}v = uv - \int v\mathrm{d}u$, where u and v are two functions. Here, we posit: $u = V_e(\xi) - V_u$, $\mathrm{d}u = V_e'(\xi)\mathrm{d}\xi$, $\mathrm{d}v = h(\xi)\mathrm{d}\xi$, and $v = -\bar{H}(\xi)$.

2.2.3 Choosing How Hard to Look

The hypothesis that both the arrival rate of job offers and the costs of the job search do not vary is unsatisfactory, since it does not allow us to take into account the fact that a job seeker may make sedulous efforts that increase the costs of the job search but at the same time increase her chances of receiving job offers. This relation is well documented by the empirical research cited in the first section of this chapter, which shows that persons who dedicate more time to looking for work exit more rapidly from unemployment.

Optimal Effort

Let us designate the intensity of the job search by the scalar e, which can be interpreted as the time and/or the intensity of the effort devoted to search. The notion that more job offers should result from greater effort devoted to search amounts to postulating that the rate at which offers arrive increases with e. For the sake of simplicity and without loss of generality, we postulate a linear relation $\lambda = \alpha e$. The parameter $\alpha > 0$ we interpret as an indicator of the state of the labor market, independent of individual efforts. This parameter is a function of, among other things, the number of vacant jobs, the number of job seekers, and objective characteristics like age, sex, and educational level. We will denote by $\phi(e)$ the cost arising from the search effort e, with $\phi' > 0$ and $\phi'' > 0$. So henceforth the instantaneous utility of a job seeker will be written $[z - \phi(e)]$. For ease of exposition, we also assume that there is no on-the-job search and for that matter the opposite assumption would change the outcome very little (see Mortensen, 1986). Thus we can follow exactly the line of reasoning worked out in the basic model in section 2.1, positing in the first stage that the amount of effort e is given.

The reservation wage x is always implicitly defined by the equation (5.6), which will henceforth be written:

$$x = z - \phi(e) + \frac{\alpha e}{r+q}\int_x^{+\infty}(w-x)\mathrm{d}H(w) \tag{5.21}$$

This relation gives the value of the reservation wage associated with a given amount of effort e. Now the optimal value of effort ought, by definition, to maximize the intertemporal utility V_u of a job seeker. Since $V_u = x/r$, this value is reached by differentiating relation (5.21) with respect to e and looking for the value of e for which $\partial x/\partial e = 0$. The result is:

$$\phi'(e) = \alpha\int_x^{+\infty}\frac{w-x}{r+q}\mathrm{d}H(w) \tag{5.22}$$

The convexity of function $\phi(.)$ guarantees that the amount of effort defined by this relation is indeed a maximum. This equation states that it is optimal to equalize the marginal cost of effort to its marginal return. The latter is equal to α (the increased probability of getting a job offer) times the expected gains associated with a job offer.

With the help of (5.21), we further obtain:

$$x = z + e\phi'(e) - \phi(e) \tag{5.23}$$

The Properties of Optimal Effort

In what follows, it will be helpful to view equations (5.22) and (5.23) as forming a system determining in an implicit manner the reservation wage and the optimal effort, respectively written $x(\alpha, z)$ and $e(\alpha, z)$. By differentiating relation (5.23) with respect to α, it is easy to show that $\partial x(\alpha, z)/\partial\alpha$ and $\partial e(\alpha, z)/\partial\alpha$ are of the same sign.[6] With the help of this property, differentiating equation (5.22) with respect to α implies:[7]

$$\frac{\partial x(\alpha, z)}{\partial \alpha} > 0 \quad \text{and} \quad \frac{\partial e(\alpha, z)}{\partial \alpha} > 0$$

We knew already that an improvement in the state of the labor market causes the reservation wage to rise—see (5.9)—and it is apparent that it also increases the intensity of the job search. In other words, when the economy is going well, or when it is easier to find a job, it pays a job seeker to look harder, which also allows him to raise his wage demands. Conversely, when the economy slows, a job seeker both lowers his reservation wage and reduces his search efforts (see also van den Berg and van Ours, 1994).

Differentiating relation (5.22) with respect to z, we deduce that $\partial x(\alpha, z)/\partial z$ and $\partial e(\alpha, z)/\partial z$ are of opposed signs. Using this result, differentiating relation (5.23) with respect to z further implies:

$$\frac{\partial x(\alpha, z)}{\partial z} > 0 \quad \text{and} \quad \frac{\partial e(\alpha, z)}{\partial z} < 0$$

Thus, as in the basic model, a rise in the income of a job seeker raises the reservation wage—see further (5.9)—but we also observe that such a rise tends to reduce the search effort. This results from the fact that an increase in z increases the intertemporal utility of the job seeker. He can thus reduce the amount of effort he puts into searching, because the marginal gain from intensified effort sinks below the level of marginal disutility that it provokes. Therefore, increases in z unambiguously increase unemployment spells, equal to $1/\alpha e[1 - H(x)]$. Finally, it should be noted that a *simultaneous* lowering of α and z has an ambiguous effect on optimal effort. It can indeed happen that certain categories of persons (the long-term unemployed in particular) find themselves facing a reduced number of job offers and a reduction in their unemployment benefits.

2.2.4 Job Search and Wealth

To this point we have neglected risk aversion and the possibility of saving or taking on debt. In actuality, individuals generally have an aversion to risk that may lead them to save when they are in work so as to lessen the impact of the drop in income that would result should they become unemployed. They may also borrow, if borrowing is possible, when they are unemployed. Such behavior modifies the expectation of utility during a

[6] Differentiating (5.23) with respect to α gives $\frac{\partial x}{\partial \alpha} = \frac{\partial e}{\partial \alpha} e\phi''(e)$.

[7] Differentiating equation (5.22) with respect to α yields:

$$\frac{\partial e}{\partial \alpha}\phi''(e) + \alpha\frac{\partial x}{\partial \alpha}\frac{1 - H(x)}{r + q} = \int_x^{+\infty} \frac{w - x}{r + q} \mathrm{d}H(w) > 0$$

spell of unemployment, since unemployed persons have an interest in dipping into their savings, and even in taking on debt, and so reducing their wealth. These modifications in the behavior of the unemployed are analyzed by Danforth (1979), Lentz and Tranaes (2005), Chetty (2008), and Lammers (2012).

To simplify, we approach this problem with a static model where the unemployed person disposes of a given initial wealth, denoted a. The utility of a job seeker is written $v(c) - \phi(e)$, where $v(c)$ is an increasing concave function corresponding to the utility of consumption c and $\phi(e)$ is an increasing convex function that represents the disutility of the search effort e. The probability of receiving an offer is equal to αe. We assume that the utility of an agent remunerated at wage w is written $v(a + w) - \psi$, where ψ designates the disutility of work. If this agent does not find work, her utility is simply equal to $v(a + z)$ where z designates the income obtained from sources other than work. At the beginning of the period, the unemployed person effects a search effort e that allows her to receive wage offers picked from a distribution $H(.)$. At the close of the period, the unemployed person works if she has received and accepted an offer and consumes all her wealth and income. In this setting, an unemployed person chooses her reservation wage x and her search effort e so as to maximize her expected utility, or:

$$\max_{(e,x)} -\phi(e) + (1 - \alpha e)\, v(a + z) + \alpha e \left\{ \int_x^{+\infty} [v\,(a + w) - \psi]\, \mathrm{d}H(w) + v(a + z)H(x) \right\}$$

The first-order conditions of this problem define the optimal effort and the reservation wage:

$$\phi'(e) = \alpha \int_x^{\infty} [v\,(a + w) - \psi - v(a + z)]\, \mathrm{d}H(w) \tag{5.24}$$

$$v(a + x) = v(a + z) + \psi \tag{5.25}$$

We see that her reservation wage depends on her wealth and that it is greater than z. Differentiating the equation (5.25) with respect to a, we get:

$$\frac{\mathrm{d}x}{\mathrm{d}a} = \frac{v'(a + z) - v'(a + x)}{v'(a + x)} > 0$$

An increase in wealth thus raises the reservation wage. Deriving and differentiating equation (5.24) yields:

$$\phi''(e)\frac{\mathrm{d}e}{\mathrm{d}a} = \alpha \int_x^{+\infty} \left[v'\,(a + w) - v'(a + z)\right] \mathrm{d}H(w)$$

The utility function v being concave, $v'\,(a + w) < v'(a + z)$ when $w > z$. Since for that matter the effort function is convex, we have $\phi''(e) > 0$, which implies that search effort decreases with wealth a. This result illustrates the intuitive view that wealthier persons have less incentive to look for work, since the marginal utility of their consumption is less.

Starting from the observation that wealth undergoes diminution with the duration of unemployment, this model suggests that the reservation wage ought to fall during a

spell of unemployment while the search effort ought to rise. Note, however, that the results of this static model are not general. Lentz and Tranaes (2005) show, in a dynamic setting, that the sign of the impact of wealth on the reservation wage and on search effort is a priori undetermined. It depends on the form of the utility function, the form of the search cost function, and the existence of liquidity constraints.

2.2.5 The Effect of Benefit Sanctions

In general, the unemployed must meet obligations in order to receive unemployment benefits. Primarily, they must take verifiable steps to look for work, and their payments may be temporarily or permanently suspended if they fail to meet their obligations or refuse a job that is offered to them. The fulfillment of these obligations is monitored with widely varying strictness across OECD countries (figure 5.2). In addition to such steps, unemployed persons may be subjected to a requirement of availability for work during participation in an active program or to demands for occupational or geographical mobility. Sanctions can also be imposed in case of refusal to participate in an active labor market program such as training or intensive placement services. The strictness of the sanctions imposed in cases of noncompliance has been summarized by Venn (2012) in a synthetic index. Figure 5.3 shows that even though sanctions are applied in principle in most countries, there are large differences in their degree of strictness.

The model of job search with endogenous effort is well adapted to analyzing the consequences of sanctions (see Lalive et al., 2005; Abbring et al., 2005; Boone and van Ours, 2006; Boone et al., 2007, 2009). The possibility of being sanctioned may be formalized with the assumption that the probability of being sanctioned decreases with search effort. The sanction consists of a *permanent* reduction in unemployment benefit

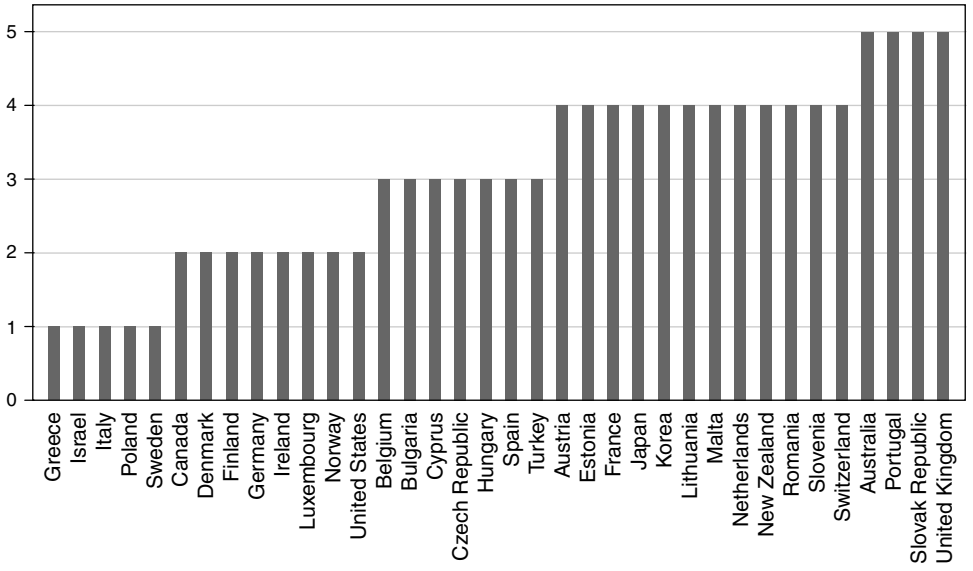


Figure 5.2
Strictness of job search monitoring, scored from 1 (least strict) to 5 (most strict).

Source: Venn (2012, figure 4, p. 18).

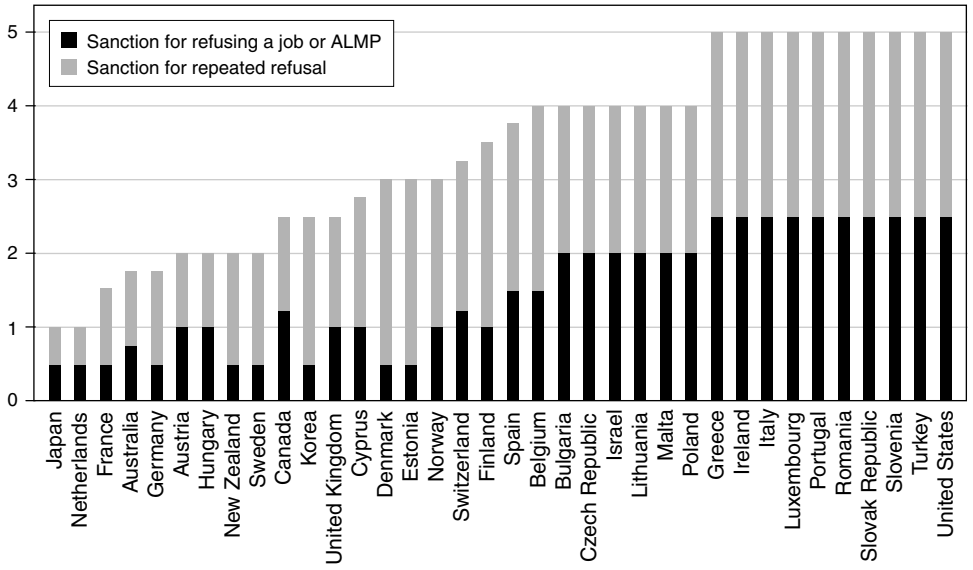


Figure 5.3
Strictness of sanctions, scored from 1 (least strict) to 5 (most strict).

Source: Venn (2012).

by an amount s. This reduction occurs at a rate of $\sigma(e)$, where σ is a decreasing and convex function of search effort e. The search effort made by a job seeker not subject to sanction we denote by e_u and that of a sanctioned job seeker e_s. The corresponding probability to receive a job offer is respectively αe_u and αe_s. To simplify the analysis, we will assume that all jobs have the same wage w and that the rate of job destruction is null, such that the value of a job, V_e, is simply equal to w/r. So that there may exist unemployed persons not subject to sanctions at stationary equilibrium, we will assume that all agents have an instantaneous death rate equal to δ and that there is a number of entrants equal to the number of the deceased at each instant. All entrants onto the labor market start out as unsanctioned unemployed. We denote by N the size, constant and exogenous, of the active population. U and S designate respectively the number of unsanctioned unemployed at stationary equilibrium and the number of sanctioned unemployed. At stationary equilibrium, the equality of the flows entering and exiting unsanctioned and sanctioned unemployment entails:

$$\delta N = \delta U + \sigma(e_u)U + \alpha e_u U \quad \text{and} \quad \sigma(e_u)U = \delta S + \alpha e_s S$$

whence:

$$U = \frac{\delta}{\delta + \alpha e_u + \sigma(e_u)} N \quad \text{and} \quad S = \frac{\sigma(e_u)}{\delta + \alpha e_s} \cdot \frac{\delta}{\delta + \alpha e_u + \sigma(e_u)} N$$

The last two equations yield the values of U and S as a function of search efforts e_u and e_s. From that we deduce the number of jobs held, denoted L, with the identity $L = N - U - S$.

In order to know the search efforts of agents, we have to specify their behavior. If ρ designates the discount rate, every agent is in fact subject to an effective discount rate $r = \rho + \delta$, given that she might die at any instant with a probability δ. Under these conditions, at stationary equilibrium, the expected utility of an unsanctioned unemployed person, denoted V_u, and that of a sanctioned unemployed person, denoted V_s, verify the two following equations:

$$rV_u = z - \phi(e_u) + \alpha e_u (V_e - V_u) + \sigma(e_u)(V_s - V_u) \tag{5.26}$$

$$rV_s = z - s - \phi(e_s) + \alpha e_s (V_e - V_s) \tag{5.27}$$

The unemployed choose levels of search effort that maximize their respective expected utilities. The first-order conditions are then written as follows:

$$\phi'(e_u) = \alpha (V_e - V_u) + \sigma'(e_u)(V_s - V_u) \tag{5.28}$$

$$\phi'(e_s) = \alpha (V_e - V_s) \tag{5.29}$$

The convexity of functions ϕ and σ defining the cost of search effort and the probability of being sanctioned ensures that the second-order conditions are satisfied.

These equations allow us to study the impact of the extent s of the sanction on search efforts. Since $V_e = w/r$ is independent of s, and taking into account the first-order condition (5.29), the derivation of (5.27) with respect to s gives $(r + \alpha e_s)\frac{\mathrm{d}V_s}{\mathrm{d}s} = -1$, and consequently we have $\frac{\mathrm{d}V_s}{\mathrm{d}s} < 0$. The function ϕ being convex, equation (5.29) thus shows that the search effort of the sanctioned unemployed increases with the sanction. This is what is called the ex-post effect of the sanction: when the sanction is applied, search effort increases on account of the reduction of unemployment benefit. But the sanction also has an ex-ante effect: the threat of its imposition modifies the behavior of the unemployed who are as yet unsanctioned. Thus, taking into account the first-order condition (5.28), the derivation of (5.26) with respect to s yields $\frac{\mathrm{d}V_u}{\mathrm{d}s} = \frac{\sigma(e_u)}{r+\sigma(e_u)+\alpha e_u}\frac{\mathrm{d}V_s}{\mathrm{d}s} < 0$. With the help of this last relation, the derivation of the first-order condition (5.28) entails, after several simple calculations:

$$\left[\phi''(e_u) + \sigma''(e_u)(V_u - V_s)\right]\frac{\mathrm{d}e_u}{\mathrm{d}s} = \frac{(r + e_u)\sigma'(e_u) - \alpha\sigma(e_u)}{\sigma(e_u)}\frac{\mathrm{d}V_u}{\mathrm{d}s}$$

The function σ being decreasing, we thus have $\frac{\mathrm{d}e_u}{\mathrm{d}s} > 0$. Note that the search effort of the unsanctioned unemployed increases with the amount of the sanction. The unemployed not (yet) subject to sanction increase their search effort to avoid the sanction, the more so to the degree the sanction is severe. In this perspective, it is interesting to compare the search effort of the sanctioned and unsanctioned unemployed. Rewriting the term $(V_e - V_s)$ in the form $(V_e - V_u) + (V_u - V_s)$ in equation (5.29), we find that:

$$\phi'(e_u) = \phi'(e_s) - \left[\alpha + \sigma'(e_u)\right](V_u - V_s)$$

As $V_u > V_s$, we observe that the search effort of the unsanctioned unemployed exceeds that of the sanctioned unemployed if (and only if) $\sigma'(e_u) < -\alpha$. Hence, when

the probability of being sanctioned rises sufficiently in response to reduced search effort, the threat of being sanctioned pushes the search effort of the unsanctioned unemployed higher than that of the sanctioned unemployed. Boone et al. (2009) observe this phenomenon in an experimental setting. The threat of the sanction may be more effective than the application of the sanction. This result is important, for it suggests that a well-managed system of sanctions may be an effective tool for giving the unemployed an incentive to search hard for a job while paying them a generous unemployment benefit (Boone et al., 2007). This problem will be analyzed in greater detail in chapter 14.

2.2.6 Nonstationary Environment

The hypothesis that a job seeker's environment is stationary does not apply in a number of cases. Financial constraints increase the longer unemployment lasts, job offers most often grow scarcer, and net income from the search falls off, since as a general rule unemployment insurance systems mandate a reduction, or even a termination, in the payment of benefits at the end of a certain period. In what follows, we focus only on this last cause of nonstationarity; van den Berg (1990) presents a model taking into account a number of causes of nonstationarity. More precisely, we assume that the net instantaneous income of a job seeker diminishes (in the broad sense) over time. We will thus have $z(t) \leq z(t')$ for all $t \geq t'$.

In this nonstationary environment, the discounted expected utility of a person entering unemployment, or $V_u(0)$, is no longer necessarily equal to the discounted expected utility $V_u(t)$ of a person who has already been unemployed for a period $t > 0$. We do however continue to assume that a job offer is a proposal of a constant wage which an employee will receive as long as he remains with the firm that makes the proposal. Thus, the discounted expected utility $V_e(w)$ of a person paid a constant wage w is stationary. Assuming for simplicity that there is no on-the-job search, it is defined by the following equation:

$$rV_e(w) = w + q\,[V_u(0) - V_e(w)] \tag{5.30}$$

The optimal job search strategy still consists of refusing all proposals that offer an expected utility less than that of an unemployed person and accepting all others. Since, following relation (5.30), $V_e(w)$ is an increasing function of w, the optimal strategy comes down to choosing, at every moment, a reservation wage such that only offers that exceed it will be accepted. Let us denote by $x(t)$ the reservation wage of a person whose duration of unemployment is equal to t; this wage is then characterized by the equality $V_e[x(t)] = V_u(t)$. Since the function $V_e(.)$ is increasing, the reservation wage $x(t)$ varies in the same direction as the discounted expected utility $V_u(t)$. Now intuition suggests that $V_u(t)$ ought to decrease with the duration t of unemployment, inasmuch as the resources $z(t)$ of a job seeker diminish with this duration. To see this clearly, we may focus on a short interval of time $[t, t+\mathrm{d}t]$ and make explicit the trade-off equation giving the value of $V_u(t)$. If λ continues to designate the rate at which job offers arrive, we then have:

$$V_u(t) = \max_s \frac{z(t)\mathrm{d}t + \lambda \mathrm{d}t \left[\int_s^{+\infty} V_e(w)\mathrm{d}H(w) + V_u(t+\mathrm{d}t)H(s)\right] + (1 - \lambda \mathrm{d}t)V_u(t+\mathrm{d}t)}{1 + r\mathrm{d}t} \tag{5.31}$$

In the maximization problem appearing in this equation, the discounted expected utility $V_u(t + dt)$ at date $(t + dt)$ has to be considered as given, for on that date the job seeker decides on a *new* reservation wage independently of the choice made at date t. The optimal reservation wage is then obtained by setting to zero the derivative with respect to s of the term between brackets in the expression (5.31) of $V_u(t)$. After several simple calculations, we arrive at $V_e[x(t)] = V_u(t+dt)$, which corresponds exactly to the characterization $V_e[x(t)] = V_u(t)$ of the reservation wage $x(t)$ when $dt \to 0$.[8]

Since the net income $z(t)$ of an unemployed person decreases over time, equation (5.31) shows that $V_u(t) \leq V_u(t')$ necessarily obtains for every $t \geq t'$. Since his reservation wage and discounted expected utility vary in the same direction, we can deduce that $x(t) \leq x(t')$ for every $t \geq t'$. Hence reservation wages fall with time spent searching for a job when unemployment insurance benefits are regressive. This result implies that the rate of leaving unemployment, or $\lambda[1 - H(x(t))]$, increases with the duration t of the unemployment spell—a conclusion confirmed by a number of observations, in particular concerning the behavior of certain categories of job seekers as the period of their entitlement to unemployment insurance benefits draws to a close (see section 4.1.2). On the other hand, the long-term unemployed have, in general, a smaller probability of exiting from unemployment than do the short-term unemployed. This phenomenon can be explained by the fact that job offers arrive less frequently the longer one is unemployed, either because the productive abilities of the individual decline or simply because employers take the view that too long a period of unemployment sends a bad "signal." In these circumstances, the fact that one's reservation wage has fallen may be offset, or more than offset, by the declining arrival rate of job offers. The rate of exit from unemployment is then no longer obliged to decrease with the duration of the job search.

The foregoing analysis can easily be applied to the case of a change in the length of time unemployment insurance benefits are paid.[9] For example, if this period is shortened, that means that the intertemporal resources of the job seeker shrink, and that diminishes both his discounted expected utility and reservation wage. Thus, for a period of unemployment of the *same length*, and for the *same amount* of benefits, a shortened period of entitlement to benefits leads to a lowering of the reservation wage and consequently a reduction in the average duration of unemployment.

3 EMPIRICAL ASPECTS OF JOB SEARCH

The job search model contains a number of predictions that it is important to test and quantify in order to have at our disposal the kind of information public policy makers require.

[8] One can check as well that the second derivative with respect to s of the term between brackets is negative when this equality is satisfied. So what we have is indeed a maximum.

[9] By way of illustration, the interested reader can characterize the reservation wages associated with a system of unemployment insurance benefit such that $z(t) = z_0$ for $0 \leq t \leq T$, and $z(t) = z < z_0$ for $t > T$, where z, z_0, and T are constant exogenous parameters. A reduction in the length of time over which benefits are paid is similar to a lowering of T.

In this perspective, the economist is faced with the classic problem of detecting causal relations. A demonstrated correlation between the generosity of unemployment insurance benefit and the duration of unemployment is not a sufficient basis for inferring a cause-and-effect sequence. Such an observed correlation might flow from a combination of two effects. The first is the actual impact on behavior of the insurance mechanism, precisely the causal phenomenon we wish to isolate. The second is the effect flowing from differences among job seekers, such as age, previous labor market experience, place of residence, date of registration as unemployed, individual motivation, life goals, and so on. Given that certain characteristics are not observed by the economist (motivation for example), it is illegitimate, even if observable characteristics are controlled for, to conclude from a correlation between the generosity of unemployment benefit and the duration of unemployment that the one caused the other.

This is a point of primordial importance. Empirical research that merely presents correlations without any convincing strategy for isolating a causal relation—what is called an "identification strategy"—must be regarded as descriptive in nature and inadequate to ground an inference of causality. Yet such research is not without value, for it may suggest approaches that can help us, where necessary, to pinpoint causal relations.

Hence empirical research on the impact of unemployment insurance on the behavior of the unemployed is implicitly focused on isolating relations of cause and effect. We will present this research with the aid of the contribution of Lalive et al. (2006), which assesses the impact of modifications in the Austrian unemployment insurance system on the duration of unemployment. We start with the identification strategy chosen by these authors, setting it in the context of the range of strategies adopted in this domain. We then proceed to the estimation properly speaking, and the empirical results. The main results of this contribution can be replicated with data and programs available at www.labor-economics.org.

3.1 The Identification Strategy

3.1.1 Controlled Experiments

Ideally, the evaluation of the impact of a measure forming part of an unemployment insurance system ought to rest on controlled experiments in which its beneficiaries would be drawn at random from the potentially eligible population, resulting in the division of the eligible population into a "test group" (or "treated group") of beneficiaries and a "control group" of nonbeneficiaries. Such a procedure offers a guarantee that the two groups are comparable and have the same characteristics on average, since they have been selected at random. It also offers a guarantee that any performance differential between the two groups is indeed attributable to the measure being tested, on condition that the control group is not indirectly affected by the measure. For example, a measure to help with job search may affect individuals who are not its beneficiaries if they are in competition with those who are. In this case, the employment outlook for the control group is negatively affected by the measure, and failure to take this effect into account may lead to an overly positive evaluation of the efficiency of the help with job search given to the treatment group (Cahuc and Le Barbanchon, 2010; Gautier et al., 2012). It is also possible that members of the control group may alter their behavior out of disappointment at not having been included among the beneficiaries. Controlled experiments focused on job search help and the monitoring of the activity of job seekers have been

carried out in the United States (see chapter 14, section 4.2.1) and in Europe (see van den Berg and van der Klaauw, 2006; Micklewright and Nagy, 2010). But such experiments are hard to set up, which is why most studies of the effects of unemployment insurance rely on "natural experiments." Such experiments take advantage of policy changes or external shocks that exert varying effects on groups of persons having characteristics as unvarying as possible. This is the approach taken by Lalive et al. (2006), which we will now present.

3.1.2 Difference-in-Differences

Lalive et al. (2006) identify the causal effect of benefit duration on the willingness of individuals to accept jobs using a policy change that took place in Austria on 1 August 1989. The replacement ratio, equal to the ratio of unemployment benefits over the previous wage, was increased by about 15% for workers earning below a certain threshold, whereas for workers above this threshold the replacement ratio remained unchanged. The potential benefit duration was increased, depending on age and experience: for workers younger than age 40 and/or for workers with little previous work experience, potential benefit duration remained unchanged; for workers with high levels of previous work experience, potential benefit duration increased, respectively, from 30 to 39 weeks for the age group 40 to 49; and from 30 to 52 weeks for workers aged 50 and older.

Accordingly, the policy change affected various unemployed workers differently, as shown in table 5.5. A first group, denoted eRR, experienced an increase in the replacement ratio. A second group, denoted ePBD, experienced an extension of potential benefit duration. A third group, denoted ePBD-RR, experienced both a higher replacement ratio and a longer potential benefit duration. The control group experienced no change in the policy parameters.

To assess the impact of changes to financial incentives on transition rates out of unemployment, Lalive et al. use longitudinal individual data from two sources: the Austrian social security database, which contains detailed information on individuals' employment, unemployment, and earnings history since the year 1972 and some information on the employer, like region and industry affiliation; and the Austrian unem-

TABLE 5.5
Changes in the replacement ratio (RR) and in potential benefit duration (PBD) on 1 August 1989 in Austria.

		Age			
		Younger than 40 work experience		40 and older work experience	
		Low	High	Low	High
Monthly income	$\leq$ 12,610 Austrian shillings	eRR	eRR	eRR	ePBD-RR
	$>$ 12,610 Austrian shillings	Control	Control	Control	ePBD

Note: Work experience "low" refers to less than 6 years of experience out of the previous 10 years and less than 9 years of experience out of the previous 15 years. Work experience "high" refers to more than 6 years of experience out of the previous 10 years and more than 9 years of experience out of the previous 15 years. ePBD: eligible for increase in potential benefit duration; eRR: eligible for increase in replacement ratio; ePDB-RR: eligible for increase in potential benefit duration and in replacement ratio.
Source: Lalive et al. (2006, table 2, p. 1018).

ployment register, which supplies information on the relevant socioeconomic characteristics. From these data Lalive et al. extract a sample that contains all unemployment entrants for the period between 1 August 1987 (two years before the policy change) and 31 July 1991 (two years after the policy change). They concentrate on job seekers in the age bracket 35 to 54, who have worked at least 52 weeks within the last two years before entering into unemployment and who reside in regions that were never eligible for a special regional extended benefit program. They end up with 225,821 unemployment spells.

The median duration of unemployment is 12 weeks. More than 85% of spells end in a job, while 14% of spells end in a non-job exit destination (long-term sickness, pension, unknown). Since spells are observed until May 1999, only 1% of unemployment entrants in the period between 1 August 1987 and 31 July 1991 are still looking for a job in May 1999. These spells are *censored*. We speak of "left censoring" when the (unknown) date of the start of the unemployment spell falls prior to the date on which the survey commences, and "right censoring" when an individual is still looking for work on the date when the survey stops. The survey simply reveals that when a spell is censored, the actual duration of unemployment is at least equal to the reported duration of this censored spell in the survey. The data exploited by Lalive et al. are of very high quality, since the proportion of censored spells is very small. But that is not always the case, and we will see below how to deal with censored data.

In this setting it becomes possible to evaluate the impact of the changes that came into effect on 1 August 1989, thanks to the difference-in-differences technique, which consists of comparing the respective trajectories of the average performances of the treated and untreated groups. Let $\bar{Y}_B^T$ be the average duration of unemployment for a treated group before the date of the reform (B for before) and $\bar{Y}_A^T$ its average duration after the date of the reform (A for after). Let $\bar{Y}_B^C$ be the average duration of unemployment for the control group before the reform and $\bar{Y}_A^C$ its average duration after the reform. The difference-in-differences estimator, denoted $\tilde{\Delta}_{DD}$, is defined by:

$$\tilde{\Delta}_{DD} = \left(\bar{Y}_A^T - \bar{Y}_B^T\right) - \left(\bar{Y}_A^C - \bar{Y}_B^C\right) \tag{5.32}$$

Thus the difference-in-differences estimator is equal to the difference between the before-after estimator of the treated group $\left(\bar{Y}_A^T - \bar{Y}_B^T\right)$, and the before-after estimator of the control group $\left(\bar{Y}_A^C - \bar{Y}_B^C\right)$. Plainly it is not possible to observe what would have happened if the reform had not taken place. Hence the validity of this estimator of the average effect of the treatment rests on the hypothesis that the difference in the duration of unemployment between the treatment group and the control group would have remained constant in the absence of the reform (see chapter 14, section 3.3.1 for a more formal presentation). This is the *common trend assumption*. For it to be credible, the researcher must strive to ensure that the observable characteristics of individuals in both groups are as alike as possible. So the goal will be to set bounds to the groups described in table 5.5 by selecting persons with age ranges and income levels as close to one another as possible. But the number of observations available frequently imposes limits. For this reason, in a first iteration, Lalive et al. take into account individuals aged 35 to 54 to present their main results, which puts at their disposal 225,821 spells of unemployment. This is a fairly wide bracket, which is why in subsequent iterations

they limit themselves to narrower brackets, at the cost of fewer observations against which to test the robustness of their results.

It is also necessary to check that the previous patterns of unemployment duration of the two groups are parallel and that their composition remains stable over time, in order to be certain that the results observed do not arise out of a change in the composition of the groups. Finally, it is important to verify that the reform was not anticipated; otherwise the behaviors of the individuals in the treatment group might have altered prior to the date of the reform.

The identification strategy chosen by Lalive et al. fits well with their natural experiment and the data available to them. It is a strategy frequently employed in labor economics. But its precondition is a change in economic policy that affects comparable groups differently. Such events are not always readily to hand. But other strategies are possible. Blundell and Costa Dias (2009) offer an overview of these strategies for the whole area of public policy.

3.2 Estimation

First we show that the difference-in-differences estimator does allow us to evaluate the impact of the unemployment insurance reform on the average duration of unemployment. Then we will see how the consequences of the reform may be studied in greater depth by examining its impact not just on the average but also on the distribution of unemployment durations, with the help of survival and hazard functions, estimations of which yield rates of exit from unemployment sorted by the duration of the spell of unemployment.

3.2.1 Unemployment Duration

It is simple to make a first pass at estimating the impact of the reform of the Austrian unemployment insurance system that took place in August 1989 using the difference-in-differences estimator defined by equation (5.32). The unemployment durations of persons who entered unemployment before 1 August 1989 serve to calculate average durations prior to the reform, since such persons were not its beneficiaries. The durations of persons who entered unemployment after 1 August 1989 serve to calculate average durations subsequent to the reform. Since these data are available only for the two years prior to the reform, durations greater than 104 weeks (or 2 years) are left out. Let t_u denote the realized duration of unemployment measured in weeks. Unemployment duration is defined as $t_u^{104} \equiv \min(t_u, 104)$. This definition discards unemployment durations greater than 104 weeks. As the latter concern no more than 1.65% of the sample, we may assume that the results obtained in this way furnish a good first approximation. They are presented in table 5.6. Column 3 shows that the duration of unemployment rises for all groups after August 1989. Column 4, however, which presents the difference-in-differences estimator, reveals that the average duration of unemployment rose more for the groups that benefited from more generous unemployment insurance after that date.

These results, which are coherent with the predictions of the job search model, permit us an initial overview of the impact of the reform. They allow us to calculate that an increase of 1% in the replacement ratio leads to a lengthening of the duration

Table 5.6
Average unemployment duration in first 104 weeks (measured in weeks).

	Before August 1989	After August 1989	Change (after-before)	Diff-in-diff (compared to control)
ePBD group	16.25 (.08)	18.67 (.09)	2.42 (.12)	1.13 (.18)
N	48,294	51,110		
eRR group	17.79 (.12)	20.*m*03 (.16)	2.24 (.20)	.96 (.24)
N	17,160	15,310		
ePBD-RR group	19.01 (.17)	23.55 (.24)	4.53 (.20)	3.25 (.24)
N	11,992	9,182		
Control group	15.24 (.08)	16.52 (.09)	1.29 (.13)	
N	33,815	38,958		

Note: Standard errors in parentheses. N: number of unemployment spells in the group. Diff-in-diff: difference-in-differences; RR: replacement rate; PBD: potential benefit duration; ePBD: eligible for increase in potential benefit duration; eRR: eligible for increase in benefit RR; ePBD-RR: eligible for both.

Source: Lalive et al. (2006, table 4, p. 1020).

of unemployment of 0.3%, in other words, an elasticity of 0.3.[10] They also allow us to show that the elasticity of the duration of unemployment to the potential duration of the payment of benefit is of the order of 0.17.[11] These results fall within the ranges obtained by many other research efforts in this domain.

It is worth noting, however, that the difference-in-differences estimates of table 5.6 are based on rather different groups. Unbiased estimates will be obtained only if there are no group-specific trends in unemployment durations. Lalive et al. (2006) provide a variety of robustness tests, including a focus on more narrowly defined groups and on groups that are just below or just above the eligibility threshold.

Postulating an upper boundary of 104 weeks on individual unemployment durations leads to a perceptible reduction in the average unemployment duration: under this postulate, the average duration amounts to 16.84 months (with a standard deviation of 0.037), whereas without this postulate, it amounts to 18.55 months (with a standard deviation of 0.067). Nor for that matter does this postulate provide any solution to the problem of censored data. It is possible, though, to take better account of the censored data and also to go beyond mere average durations by estimating the impact of the

[10] The unemployment rate of group eRR is in fact 20.03 after the reform. The last column of table 5.6 indicates that it would have been weaker by 0.96 point in the absence of a rise in the replacement ratio, or $20.03 - 0.96 = 19.07$. Hence the augmentation of the replacement ratio raised the unemployment rate by $0.96/19.07 = 5\%$. Since the replacement ratio increases by around 15%, the elasticity of unemployment duration with respect to the replacement ratio is equal to $5\%/15\% = 0.3$.

[11] Persons 50 and older compose 16% of the sample. The potential duration of benefit payments thus grew by $(0.16)(22/30) + (0.84)(930) = 37\%$. Table 5.6 indicates that the increase in the duration of benefit caused the duration of unemployment to lengthen by $(1.13)/(18.67 - 1.13) = 6.4\%$. Consequently the elasticity of the duration of unemployment to the potential duration of benefit payment is $6.4/37 = 0.17$.

reform on the rates of exit from unemployment as a function of the amount of time spent unemployed. This is the procedure adopted by Lalive et al. in the following part of their article. It consists of estimating hazard and survival functions.

3.2.2 Hazard Function and Survival Function

Let us illustrate the "hazard function," which is a basic concept of duration models. In what follows, we will denote the continuous random variable representing the duration of unemployment by T. Like every random variable, the duration of an individual's unemployment spell is characterized by knowledge of its cumulative distribution function denoted $F(t)$, or its probability density $f(t) = F'(t)$. Recall that the cumulative distribution function is defined by $F(t) = \Pr\{T < t\}$ and so represents the probability that the unemployment spell lasts less than t units of time. Theoretical job search models are capable of producing a certain number of predictions about this function, but they most naturally lead to characterizations of the hazard function. The latter represents, for an individual, the instantaneous conditional probability of exiting from unemployment when she has been unemployed for at least a period of length t. For example, in the model in section 2.2.6, in which unemployment insurance benefits are not stationary, the hazard function is equal to $\lambda[1 - H(x(t))]$, where $x(t)$ designates the reservation wage after an unemployment spell equal in length to t. More generally, designating the hazard function by $\varphi(.)$ and knowing that the individual has been unemployed for at least a period of length t, the conditional probability $\varphi(t)\mathrm{d}t$ that the duration of unemployment is located within the small interval of time $[t,\ t+\mathrm{d}t]$ is defined by $\varphi(t)\mathrm{d}t = \Pr\{t \leq T < t + \mathrm{d}t \mid T \geq t\}$. Applying the definition of conditional probabilities[12] gives us:

$$\varphi(t)\mathrm{d}t = \frac{\Pr\{t \leq T < t + \mathrm{d}t\}}{\Pr\{T \geq t\}} = \frac{f(t)\mathrm{d}t}{1 - F(t)}$$

The hazard function is thus characterized by the equality:

$$\varphi(t) = \frac{f(t)}{\overline{F}(t)}, \quad \text{with} \quad \overline{F}(t) \equiv 1 - F(t) \tag{5.33}$$

In this expression there appears the *survival function* $\overline{F}(t)$, representing the probability that the unemployment spell lasts at least a period of length t. Obviously, there is a relation between the survival function and the expected duration of unemployment: $\mathbb{E}(T) = \int_0^{T_u} tf(t)\mathrm{d}t$, where T is defined on the support $[0, T_u]$. Integrating this equation by part, we get:[13]

$$\mathbb{E}(T) = \int_0^{T_u} \overline{F}(t)\mathrm{d}t$$

[12] Given two events A and B, this definition is written:

$$\Pr\{A \mid B\} = \frac{\Pr\{A \cap B\}}{\Pr\{B\}}$$

With $A = \{t \leq T < t + \mathrm{d}t\}$ and $B = \{T \geq t\}$, we find the formula given in the text.

[13] We use the formula $\int u\mathrm{d}v = uv - \int v\mathrm{d}u$, and posit $u = t$, $\mathrm{d}u = \mathrm{d}t$, $\mathrm{d}v = f(t)\ \mathrm{d}t, v = -[1 - F(t)]$.

so that the expected duration of the spell of unemployment is equal to the integral of the survival function.

It is also useful to link the survival function to the integral $\Phi(t)$ of the hazard function. This integral, also called the "integrated hazard," is defined by $\Phi(t) = \int_0^t \varphi(\xi)\mathrm{d}\xi$. Relation (5.33) can also be written $\varphi(t) = -\partial[\ln \overline{F}(t)]/\partial t$; integrating this equality, we find:

$$\Phi(t) = -\ln \overline{F}(t) \tag{5.34}$$

The integrated hazard is thus equal to the opposite of the logarithm of the survival function.

In practice it is important to know if the duration of the phenomenon under study, in this case the duration of an unemployment spell, increases, diminishes, or remains constant with time already spent unemployed. The hazard function allows us to characterize this notion of "duration dependence" very easily. If $\varphi'(t) > 0$, the probability of exiting from unemployment increases with the amount of time t already spent unemployed, and we refer to "positive duration dependence." Conversely, if $\varphi'(t) < 0$, the probability of exiting from unemployment diminishes with the amount of time t already passed in this state, and we then refer to "negative duration dependence." The model presented in section 2.2.6, for example, in which unemployment benefits tail off as the time spent looking for a job lengthens, exhibits positive duration dependence. It should be noted that the hazard function is not necessarily monotonic: it may increase for certain values of t and diminish for others. The hazard function may equally be independent of the length of an unemployment spell, as is the case in the basic job search model in section 2.1, where the exit rate from unemployment $\lambda[1 - H(x)]$ is a constant.

3.2.3 Nonparametric Estimation

It is possible to estimate the survival function by adopting what is called a nonparametric approach, which makes no hypothesis about the form of the distribution of the durations of unemployment spells. This approach is very useful in the first stage of data exploitation. Following Lalive et al. (2006), we most often adopt the Kaplan-Meier (1958) estimator of the survival function. Table 5.7 presents an extract from the data concerning the 225,821 spells of unemployment that began between 1 August 1987 and 31 July 1991. Column 1 assigns an identifier to each spell of unemployment. Column 2 gives its duration, expressed in weeks. The durations have been ranked in ascending order. The shortest duration is equal to .0712128, which corresponds to half a day (6 hours, assuming that the week comprises 7 days of 12 hours each). There are 17 unemployment spells with a duration of half a day. The variable in column 3 takes the value 1 if the spell is censored, and zero if not. We observe that unemployment spell 189540 is censored and that it matches an individual 36.2 years old at the date he entered into unemployment, as shown in column 4.

Let K designate the number of different durations of unemployment inventoried in the sample of the $n = 225{,}821$ observations and let us rank these durations in ascending order, $\tau_1 < \tau_2 < \ldots < \tau_K$. Let us denote n_i the number of unemployment spells the duration of which is at least equal to τ_i. If there is no censoring, the Kaplan-Meier estimator

TABLE 5.7
Extraction from the data set of Lalive et al. (2006).

id	dur	uncc	age
1	0.712128	0	49.99863
2	0.712128	0	49.99863
...	...	...	...
189540	25.78669	1	36.21355
...	...	...	...

Note: id: identification number of unemployment spells; dur: duration of the unemployment spell; uncc equals 1 if the spell is censored and equals zero otherwise; age: age of the individual at the beginning of the unemployment spell.

of the survival function, $\bar{F}(\tau_i)$, in other words the probability that the duration of unemployment is at least equal to τ_i, is simply equal to the proportion of persons whose unemployment duration is greater than τ_i, or $\hat{S}(\tau_i) = n_i/n$. But this estimator requires modification when there are censored observations. Such is the case here, since some persons remained unemployed in May 1999. Let d_j be the number of unemployment spells with a duration equal to τ_j and let c_j be the number of censored spells lying between τ_j and τ_{j+1}. We may then define the number of spells of unemployment with a duration at least equal to τ_i by $n_i = \sum_{j=i}^{K}(d_j + c_j)$. In this case, an estimation of the hazard function, $\hat{\varphi}(\tau_i)$, which here corresponds to the probability that the unemployment spell has a duration of exactly τ_i, is given by:

$$\hat{\varphi}(\tau_i) = \frac{\text{Number of spells with duration equal to } \tau_i}{\text{Number of spells with duration at least equal to } \tau_i} = \frac{d_i}{n_i}$$

The quantity

$$1 - \hat{\varphi}(\tau_i) = \frac{\text{Number of spells with duration greater than } \tau_i}{\text{Number of spells with duration at least equal to } \tau_i} = \frac{n_i - d_i}{n_i}$$

will then represent an estimation of the probability that a spell of unemployment has a duration greater than τ_i. Now, if one's spell of unemployment has lasted longer than τ_i, one has to have been unemployed for all the durations τ_j, $j \leq i$. The Kaplan-Meier estimator of the survival function—in other words, an estimation of the probability that the duration of unemployment is at least equal to τ_i—may then be defined as follows:

$$\hat{S}(\tau_i) = \prod_{j<i}[1 - \hat{\varphi}(\tau_j)] = \prod_{j<i}\frac{n_j - d_j}{n_j}$$

The Kaplan-Meier estimators of the survival function and the hazard function are programmed into the standard software used in econometrics. They prove highly useful in describing the form of the survival and hazard functions of different groups.

Figure 5.4 compares the Kaplan-Meier estimators of the survival functions before and after August 1989 of persons more than 40 years old whose wage prior to their entry into unemployment was greater than 12,610 Austrian shillings. Members of this group

benefited from the reform of August 1989 (see table 5.5), which raised the potential duration of unemployment benefit from 30 to 39 weeks for those younger than 50, and from 30 to 52 weeks for those 50 and older. Each curve in figure 5.4 portrays an estimation of the probability of still being unemployed as a function of the number of weeks already spent being unemployed. We observe that the two survival functions diverge after 15 weeks. The gap widens until the 40-week point, then narrows and becomes constant from 65 weeks on. These results are coherent with the predictions of the job search model, which indicate that the unemployed raise their reservation wage and reduce their search effort when the unemployment insurance system becomes more generous. Under these conditions, the probability of remaining unemployed increases.

The Kaplan-Meier estimator of the hazard function, shown in figure 5.5, allows us to visualize the rates of exit from unemployment during spells of unemployment. This figure shows a significant peak in the exit rate at week 30 for unemployment spells that began before the reform, and that two significant peaks appear at weeks 39 and 52 for spells that began after the reform. Such a pattern of movement in the dates of the peak rates of exit from unemployment suggests that the reform did indeed have a causal impact on exit rates from unemployment. This can be seen with greater precision in figure 5.6, which represents the difference-in-differences of the rates of exit from unemployment as between this group and the control group. At week 30, we observe a drop in the exit rate from unemployment for the test group relative to the control group. The peaks in the differences between the exit rates shift toward week 39 and especially toward week 52. We also note that the difference-in-differences are large for the spell situated before week 60 and subsequently become quite small, the reason being that the reform did not alter the parameters of the benefits scheme for spells of unemployment that exceed 52 weeks.

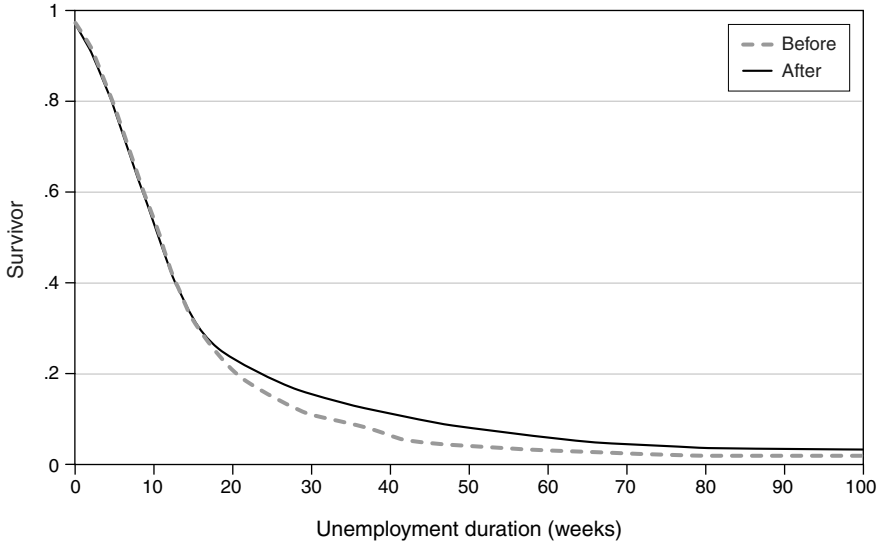


Figure 5.4
Kaplan-Meier survivor functions before and after the reform for the group of individuals potentially eligible to the extension of potential duration of benefits.

Source: Data from Lalive et al. (2006).

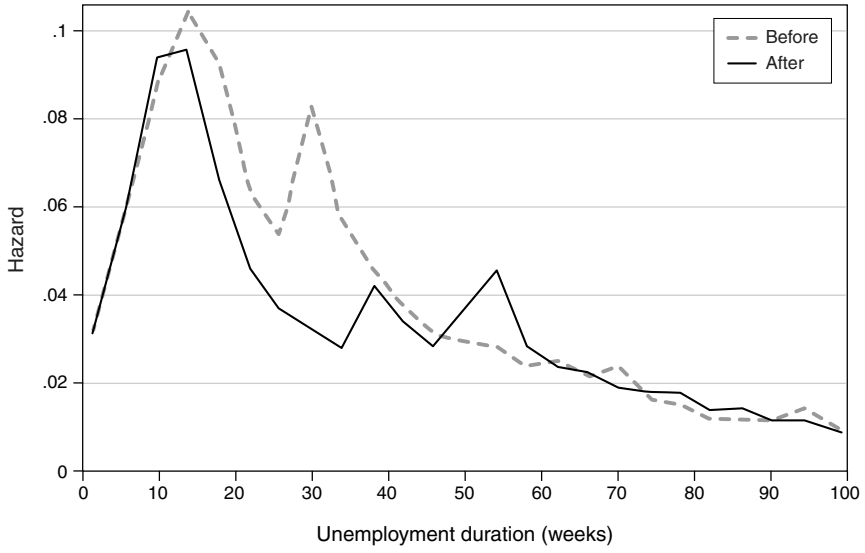


Figure 5.5
Kaplan-Meier hazard functions before and after the reform for the group of individuals potentially eligible to the extension of potential duration of benefits. Hazard functions are smoothed on 4-week windows.

Source: Data from Lalive et al. (2006).

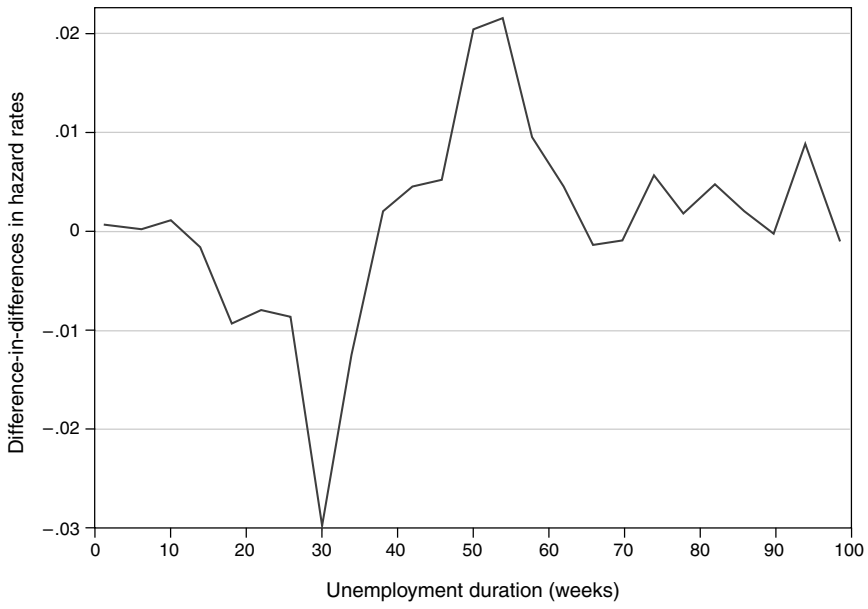


Figure 5.6
Difference-in-differences in Kaplan-Meier hazard functions for the group of individuals potentially eligible to the extension of potential duration of benefits and the control group. Hazard functions are smoothed on 4-week windows.

Source: Data from Lalive et al. (2006).

The nonparametric estimation of the survival and hazard functions constitutes an important stage in the process of evaluation. It allows us to observe the alterations in the average rates of exit from unemployment for the treated groups and the control group. Still, it does not allow us to estimate the impact of the unemployment insurance modifications conditional upon an array of explanatory variables. Of course, it is possible to calculate the survival and hazard functions for groups of either gender, different age, and different educational level. But in order to detect the behavior of the unemployed conditional upon their individual characteristics, it is necessary to estimate a model that expresses the survival and hazard functions as a function of these characteristics. This step is highly instructive, given that the characteristics of the persons constituting the treatment and control groups differ. To accomplish it, Lalive et al. (2006) adopt (as most research in this domain does) a parametric approach: they make hypotheses about the form of the distribution of the probabilities of unemployment duration. It then becomes possible to estimate the rate of exit from unemployment taking into account individual heterogeneity. This approach is called "the reduced form approach" because it simply relies on the predictions of some predefined models. This approach is different from the structural approach, in which each study estimates parameters of a specific model explicitly derived from the theory. For a good example of the structural model, consult Wolpin (1987), Devine and Kiefer (1991, chapter 5), and the survey of Eckstein and van den Berg (2007).

3.2.4 Parametric Estimation

To estimate the impact of financial incentives on the rate of exit from unemployment while taking into account the observable heterogeneity of individual characteristics, Lalive et al. (2006) use a model with proportional hazard.

The Likelihood Function

In the proportional hazard model, we assume that the vector $\boldsymbol{\theta}$ of the explanatory variables is composed of two subsets, $\boldsymbol{\theta_0}$ and $\boldsymbol{\theta_x}$, and that the hazard function takes the following form:[14]

$$\varphi(t, \mathbf{x}, \boldsymbol{\theta}) = \varphi_0(t, \boldsymbol{\theta_0})\rho(\mathbf{x}, \boldsymbol{\theta_x}) \tag{5.35}$$

Function φ_0 is called the "baseline hazard" because it is identical for all individuals, so that $\boldsymbol{\theta_0}$ is a vector of parameters to be estimated, independent of individuals' characteristics. Most often we utilize a well-specified function, for example, the Weibull distribution or, as Lalive et al. do, a piecewise constant function of elapsed duration.[15]

Relation (5.35) shows that, in the proportional hazard model, the effect of the vector $\mathbf{x}$ of individual characteristics is to multiply the baseline hazard by the scale

[14] This hypothesis amounts to assuming that the cumulative distribution function of the random variable T takes the expression $F(t) = 1 - e^{-\rho(\mathbf{x}, \boldsymbol{\theta_x})\int_0^t \varphi_0(\tau, \boldsymbol{\theta_0})\mathrm{d}\tau}$.

[15] Note that for proportional hazard models, it is possible to proceed to a semiparametric estimation by specifying the scale factor a priori while not imposing any particular form for the baseline hazard (in that case, we must utilize the empirical distribution of the unemployment durations). This so-called partial-likelihood approach was suggested by Cox (1975); one may consult Kiefer (1988, IV-C) for a good introduction to it.

factor $\rho(\mathbf{x}, \boldsymbol{\theta}_x)$ independent of the duration t of unemployment. A specification frequently used, as Lalive et al. do, for the scale factor is $\rho(\mathbf{x}, \boldsymbol{\theta}_x) = \exp(\mathbf{x}\boldsymbol{\theta}_x)$, which has the advantage of being positive and supplying a simple interpretation of the components of the vector $\boldsymbol{\theta}_x$. If we denote by x_k the k^{th} component of vector $\mathbf{x}$ of individual characteristics, relation (5.35) defining the hazard function shows that $(\partial \ln \varphi / \partial x_k) = \theta_{xk}$, where θ_{xk} designates the k^{th} component of vector $\boldsymbol{\theta}_x$. If we have been careful to specify the explanatory variables in terms of logarithms, vector $\boldsymbol{\theta}_x$ then represents the vector of the elasticities of the hazard function, that is, the elasticities of the conditional probability of exiting unemployment with respect to the explanatory variables.[16]

The estimators of vectors $\boldsymbol{\theta}_x$ and $\boldsymbol{\theta}_0$ are obtained by maximizing the likelihood function of the sample with respect to the components of vectors $\boldsymbol{\theta}_x$ and $\boldsymbol{\theta}_0$. If there is no censoring, the likelihood function of the sample is written $\prod_{i=1}^{n} f(t_i, \mathbf{x}, \boldsymbol{\theta})$, where f denotes the probability density of unemployment durations and t_i designates the length of observation i. This likelihood can be computed from the hazard function and the survival function using equation (5.33). The likelihood is merely $\prod_{i=1}^{n} \varphi(t_i, \mathbf{x}, \boldsymbol{\theta}) \bar{F}(t_i, \mathbf{x}, \boldsymbol{\theta})$. But in reality some spells are censored. If observation t_i is censored, the survey simply reveals that the duration of unemployment T_i is *at least* equal to t_i. The contribution of this observation to the likelihood of the sample is then equal to $\Pr\{T_i \geq t_i\} \equiv \overline{F}(t_i|\mathbf{x}, \boldsymbol{\theta})$. Let us define the dummy variable k_i by $k_i = 1$ if the observation is not censored, and by $k_i = 0$ if it is. Then the likelihood function becomes[17] $\prod_{i=1}^{n} [\varphi(t_i, \mathbf{x}, \boldsymbol{\theta})]^{k_i} \bar{F}(t_i, \mathbf{x}, \boldsymbol{\theta})$. It is possible to express this likelihood function solely with the help of the hazard function $\varphi(t, \mathbf{x}, \boldsymbol{\theta})$ and its integral, the integrated hazard $\Phi(t, \mathbf{x}, \boldsymbol{\theta})$. Relations (5.33) and (5.34) thus give $\ln f(t_i, \mathbf{x}, \boldsymbol{\theta}) = \ln \varphi(t_i, \mathbf{x}, \boldsymbol{\theta}) - \ln \overline{F}(t_i, \mathbf{x}, \boldsymbol{\theta})$ with $\Phi(t_i, \mathbf{x}, \boldsymbol{\theta}) = -\ln \overline{F}(t_i, \mathbf{x}, \boldsymbol{\theta})$, and the likelihood of the sample becomes, in logarithmic form, which is generally used to proceed to maximization:

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} k_i \ln \varphi(t_i, \mathbf{x}, \boldsymbol{\theta}) - \sum_{i=1}^{n} \Phi(t_i, \mathbf{x}, \boldsymbol{\theta}) \tag{5.36}$$

In practice the estimator $\widehat{\boldsymbol{\theta}}$ of vector $\boldsymbol{\theta}$ of the parameters corresponds to the value of $\boldsymbol{\theta}$ that maximizes this log-likelihood function. This maximization most often gives no analytical solution, and it is necessary to fall back on numerical methods. Maximum likelihood-based methods are now so common that standard econometric software packages have routines for many of these methods.

In the hazard proportional model, it is easy to check that if we assume that the scale factor takes the form $\exp(\mathbf{x}_i \boldsymbol{\theta}_\mathbf{x})$, the log-likelihood function is:

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} k_i \left[(\mathbf{x}_i \boldsymbol{\theta}_\mathbf{x}) + \ln \varphi_0(t_i, \boldsymbol{\theta}_\mathbf{0}) \right] - \sum_{i=1}^{n} \Phi_0(t_i, \boldsymbol{\theta}_\mathbf{0}) \exp(\mathbf{x}_i \boldsymbol{\theta}_\mathbf{x})$$

In this expression, function Φ_0 represents the integrated hazard of the baseline hazard φ_0.

[16] Recall that the elasticity of function $f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$, with respect to x_i is $(x_i / f(\mathbf{x}))\,(\partial f(\mathbf{x}) / \partial x_i) = \partial \ln(f(\mathbf{x})) / \partial \ln(x_i)$.

[17] This expression of the likelihood function assumes that the censoring mechanism is independent of the duration T_i of unemployment.

The Estimation of Parameters

The set of individual characteristics **x** accounted for by Lalive et al. comprises age, marital status, gender, education, log (previous monthly income), recall status, blue collar (versus white collar), seasonal industry, manufacturing industry, time spent unemployed, tenure, and quarter of inflow into unemployment. Their baseline hazard φ_0 is a piecewise constant function of elapsed duration defined as follows:

$$\varphi_0(t_i, \boldsymbol{\theta_0}) = \exp\left(\sum_{j=0}^{14} \theta_j \mathbb{I}(4j < t_i \leq 4(j+1)) + \theta_{15}\mathbb{I}(t_i > 60)\right)$$

where $\mathbb{I}(x)$ is an indicator function equal to 1 if condition x is verified, and zero if not. This expression of the baseline hazard incorporates the assumption that the hazard rate shifts in every four-week interval until week 60 and, because there are very few transitions beyond week 60, the last time interval covers the entire remaining duration of the spell as of week 60. Parameters θ_j allow us to detect the evolution of the hazard according to the duration of the spell of unemployment, and thus furnish a measure of the "duration dependence" of the process of exiting unemployment and hence serve to pinpoint the impact of the reform of August 1989.

Parameters θ_j can be estimated with the help of a specification of the difference-in-differences type. Let us limit ourselves to considering only the impact of the prolongation of the potential duration of unemployment benefit. Eligibility for the increase in potential benefit duration is denoted by a function *ePBD* equal to 1 if individuals belong to the group *ePBD* defined in table 5.5 and to zero otherwise, that is, $ePBD = I(ePBD = 1)$. Second, we introduce the function $A89$ equal to 1 if the unemployment spell of length t_i began after 1 August 1989 and to zero otherwise for all individuals. Thus, the interaction term $ePBD \times A89$ indicates that an individual satisfying all eligibility criteria for the increase in potential benefit duration has entered the period when this policy change has been enacted. The duration dependence of the hazard rate is defined as follows:

$$\theta_j = \beta_{0j} + \beta_{1j} ePBD + \beta_{2j} A89 + \delta_j ePBD \times A89$$

$$j = 0, \ldots, 15$$

Assuming at this stage that the sample includes only individuals belonging to the control group and to the *ePBD* group, parameters β_{1j} capture ex ante differences between the *ePBD* group and the control group, parameters β_{2j} capture changes to duration dependence occurring over time for all individuals, and parameters δ_j measure the change in the duration dependence of the hazard rate due to changes in the potential benefit duration in August 1989.

Lalive et al. (2006) use this type of specification for all groups defined in table 5.5. They estimate the parameters β and δ proper to each group by maximizing the likelihood function of the sample. With the parameters estimated, it becomes possible to simulate the impact of the reform on the hazard and survival functions. Figure 5.7 represents the simulation of the impact of the reform on the hazard rate of group *ePBD*, which benefited from the prolongation of the potential duration of unemployment benefit from 30 to 39 weeks. Comparison with the control group reveals that this prolongation shows

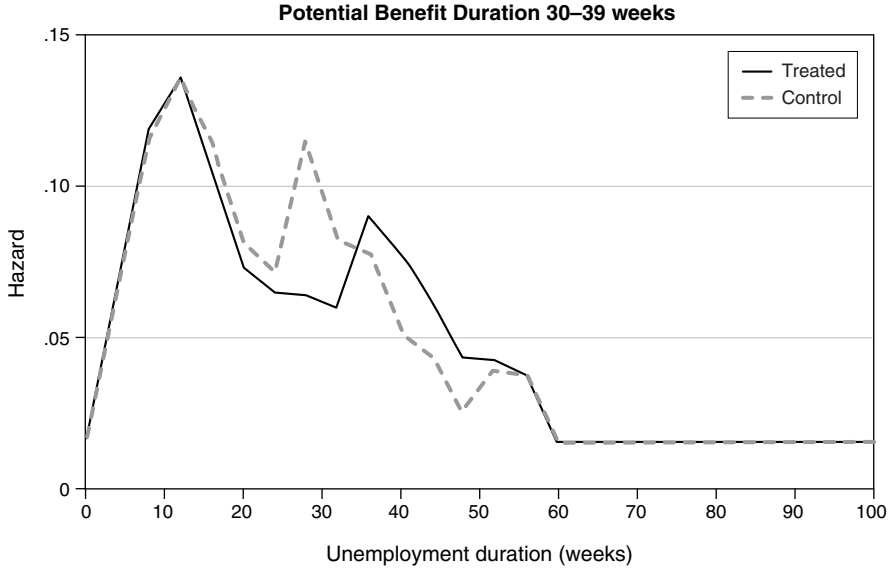


Figure 5.7
Estimated average treated and control hazard rates for the group who benefited from the extension of the potential benefit duration.

Source: Lalive et al. (2006, figure 5, p. 1026).

up in a shift of the peak point for exit from unemployment from week 30 to week 39. Taking individual heterogeneity into account thus confirms the result obtained by the nonparametric Kaplan-Meier estimation (see figure 5.5).

Parametric estimation makes it possible to perceive the impact of modification of the parameters of unemployment insurance as a function of observable individual characteristics. For example, it turns out that older unemployed persons react more strongly to these modifications. This might result from the greater difficulty they experience in reinserting themselves into the labor force or from their nearness to retirement age, which might make their search effort and reservation wage more sensitive to unemployment insurance (Hairault et al., 2010).

3.2.5 Unobserved Heterogeneity

Explanatory variables such as sex, educational attainment, and past experience allow us to control to a degree the heterogeneity among individuals. But unobserved heterogeneity always remains: for example, personal motivation is generally not observed. The omission of some variables, or specification errors in the impact of the exogenous ones, are formally much like unobserved heterogeneity. Failure to take this type of heterogeneity into account leads to bias in the estimation of time dependency.

To see this clearly, consider an example in which there is a fraction p of the population which has a constant hazard function γ_1 and a fraction $(1-p)$ which has a constant hazard function γ_2. The hazard function of the whole sample is equal to:

$$\varphi(t) = \frac{p\gamma_1 e^{-\gamma_1 t} + (1-p)\gamma_2 e^{-\gamma_2 t}}{pe^{-\gamma_1 t} + (1-p)e^{-\gamma_2 t}}$$

It is easy to verify that $\varphi'(t) < 0$. Consequently, the omission of unobserved heterogeneity can falsely introduce a negative duration dependence, since in reality the individual probability of finding a job is independent of the amount of time spent unemployed.

The presence of unobserved heterogeneity can also lead to a selection bias prejudicial to the estimation of hazard functions. Such is the case, for example, in Lalive et al. (2006), if the unemployment insurance reform exerts different effects on the behavior of persons whose *unobservable* characteristics are different. Let us suppose, for example, that members of group *ePBD*—beneficiaries of the increased potential duration of unemployment insurance—have on average observable and unobservable characteristics identical to those of the control group at the moment of the onset of the reform, 1 August 1989. Let us also suppose that those persons with the most motivation to look for work are less sensitive to the increased potential duration of unemployment insurance and that this motivation is unobservable. In this setting, the unobservable characteristics of the group that benefited from the increased potential duration of unemployment insurance will play a different role than the unobservable characteristics of the control group during spells of unemployment. If these (plausible) hypotheses are verified, then the ratio between hazard rates at any given instant will simultaneously reflect the causal effect of the potential duration of benefit payments and a selection bias on unobservable characteristics.

To get around these difficulties we may assume that the probability density of the dependent variable is written (leaving out vectors $\mathbf{x}$ and $\boldsymbol{\theta}$ for the sake of simplicity) $f(t, v)$, where v is a random variable of density $p(.)$ marking the unobserved heterogeneity among agents. For example, in the proportional hazard model, it is possible to introduce this form of heterogeneity by assuming that the hazard function takes the form $\varphi(t, \mathbf{x}, \boldsymbol{\theta}) = \rho(\mathbf{x}, \boldsymbol{\theta}_x)\varphi_0(t, \boldsymbol{\theta}_0)v$. We thus obtain the *mixed proportional hazard model* studied in detail by Lancaster (1979), van den Berg (2001), and Abbring and van den Berg (2003). The probability density function $p(.)$ of the random variable v is unknown and must therefore be estimated. In practice it can be assumed to follow a Gamma distribution, with mean normalized to 1 and variance equal to $1/\sigma$. This adds a single parameter to estimate in the likelihood. A discrete law (v_k, p_k) is also often used, with $p_k = \Pr\{v = v_k\}$ for $k = 1, .., K$, and we estimate the vector $(v_1, .., v_K ; p_1, .., p_K)$ along with all the other parameters of the model. Lalive et al. (2006) did robustness checks of their estimates with a mixed proportional hazard model allowing for a discrete distribution of unobserved heterogeneity with two mass points. Although they find evidence of the presence of these two mass points, the estimated effects of a change in the replacement ratio and in the potential benefits duration are not affected by taking unobserved heterogeneity into account.

3.3 Main Results on the Determinants of Unemployment Duration

The study of Lalive et al. (2006) allows readers to grasp the procedure generally followed in estimating the impact of changes to the unemployment insurance system on the duration of unemployment. But it gives no more than a glimpse of the large body of results obtained in this domain and reported in a burgeoning literature. The average duration of spells of unemployment proves to be strongly linked to the replacement

rate and potential benefit duration. There exists, however, a strong heterogeneity in the duration of spells of unemployment in relation to the individual characteristics of job seekers. We present here a synthesis of the main results.

3.3.1 The Effects of Potential Benefit Duration and Replacement Rate

In line with the predictions of the job search model, empirical studies find that potential benefit duration and the replacement rate exert significant effects on the duration of unemployment. They also find that the quality of the jobs that are found may be affected by these two parameters, but here the results are less striking.

The survey of Tatsiramos and van Ours (2012), which selects studies with relevant identification strategies, shows that the magnitude of the effects of unemployment benefits differs for different countries and different types of policy changes but that the effects themselves differ less. Their survey warrants us to accept the two following orders of magnitude: first, the elasticity of the duration of unemployment with respect to the replacement ratio varies between 0.4 and 1.6; and second, an increase of a week in the potential duration of benefit payments leads to an increase in the duration of unemployment ranging between 0.1 and 0.4 of a week.

Numerous studies, following the contribution of Meyer (1990), highlight a significant discontinuity in the exit rate from unemployment in the period immediately preceding the exhaustion of entitlement to unemployment insurance benefits, as displayed in Figure 5.8. The size of the spike is influenced by the characteristics of workers and by the institutional environment. This is illustrated by the studies of Dormont et al. (2001) on French data. They show as well that the exit rate from unemployment to employment rises more at the end of the entitlement period for better-qualified job seekers. Figure 5.8 clearly illustrates this phenomenon. It traces the exit rate from unemployment for individuals whose benefits drop significantly in the 14th month of unemployment. At that

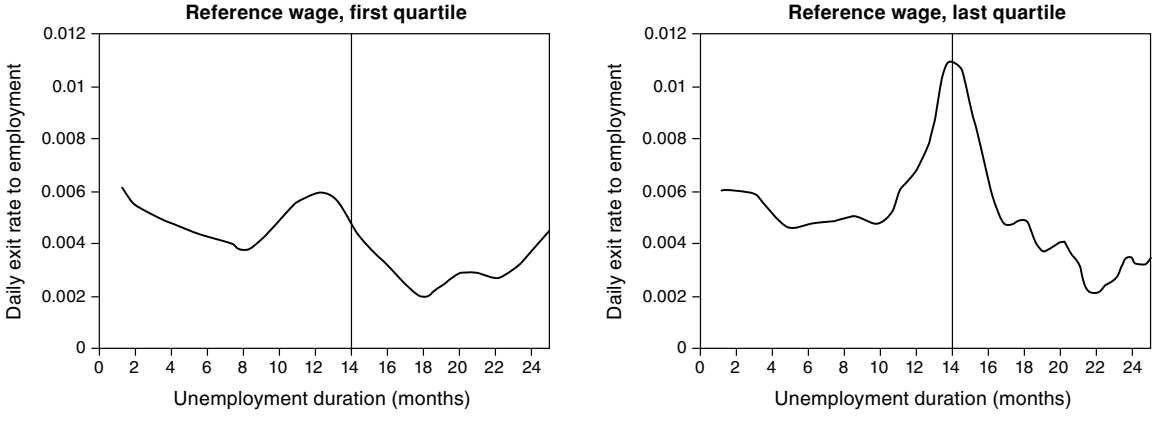


Figure 5.8

Exit rate from unemployment into employment and the end of entitlement to benefits. Period: 1986–1992. Population: individuals aged 25 and older. The reference wage corresponds to the average wage for the 12 months immediately preceding job loss.

Source: Dormont et al. (2001).

time, benefits pass from a magnitude of 57% to 75% of the *previous* wage to a fixed sum corresponding to roughly 60% of the *minimum* monthly wage. Figure 5.8 shows that the probability of exit rises significantly as the 14th month approaches. Further, this effect is much more marked for job seekers who previously earned high wages. Two causes contribute to this phenomenon. First, better-qualified workers, ones earning higher wages, are also those who can find jobs more easily and behave in a more opportunistic manner. Second, the fall in income in the 14th month is weaker to the extent that the reference wage was low to begin with. The question of the relative importance of these two causes remains open.

The studies of Dormont et al. (2001) on French data limit themselves to exits from unemployment into employment. But Card et al. (2007b) have pointed out that the peak observed is markedly more pronounced if all types of exit from unemployment are taken into account. The fact is, a large number of such exits are transitions into other states such as inactivity or training. Card et al. find, on Austrian data, that the rate of exit from unemployment is multiplied by 2.4 around the maximum duration of benefit, compared to the rate observed at the onset of an unemployment spell. The rate of return to employment, in contrast, is multiplied by only 1.15. In the case of seniors, the results of Hunt (1995) suggest that a substantial portion of the transitions observed at the exhaustion of benefit entitlement amount to exits from the labor market altogether. More generally, Card et al. (2007b) stress that the magnitude of any spike in the reemployment rate depends on institutional factors and labor market conditions that include the availability of post-exhaustion benefits (Pellizzari, 2006), the participation of UI recipients in the informal sector, and the incentives for firms to cycle workers through temporary unemployment.

3.3.2 Post-unemployment Outcomes

The job search model predicts that an increase in the amount of unemployment benefit or in its potential duration ought, by raising the reservation wage, to entail an improvement in the quality of the jobs accepted by the unemployed. The empirical results are very mixed. The first research published on this topic by Burgess and Kingston (1976) and Ehrenberg and Oaxaca (1976) estimated that the generosity of unemployment insurance had a positive impact on the wages of jobs accepted upon exit from unemployment. Belzil (2001) for Canada and Centeno (2004) for the United States find, on one hand, that the jobs accepted at the close of the potential duration of unemployment insurance are more unstable and on the other that a higher replacement ratio leads to more stable jobs. These predictions fit the job search model: as the end of the potential duration of benefit approaches, the lowering of their reservation wage induces the unemployed to accept lower-quality, more unstable jobs. An increase in the amount of unemployment benefit has the contrary effect.

More recent studies, using difference-in-differences or regression discontinuity approaches, which provide more reliable identification, find much more mixed results. Card et al. (2007b) detect no impact of the generosity of unemployment insurance on the wages or the stability of jobs in Austria. Van Ours and Vodopivec (2006) arrive at the same conclusion for Slovenia. Centeno and Novo (2009) detect very weak effects on wages in Portugal.

Hence, in the present state of knowledge, the impact of the generosity of unemployment insurance on the quality of jobs remains an open question.

3.3.3 The Effects of Wealth and Liquidity Constraints

The theoretical model elaborated in section 2.2.4 above showed that a rise in the total wealth of an unemployed person ought to have a disincentive effect on job search by raising her reservation wage and reducing her search effort. From this it follows that a rise in the wealth of an unemployed person may cause spells of unemployment to be more prolonged.

Using a portion of the Austrian data already exploited by Lalive et al. (2006), which we described above, Card et al. (2007a) examine the effect of an entitlement to severance pay (which augments the wealth of the unemployed person) on the intensity of job search. To be entitled to severance pay, one must have worked for at least 36 months. Card et al. use a regression discontinuity design to estimate the effects of severance pay, essentially comparing the search behavior of people who were laid off just before and just after the 36-month cutoff for severance pay eligibility. They verify that the firing decisions of firms are not linked to the payment of severance (the regulations for firing employees are very strict in Austria, and the law forbids any strategic behavior in the timing of such separations). In this case, one may reasonably suppose that there is indeed a random selection around the eligibility threshold. Card et al. find that the rate of return to employment (during the normal period of entitlement to unemployment insurance, which means the first 20 weeks of being unemployed) is between 8% and 12% lower for those who are just eligible to receive a severance payment than for those who are just ineligible.

The contribution of Lammers (2012) brings the behavior of unemployed persons following a change in their wealth into sharper focus. It relies on a Dutch panel that contains detailed information on individual wealth and income, subjective reservation wages, and proxies for search effort. Lammers finds that the wealth of an unemployed person has a positive effect on her reservation wage but has no significant effect on her search effort. So the prolongation of a spell of unemployment would be the consequence of her rejecting more of the job offers she receives rather than her receiving fewer offers because her effort has slackened.

Recent empirical studies, therefore, arrive at strongly convergent results regarding the effects of wealth and liquidity constraints. But they are not yet sufficiently numerous to yield definitive conclusions (Tatsiramos and van Ours, 2012).

3.3.4 The Reservation Wage

The survey of Krueger and Mueller (2011), mentioned above in connection with the use of their time by the unemployed, also supplies interesting indications concerning reservation wages. More than 6,000 unemployment insurance recipients in New Jersey were regularly queried each week throughout autumn 2009 and into spring 2010. Their answers when asked what the lowest wage is they would be willing to accept if they received a job offer are treated as their reservation wages. Table 5.8 reports the average ratio of the reservation wage to the pre-unemployment wage.

We observe that the average values of the reservation wage are very close to the previous wage, whatever the duration of the unemployment spell. But these average values conceal a steep variability. In Krueger and Mueller's survey the 25th percentile reservation wage ratio is 0.70, the median is 0.91, and the 75th percentile is 1.17. We further observe that in 6.7% of cases the reservation wage is below the unemployment benefit.

Table 5.8
Reservation wage ratio by duration of unemployment.

All durations	< 5 weeks	5–9 weeks	10–14	15–19	20–24	25–49	> 50
0.99	1.04	1.02	1.01	1.00	1.06	0.95	0.94

Source: Krueger and Mueller (2011, table 4.1).

Table 5.9
Elasticities of the reservation wages with respect to the income of unemployed persons.

Authors	Data	Elasticities
Lynch (1983)	UK (youth)	0.08 – 0.11
Holzer (1986)	US (youth)	0.018 – 0.049
van den Berg (1990)	Netherlands (30–55 years)	0.04 – 0.09

Source: Devine and Kiefer (1991, table 4.2, p. 75).

An initial series of studies attempted to make direct estimates of relations like equation (5.6) giving the value of the reservation wage in the basic model. To that end, as with the recent work of Krueger and Mueller (2011), they relied on data from surveys in which unemployed persons were asked to answer more or less directly the question, "What for you is the lowest acceptable wage?" Table 5.9 gives the magnitudes of the elasticity of the reservation wage with respect to the income of an unemployed person for three studies that use this type of data.[18] It results that, as the basic model predicts, this elasticity is positive. Its magnitude is however very slight.

3.3.5 Helping and Monitoring the Unemployed

The majority of OECD countries have adopted measures aimed at increasing the efficiency of the job search by those receiving unemployment insurance benefits. In the United States, Denmark, the Netherlands, and the United Kingdom, starting in the 1980s, these measures combine help in looking for a job with sanctions, generally consisting of a reduction in benefit, when the rules imposed by the body administering unemployment insurance are not adhered to (Venn, 2012). More precisely, we can distinguish three types of instruments that are generally used in combination: programs giving individual counseling to job seekers, stronger measures to check that eligibility conditions have been met, and stronger measures to check that suitable efforts to find a job are being made. Studies of experimental and nonexperimental programs usually find that the surveillance and counseling programs may have a significant effect on unemployment exit rates among those who need help. They also exert pressure on a percentage of the eligible unemployed who are not experiencing any real difficulty in finding work. The impact of these programs on unemployment duration and wages is reviewed in detail in chapter 14, section 4.

[18] These are averages of estimates for Lynch (1983) and Holzer (1986). The study by van den Berg (1990) estimates the value of reservation wage elasticity at the onset of a spell of unemployment in relation to the future income of an unemployed person.

4 SEARCH FRICTIONS AND WAGE DIFFERENTIALS

In this section we extend the basic job search model studied above to render *endogenous* the dispersion of wages for individuals endowed with identical productive abilities and preferences. This perspective is important inasmuch as it allows us to understand how individuals with identical productive abilities and preferences and with identical jobs can receive different wages. We begin by providing empirical facts about wage differentials which suggest that workers of identical productivity are paid differently. Then, we will see how the equilibrium search model can explain this phenomenon.

4.1 Empirical Facts About Wage Differentials

In a perfectly competitive labor market, with given productive abilities and working conditions, the wage of any individual ought to be independent of the firm or industry in which she is employed. If one industry or firm pays better than others, perfect mobility of workers ought to lead to a flow of labor supply toward that firm or industry and a consequent drop in remuneration. But the existence of persistent wage differentials among industries and firms is a stark, and abundantly documented, fact. Slichter (1950) had already established that this was the case for American workers between 1923 and 1946.

4.1.1 Interindustry Wage Differentials

Let w_{it} be the hourly wage of an individual i at date t; let $\mathbf{x}_{it}$ be the vector of her personal characteristics and those of her job at the same date; let $J(i,t)$ define worker i's industry at date t and let $d_{ijt} = \mathbb{I}[J(i,t) = j]$ be variable indicators equal to 1 if worker i works in industry j at date t and to zero otherwise. Interindustry wage differences are generally highlighted by estimating an equation of the form:

$$\ln w_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \sum_{j=1}^{J} \gamma_j d_{ijt} + \varepsilon_{it} \tag{5.37}$$

In this equation, ε_{it} designates a residual with zero mean. The coordinates of vector $\boldsymbol{\beta}$ and the γ_j are parameters to be estimated. Assuming that the residuals ε_{it} are independent of the d_{ijt}, the ordinary least squares estimators of the γ_j indicate the impact of the industries on wages. Note that the hypothesis of independence between the ε_{it} and the d_{ijt} is strong, since it means that the allocation of workers among sectors is independent of the error term, which includes unobserved heterogeneity. In other words, we assume that the allocation of workers among sectors is independent of their efficiency, something not observed by the econometrician.

Under these hypotheses, if the γ_j coefficients are significantly different from 0, we must conclude either that there are omitted variables or that there are interindustry wage differentials. Traditionally, the estimation of this type of equation gives coefficients γ_j that are significantly different from 0. The first column of table 5.10 presents the results obtained by Goux and Maurin (1999) for France in 1990–1995, with a breakdown into

39 industries. According to these estimations, the standard gap due to industry is of the order of 8% to 9%. We also observe that the industries that pay the least (agriculture, food retail, and hotel, bar, and restaurant) offer wages 15% lower on average than the rest of the economy. The industries that pay the most (petroleum, mining, chemicals) have wages that are on average 15% above those in the rest of the economy. Goux and Maurin note as well that these wage differences persist over time.

4.1.2 The Importance of Unobserved Worker Ability Differences

The results presented in the first column of table 5.10 are similar to those obtained for France and the United States by Dickens and Katz (1987), Krueger and Summers (1988), Katz and Summers (1989), and Abowd et al. (1999). At first glance, these results suggest that the labor markets are far from being perfectly competitive. Their interpretation is a delicate matter, however. It is quite possible that wage differentials are caused by an unobserved heterogeneity of workers. If those with the greatest productive abilities (unobserved) are concentrated in the same industries, those industries must pay higher wages. In that case, the model explaining wages is not described by equation (5.37), but by:

$$\ln w_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \sum_{j=1}^{J} \gamma_j d_{ijt} + \alpha_i + \varepsilon_{it} \qquad \textbf{(5.38)}$$

where α_i designates the unobserved time-invariant characteristics of individual i. If data for several periods are available, it is possible to eliminate this term by estimating equation (5.38) in differences. Such estimations, carried out on U.S. and French data, and covering a sufficiently large number of industries (Murphy and Topel, 1987; Abowd et al., 1999; Goux and Maurin, 1999), find that interindustry wage differentials are to

Table 5.10
Estimates of interindustry wage differences in France, 1990–1995. Hourly wages. Field: men, wage earners. Selected industries.

Industry	Model without fixed individual effect	Model with fixed individual effect
Agriculture	−.101 (.07)	−.017 (.016)
Mining (coal)	.139 (.020)	.058 (.056)
Petroleum	.210 (.018)	.049 (.027)
Electricity	.108 (.007)	.058 (.019)
Chemical	.163 (.009)	.016 (.019)
Food retail	−.112 (.007)	−.043 (.014)
Hotels, bars, and restaurants	−.175 (.006)	−.008 (.012)

Note: The figures in parentheses are standard errors. Aside from industry, the variables taken into account are experience in the labor market, job seniority, place of residence, education, nationality (French or foreign), and profession. How to read the table: according to the model without fixed individual effect, a wage earner in agriculture receives, on average, a wage 10.1% lower than in the reference sector (public utilities).
Source: Goux and Maurin (1999, table 3).

a very large extent explained by the characteristics of workers. The second column of table 5.10, taken from Goux and Maurin (1999), gives a striking illustration of this. It shows that the contribution of industry to wage setting is much smaller, and often not significantly different from 0, at the threshold of 5%. Further, Goux and Maurin point out that there is a very weak correlation, less than 0.25, between the coefficients estimated by the models with and without fixed individual effect. Finally, they find that the wage variations incurred by an individual who changes industries do not exceed 2% to 3%.

These results point to the conclusion that interindustry wage differences are essentially explained by individual effects. They are very different from the results obtained by Krueger and Summers (1988) and Gibbons and Katz (1992), who, with a smaller number of industries (around 20, rather than around 40), show that industry makes a significant contribution to wage formation, after controlling for fixed individual effects. But if they adopt a breakdown similar to that of Krueger and Summers (1988) and Gibbons and Katz (1992), Goux and Maurin (1999) arrive at conclusions close to theirs for French data. That being so, the results that tend to prove the importance of the industry component obtained by Krueger and Summers (1988) and Gibbons and Katz (1992) are most likely the result of an aggregation bias caused by using too few industries.

4.1.3 Industry Effect and Firm Effect

Although the impact of industry on wage formation appears very slight, it is quite possible that wages are heterogeneous among firms in the same industry. An approach like the one just described must be used to study this problem, but with identification of the firms, and not just the industries, in which individuals work. This is the aim of Abowd et al. (1999), who have estimated the importance of firm-fixed effects with equation (5.38), where the "j" now represents indexes of firms rather than sectors. The estimation of this equation by ordinary least squares permits an evaluation of the contribution of firm-fixed effects to wage dispersion under the hypothesis that the mobility of wage earners among firms is independent of the residue ε_{it}. It is important to note that the identification of firm-fixed effects rests on the observation of wage earners who change firms. It is therefore a requirement to observe a sufficient number of entries into and exits from each firm in order to be able to identify these fixed effects with precision. Thus the panel must have a duration sufficiently protracted to satisfy this condition.

Use of this method shows that firm-fixed effects explain a significant part of wage distribution. Table 5.11 shows that in France firm-fixed effect explains 17% of the variation in the logarithm of wages, while individual-fixed effect explains 29%. In the

TABLE 5.11
Components, in percentage, of variations of (log of) real annual wages in France and in the United States.

	France (1976–1999)	United States (1990–1999)
Experience and experience squared	17	5
Person effect	29	24
Firm effect	23	24
Residual	31	47

Source: Abowd et al. (2003).

United States, firm- and individual-fixed effects each explain 24% of wage distribution. Abowd et al. (2013) confirmed these results using more recent data and improved estimation techniques.

Studies that take this approach show unambiguously that the impact of the firm is greater than that of the industry. Goux and Maurin (1999), for example, assess the average of the difference in wages paid to an identical worker employed at two different firms in France at 20% to 30%, whereas it does not exceed 2% to 3% for a change of industry. They show, moreover, that these differences are positively correlated with the size and the capital/labor ratio of firms. The correlations with productivity and profitability are much smaller and much less significant.

All in all, these studies show that individuals with identical time-invariant characteristics are paid differently when they work in different firms. Once again, the interpretation of these results is a delicate matter. It is possible that wage differences among firms are the result of unobserved differences linked to working conditions on a perfectly competitive labor market. The compensating differential theory of wages (see chapter 3, section 2) indicates that a wage reflects not just productive ability but also the content of the tasks an employee must carry out at her workplace: more dangerous, more unstable, and more laborious jobs are offset by higher wages. As these characteristics of jobs are generally poorly measured, it remains possible that the unobserved heterogeneity of jobs does explain wage differences among firms, according to a perfectly competitive logic.

The dispersion of wages may also be the consequence of market imperfections. It may for instance be the consequence of coalitions of employees and employers deciding wages and working conditions jointly, through negotiation. We study this question in chapter 7, dedicated to collective bargaining. In this chapter, we examine the consequences of limited mobility. When there are barriers to mobility, monopsony situations (those in which a single employer confronts a large number of suppliers of labor) can arise (see chapter 12, section 2.2.1). Fundamentally, monopsony is one of the textbook cases in which mobility costs work to the disadvantage of wage earners. If these costs did not exist, the monopsony would be powerless vis-à-vis employees who could quit their jobs at any time. But the converse is just as valid: the costs of hiring, firing, and training are obstacles to mobility of jobs that can be exploited by employees in such a way as to capture a share of the rent. No matter what their source, mobility costs and the rent-sharing that attends them generate wage differentials that are unrelated to productivity differentials and that hinder the efficiency of the competition mechanism. From this viewpoint, job search models generalize the monopsony model: they take into account the costs of mobility, called "friction costs," for employers and for workers. We will see that by making possible a deeper understanding of wage formation, these models have made a profound advance in our conception of how the labor market functions.

4.2 The Equilibrium Search Model

The basic job search model focuses solely on the behavior of job seekers and takes the distribution of wages as given. This approach leaves the setting of wages unexplained. Equilibrium search models have as their goal the explanation of how wages are set through the attribution of well-defined strategic behavior to firms. Labor market equilibrium is therefore characterized by an *endogenous* distribution of wages.

4.2.1 Diamond's Critique

In the job search models presented above, the cumulative distribution function $H(.)$ of wage offers is exogenous. This hypothesis must be abandoned if we wish to understand how wages are determined. To achieve that, we must make explicit the behavior of employers. This poses a conceptual difficulty known as *Diamond's critique*. Diamond (1971) was the first to emphasize that if the reactions of employers are introduced into the basic job search model, the outcome is necessarily a labor market equilibrium in which the distribution of wages is concentrated at a single point. To better understand this result, let us assume that the economy is composed of a large number of identical suppliers of labor and a large number of firms, likewise identical, and let us suppose that workers sequentially receive job offers (i.e., cannot recall former offers if declined) and accept any offers with a wage that is above their reservation wage, as shown above in section 2.1. Since the workers accept *without distinction* all proposals that equal or exceed the reservation wage, the firms gain no advantage by offering wages that exceed it (because as a general rule the profit per capita diminishes with the cost of labor). At equilibrium, the distribution of wages is thus concentrated at value x of the reservation wage. The definition of the reservation wage, which was obtained above in equation (5.6), is:

$$x = z + \frac{\lambda_u}{r+q}\int_x^{+\infty}(w-x)\mathrm{d}H(w) \tag{5.39}$$

where λ_u denotes the arrival rate of job offers and q the job destruction rate, implies that the distribution of wages is thus concentrated at a value equal to the instantaneous gain z of the unemployed workers. This result arises essentially out of the hypothesis that workers never (voluntarily) leave their employers and from the fact that employers have no power over the reservation wage level of workers. Indeed, x does not depend on the wage offered by any firm in particular but on the distribution H on the market. Hence firms have no incentive to post wages superior to the minimum acceptable z, for doing so allows them neither to attract nor to retain more workers. At first sight, Diamond's critique appears to deprive the basic job search model of all its relevance, since within this model we cannot explain why the distribution of wages does not degenerate to a single point.

In reaction to the critiques directed at the basic job search model, *equilibrium* search models have been elaborated, in which the distribution of wages becomes an endogenous variable dependent on, among other things, the wage strategies of employers. An initial approach consists of extending the basic model—often termed the partial model for clarity—by introducing heterogeneity among the workers and assuming that firms post non-renegotiable wages that cannot be contingent on the reservation wages (Albrecht and Axell, 1984). Under certain conditions, labor market equilibrium is compatible with a nondegenerated distribution of wages that coincides with that of the reservation wages of different categories. This means that an individual can encounter a distribution of wages because he is surrounded by individuals with different reservation wages.

4.2.2 The Mean-Min Wage Ratio

The solution to Diamond's critique based on different reservation wages is not totally satisfactory, inasmuch as Hornstein et al. (2011) have shown that the basic search model cannot generate large wage differentials for identical individuals for plausible values

of preference parameters and of labor flows. More precisely, they argue that the basic search model predicts that the ratio between the mean wage and the reservation wage is very small for plausible parameters values. In order to show this, let ρ denote the replacement ratio, equal to the ratio between the instantaneous gains of unemployed workers and the average wage, so that $z = \rho\overline{w}$, where $\overline{w} = \int_x^\infty \frac{w}{1-H(x)} \mathrm{d}H(w)$ stands for the mean wage. Then equation (5.39) can be written:

$$x = z + \frac{\lambda_u^*}{r+q}(\overline{w} - x)$$

where $\lambda_u^* = \lambda_u [1 - H(x)]$ is the job finding rate. This equation yields a relation between the minimum observed wage x and the mean wage, which can be written under the form of the mean-min wage ratio:

$$Mm \equiv \frac{\overline{w}}{x} = \frac{\frac{\lambda_u^*}{r+q} + 1}{\frac{\lambda_u^*}{r+q} + \rho} \tag{5.40}$$

which is a measure of the dispersion of wages. The distribution of wages is degenerated to a single mass point if this ratio is equal to 1. And the bigger this ratio, the wider the wage dispersion. Hornstein et al. (2007) have estimated on U.S. data that the mean-min ratio of the frictional distribution of wages, the distribution of wages for identical workers, belongs to the interval $[1.5, 2]$.

An interesting feature of the mean-min wage ratio defined in equation (5.40) is that it does not depend *directly* on the wage offer distribution H. Accordingly, one does not need to have direct information about this distribution to know its mean-min ratio. The mean-min wage ratio is merely a function of four parameters: the job finding rate, the job separation rate, the discount rate, and the replacement rate. Hornstein et al. (2007) have shown that the mean-min wage ratio is very close to 1 for plausible values of these parameters. Setting the period to one month and calibrating the model on the U.S. economy over the period 1991–2007 implies a monthly job finding rate of 0.43, a monthly job separation rate of 0.03, and a replacement ratio ρ equal to 0.4. Assuming that the annual discount rate is equal to 5%, which implies a monthly discount rate $r = 0.0041$, the mean-min wage ratio would amount to 1.04. Hence the predicted mean-min wage ratio is very close to 1, meaning that the baseline job search model predicts that there is very little wage dispersion.

It is easy to see that this result arises out of the fact that the term $\lambda_u^*/(r+q)$ that appears in the numerator and in the denominator of the expression of the mean-min wage ratio is much bigger than 1 to the extent that the job finding rate is much bigger than the job separation rate; it is about 10 times bigger on most labor markets. This can be seen more clearly if we express the mean-min wage ratio as a function of the unemployment rate, denoted u. Normalizing the size of the labor force to 1, equality between entries into and exits out of unemployment, respectively equal to $q(1-u)$ and $\lambda_u^* u$, implies that $u = q/(q + \lambda_u^*)$. Now the mean-min wage ratio can be written:

$$Mm = \frac{\frac{q}{r+q}\frac{1-u}{u} + 1}{\frac{q}{r+q}\frac{1-u}{u} + \rho} \simeq \frac{\frac{1-u}{u} + 1}{\frac{1-u}{u} + \rho}$$

where the approximation is justified by the fact that the discount rate r is small with respect to the job separation rate, about 10 times smaller. In this expression of the mean-min ratio, we see that the term $(1-u)/u$ goes from 19 to 9 when the unemployment rate goes from 5% to 10%, driving the mean-min wage ratio up from 1.03 to 1.06. Therefore, on average, the mean-min wage ratio predicted by the baseline job search model is in the neighborhood of 1.05, between 10 times and 20 times smaller than the actual mean-min ratio. Hornstein et al. (2011) show that the mean-min wage ratio remains small under a large set of changes in the assumptions of the model, such as the introduction of risk aversion, of compensating wage differential, of return to experience, and of endogenous search effort.

Hornstein et al. (2011) convincingly show that the basic job search model does predict a narrow dispersion of wages. However, they also show that the on-the-job search models, initiated by Burdett and Mortensen (1998), can explain larger wage dispersion. On-the-job search implies that there is competition among employers to attract workers, which alters the way wages are determined. The model that follows sheds light on how this works.

4.2.3 Worker Turnover and Wage Dispersion

We consider an economy composed of a continuum of firms and a continuum of workers. For simplicity, these two continuums are assumed to be of unitary mass. This hypothesis allows us to account simply for the fact that there exists a large given number of firms and workers. The job search behavior of suppliers of labor is identical to that in the model with on-the-job searching studied above, in section 2.2.2. In particular, q always designates the instantaneous job destruction rate, and the parameters λ_u and λ_e represent respectively the arrival rate of job offers for an unemployed job seeker and for one who has a job. The reservation wage of the former, always denoted x, is then given by relation [as shown above in equation (5.20)]:

$$x = z + (\lambda_u - \lambda_e)\int_x^\infty \frac{\bar{H}(\xi)}{r + q + \lambda_e \bar{H}(\xi)} \mathrm{d}\xi \quad \text{with} \quad \bar{H}(\xi) \equiv 1 - H(\xi) \tag{5.41}$$

This equation implicitly defines the reservation wage as a function of the parameters λ_u, λ_e, and the cumulative distribution function H. When $\lambda_e = 0$, that is, when there is no on-the-job search, we come back to the reservation wage of the baseline model. Vice versa, if $\lambda_e > 0$, the job seeker takes account of the possibilities of future income associated with continuing to look for a job while employed. Now and henceforth, contrary to the model of section 2.2.2, the cumulative distribution function $H(.)$ is an endogenous variable.

Let us designate by $G(w)$ the cumulative distribution function of wages among employees, that is, the share of employees paid a wage lower than w. It is important to remark that $G(w)$ is different from $H(w)$, which is also called the *sampling* distribution, the cumulative distribution function of wage offers that job seekers receive. We can get a relation between these two cumulative distribution functions using the equality between exits and entries in jobs that pay wages lower than w. Let us denote by u the unemployment rate. The entries into jobs that pay less than w are composed of unemployed job seekers who have received a wage offer inferior to w. Now at each date an unemployed job seeker receives offers at rate λ_u, and these offers are lower than w with a probability $H(w)$. Entries of unemployed job seekers into jobs offering a wage

lower than w then amount to $\lambda_u u H(w)$. These comprise all the entries into these jobs because we only consider the net number of entries; we do not count the entries and exits into and out of jobs that pay less than w. As regards exits, employment in the jobs paying less than w is equal to $(1-u)G(w)$. These jobs are destroyed at rate q, and employees get job offers at rate λ_e that they accept only if the wage offered is above w, which occurs with probability $1 - H(w) = \bar{H}(w)$. Therefore, at stationary equilibrium, the equality of the flows of entries and exits is given by the following equation:

$$\lambda_u u H(w) = (1-u)G(w)\left[\lambda_e \bar{H}(w) + q\right] \tag{5.42}$$

At stationary equilibrium the value of the unemployment rate results directly from the equality between the flows of workers entering into and exiting from unemployment. The former amounts to $q(1-u)$ and the latter is equal to $\lambda_u[1 - H(x)]u = \lambda_u u$ because the reservation wage x is the lower bound of the sampling distribution to the extent that it cannot be optimal to make wage offers that are always rejected. The stationary unemployment rate is then given by:

$$u = \frac{q}{\lambda_u + q} \tag{5.43}$$

Using equations (5.42) and (5.43), we get:

$$G(w) = \frac{H(w)}{1 + \kappa \bar{H}(w)}, \quad \text{where } \kappa = \frac{\lambda_e}{q} \tag{5.44}$$

This equation implies that $G(w) < H(w)$, when $\lambda_e > 0$. In other words, the probability of occupying a job with a wage less than w is lower than the probability of receiving a job offer with a wage less than w. The distribution of wages among employees thus dominates the distribution of wages offered. This stems simply from the fact that wage earners holding a job accept only wages higher than their current wage. Wage earners holding a job thus obtain higher and higher wages every time they change jobs. Moreover, equation (5.44) shows that the gap between the sampling distribution and the distribution of wages among employees grows with parameter κ, which represents the average number of job offers received in the interval between two job-destroying shocks. Parameter κ is an inverse measure of search frictions, which are small when there are many job offers between two consecutive job-destroying shocks.

Jolivet et al. (2006) have compared the distribution of wages of persons holding jobs with the wages at which unemployed persons are hired in 10 European countries and in the United States. They find that the distribution of the hiring wages of the unemployed is systematically dominated by that of the wages of persons who hold jobs. Figure 5.9 illustrates this phenomenon: it shows that the cumulative distribution functions of the hiring wages of the unemployed are systematically situated below those of the wages of job holders. Of course, the difference between these distributions may be influenced by factors other than job search. In particular, employers may have an interest in offering contracts in which the wage rises with seniority in order to raise the performance of their employees. The lower wages of entrants are then explained by reasons of incentivation, which we will study in chapter 6. It remains the case that the job search model with on-the-job search furnishes a highly convincing explanation of an empirical phenomenon that is well identified.

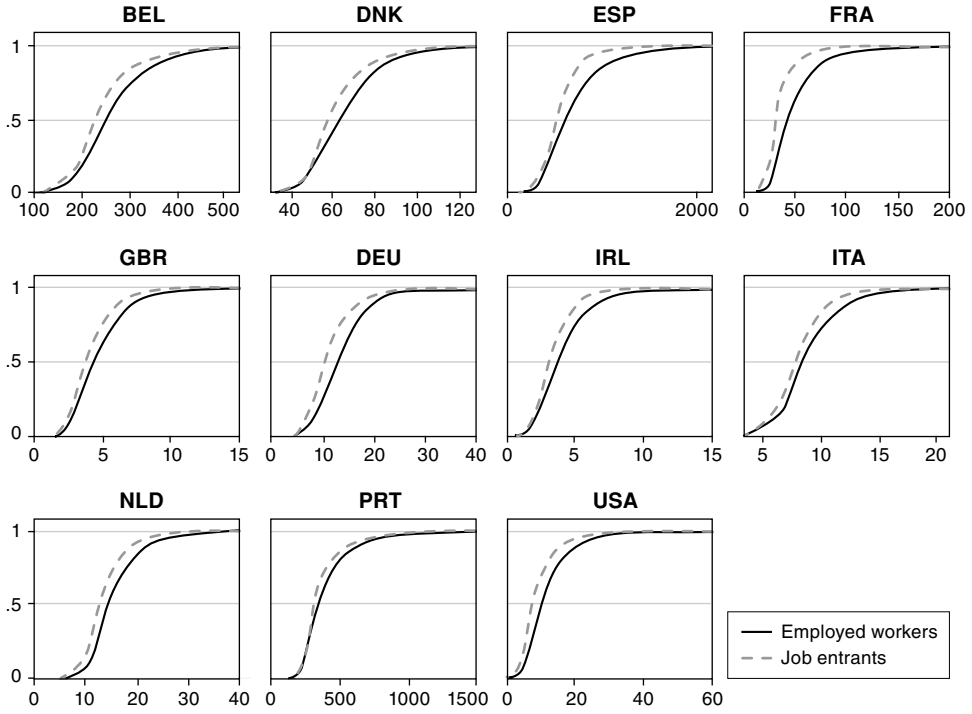


FIGURE 5.9
Cumulative distribution functions of hourly wages (in local currency) in 10 European countries (1994–1997) and in the United States (1993–1996).

Source: Jolivet et al. (2006).

4.2.4 WAGE-POSTING MODELS

To explain wage distribution, it is necessary to specify the manner in which wages are set. Let us begin by presenting the wage-posting model of Burdett and Mortensen (1998), where every firm unilaterally proposes an identical wage for all the workers it hires. In this model, workers of the same productivity who hold jobs of the same degree of difficulty obtain different wages.

The Behavior of Firms

Burdett and Mortensen (1998) assume that firms compete by posting wages to attract workers. At every instant, the probability of encountering a worker is identical for all firms. Each firm decides unilaterally on the *constant* wage that will be paid to its employees. It is thus assumed that the workers at one firm all receive the same wage. It is also assumed that at each moment a worker is capable of producing, if she is employed, a constant exogenous quantity y of goods. If there are $\ell(w)$ workers in a firm that pays wage w, the instantaneous profit of this firm works out to $(y - w)\ell(w)$. For simplicity, we assume that the real rate of interest r is close to 0 (an approximation we can justify by noting that in practice r is clearly smaller than the rates λ_u, λ_e, and q). Under this hypothesis, each firm sets its wage in such a way as to maximize its stationary instantaneous profit $(y - w)\ell(w)$, with the wages being paid in the other firms being

taken as given (so what we have is a noncooperative equilibrium of the Cournot-Nash type). Let us first note that each firm must necessarily propose a wage w higher than the reservation wage x, or $w \geq x$, so as to be able to attract the unemployed at least. The optimal wage is then defined by the equality:

$$\frac{\ell'(w)}{\ell(w)} = \frac{1}{y - w}, \quad w \geq x \tag{5.45}$$

The employment function $\ell(w)$ is obtained using relation (5.45), which characterizes the optimal behavior of a firm. As this relation is true for all wages belonging to the support of H, it can be considered a first-order differential equation in $\ell(w)$. Quantity $\ell'(w)/\ell(w)$ representing the derivative of $\ln \ell(w)$, and the integral of $(y - w)^{-1}$ being equal to $-\ln(y - w)$, this equation is written $\int \ln \ell(w) \mathrm{d}w = -\int \ln (y - w) \mathrm{d}w$, or $\ln \ell(w) = -\ln (y - w) + a$, and then $\ell(w) = \exp(a)/(y - w)$, where a is a constant that does not depend on w. The intuitive view proves correct: employment does indeed increase with wages. We observe as well that the wage is lower than the marginal productivity of work, as in the monopsony model, and that the profits $(y - w)\ell(w)$ of the different firms are all equal to $\exp(a)$ at equilibrium. In other words, there exists a distribution of wages such that at equilibrium firms can realize the same level of profit with low wages and a small workforce, or with high wages and a large workforce. Consequently, firms that pay low wages face a relatively low hiring rate and a relatively high quit rate, which results, at stationary equilibrium, in a small workforce.

The Equilibrium Wage Distribution

The wage-setting policy of firms allows us to obtain a relation between the wage and employment. We can obtain another relation between employment and wages, deriving from worker flows, from equation (5.44). There are $\ell(w)$ employees in each firm that pays wage w and there are $H'(w)$ firms that pay wage w. Therefore, the total mass of employees paid wage w amounts to $H'(w)\ell(w)$. By definition, the mass of employees paid wage w is also equal to $G'(w)(1 - u)$. Thus:

$$G'(w)(1 - u) = \ell(w)H'(w) \tag{5.46}$$

Using this equality with equation (5.44) yields:[19]

$$\frac{\ell'(w)}{\ell(w)} = \frac{2\kappa H'(w)}{1 + \kappa \bar{H}(w)} \tag{5.47}$$

[19] Deriving equation (5.44) with respect to w yields:

$$G'(w) = \frac{H'(w)\,[1 + \kappa G(w)]}{1 + \kappa \bar{H}(w)} \tag{i}$$

Substituting this expression of $G'(w)$ into (5.46) gives:

$$[1 + \kappa G(w)]\,(1 - u) = \ell(w)\left[1 + \kappa \bar{H}(w)\right]$$

The derivation of the logarithm of this equation with respect to w gives:

$$\frac{\kappa G'(w)}{1 + \kappa G(w)} = \frac{\ell'(w)}{\ell(w)} - \frac{\kappa H'(w)}{1 + \kappa \bar{H}(w)}$$

This equation, together with equation (i), implies equation (5.47).

This differential equation implicitly defines the relation between functions $\ell(.)$ and $H(.)$ compatible with equilibrium of flows on the labor market. Comparison of equations (5.45) and (5.47) reveals that distributions of wage offers compatible with both equilibrium of flows on the labor market and strategic behavior by firms in setting wages necessarily satisfy, for any value of w, relation:

$$2(y-w)H'(w)+H(w)=\frac{1+\kappa}{\kappa} \tag{5.48}$$

This equality, which holds for all w, is interpretable as a first-order differential equation in $H(w)$. If A designates any constant, then the general solution of this differential equation is written:[20]

$$H(w)=A\sqrt{y-w}+\frac{1+\kappa}{\kappa}$$

The constant A is obtained using the fact that firms have no interest in offering a wage smaller than the reservation wage x of unemployed job seekers. Thus it is certain that $H(x)=0$. Utilizing this property, we find that the *unique* possible equilibrium wage distribution is expressed by:

$$H(w)=\frac{1+\kappa}{\kappa}\left[1-\sqrt{\frac{y-w}{y-x}}\right] \tag{5.49}$$

The upper bound of the distribution of wages, denoted $w_{\sup}$, satisfies $H(w_{\sup})=1$. It is defined as a function of the reservation wage by the formula:

$$w_{\sup}=y-(y-x)\left(\frac{1}{1+\kappa}\right)^2 \tag{5.50}$$

If the reservation wage is less than the instantaneous production of a worker y (which is a necessary condition of the existence of equilibrium), we can verify that the upper bound $\overline{w}$ of wages is likewise smaller than individual production y. Taking into account (5.49), the equilibrium wage distribution takes the form:

$$H'(w)=\frac{1+\kappa}{2\kappa}\frac{1}{\sqrt{(y-x)(y-w)}} \tag{5.51}$$

The equilibrium density of the sampling distribution $H'(.)$ of this model turns out to *increase* as the level of wages rises. This result is a consequence of both the property that all agents are homogeneous and the firms' strategy of simply proposing an invariable wage. Under these conditions, a firm that raises its wage w increases its volume of employment to the detriment of employment in the other firms. This movement leads to an increasing relation between the wage and the size of the firms.

[20] Recall that the general solution of a linear differential equation is obtained by adding a particular solution to the general solution of the *homogeneous* equation. The latter is written $H'(w)/H(w)=-1/2(y-w)$; it is integrated exactly like equation (5.45), which gives us $H(w)=A\sqrt{y-w}$, where A is an arbitrary constant. We get a particular solution of equation (5.48); by making $H'=0$ in this equation, we immediately find $H(w)=(q+\lambda_e)/\lambda_e$, and from that the general solution of equation (5.48).

All the relations giving the equilibrium values of the endogenous variables of the model depend on the reservation wage x, which is itself an endogenous variable. Now the reservation wage is always defined by equation (5.41) of the partial model, on condition of positing $r = 0$. Taking account of expression (5.49) of the equilibrium wage distribution, it is possible to obtain an explicit analytic form of this wage. After several calculations, we arrive at:

$$x = \frac{z(q + \lambda_e)^2 + (\lambda_u - \lambda_e)\lambda_e y}{(q + \lambda_e)^2 + (\lambda_u - \lambda_e)\lambda_e} \tag{5.52}$$

If there is no possibility of on-the-job search, or $\lambda_e = 0$, we have $x = z$ and, following (5.50), $w_{\text{sup}} = z$. We thus come back to the paradox pointed out by Diamond (1971)—the only possible equilibrium in the partial job search model occurs when the distribution of wages is entirely concentrated at the level of the instantaneous gain z of an unemployed job seeker. When $\lambda_u \to +\infty$, there is no friction in the labor market and the workers obtain the totality of product. The wage is thus uniform and equal to the value of production ($x = w_{\text{sup}} = y$). Searching while working is thus pointless ($\lambda_e = 0$). These characteristics describe a perfectly competitive equilibrium where there is no unemployment ($u = 0$) and where the wage equals the marginal productivity of labor.

The Empirical Implications of the Wage-Posting Model

The version of the wage-posting model we have just presented yields a certain number of pertinent empirical predictions. First, in the equilibrium search model, the wage of an individual employee rises when she moves from one job to another. Although that is not in practice the only reason for individual pay to rise, this phenomenon is in fact observed in the majority of transitions of this type (see for example Topel and Ward, 1992). Moreover, in this model the wage is positively correlated with the size of the firm, which fits well with observations that tell us that even after controlling for the heterogeneity of workers and firms, bigger firms pay higher wages than do smaller ones (Abowd et al., 1999).

Second, wages rise, on average, as workers gain experience. Assuming that new entrants begin as job seekers, the wage at which they are hired is a minimum corresponding to the reservation wage x. After that, their wage rises every time they change firms. More senior employees, who have on average had the most job offers, thus enjoy the highest wages. This prediction of the equilibrium search model agrees with the observation that a worker's wage increases with the time she has spent in the labor market (Abowd et al., 1999).

Third, the lower bound of the equilibrium wage distribution being equal to the reservation wage, an unemployed job seeker accepts all the offers he receives. This conclusion fits very well with that of empirical studies, which do in fact find that the probability of accepting an offer is close to 1.[21]

Fourth, the model with on-the-job search predicts more wage dispersion than the baseline job search model. In particular, the predicted mean-min wage ratio is larger.

[21] See Devine (1988), van den Berg (1990), and Wolpin (1987).

When there is on-the-job search, the mean-min wage ratio can be approximated by the expression (assuming as above that $r \to 0$):[22]

$$Mm \simeq \frac{\frac{\frac{1-u}{u}-\kappa}{1+\kappa}+1}{\frac{\frac{1-u}{u}-\kappa}{1+\kappa}+\rho}$$

This formula shows that the mean-min wage ratio increases with κ, which means that the mean-min ratio increases when there is more competition among firms. Jolivet et al. (2006) provide estimates of parameter κ for 10 European countries and for the United States. They find that κ is between 0.27 (for Portugal) and 2.03 (for France). Assuming that the unemployment rate equals 10% and that $\rho = 0.4$, the mean-min ratio is equal to about 1.2 when κ equals 2. Therefore, the on-the-job search model with wage posting can predict a value of the mean-min wage ratio that is significantly larger than the baseline job search model but that is still below the empirical mean-min wage ratio.

The search equilibrium model does present one major flaw: the density of wage distribution—see (5.51)—is an *increasing* function of the wage. This prediction turns out to conflict with all observations, which reveal that this density is increasing at first, then decreasing, with a maximum generally not too far from the lower bound, as shown by figure 5.10.

To remedy this flaw, one solution lies in introducing heterogeneity among agents. Assume that upon receiving a job offer, workers draw the productivity of the firm from which the offer comes from an exogenous distribution denoted $\Gamma(y)$, where y stands for the time-invariant productivity of each firm. Let us denote $H(w)$ the corresponding sampling distribution of wage offers. From equations (5.43), (5.44), and (5.46) we get:

$$\ell(w) = \frac{\lambda_u(1+\kappa)}{(\lambda_u+q)\left[1+\kappa\bar{H}(w)\right]^2} \tag{5.53}$$

In equilibrium, the sampling distribution of wages and firm types must be identical, $H[w(y)] = \Gamma(y)$, and the firms with the smallest productivity offer workers their reservation wage and hire workers only from the unemployment pool. This distribution is an equilibrium if each firm offers a wage $w(y)$ that maximizes its steady-state profit flow, $\pi(y,w) = (y-w)\ell(w)$, which implies that $\mathrm{d}\pi[y,w(y)]/\mathrm{d}y = \delta\pi[y,w(y)]/\partial y = \ell[w(y)] > 0$. Or, utilizing equation (5.53):[23]

$$\frac{\mathrm{d}\pi[y,w(y)]}{\mathrm{d}y} = \frac{\lambda_u(1+\kappa)}{(\lambda_u+q)\left[1+\kappa\bar{\Gamma}(y)\right]^2}$$

If we assume that there is free entry into the market for goods, the profit of the firm with the weakest productivity is zero, or $\pi(y_{\inf}, x) = 0$, where x designates the reservation wage and $y_{\inf}$ represents the lower bound of the productivity distribution. This condition entails that the firm with the weakest productivity, equal to $y_{\inf}$, hires job seekers by

[22] See Hornstein et al. (2011).

[23] Note that individual firm decisions have no impact on $H[w(y)]$.

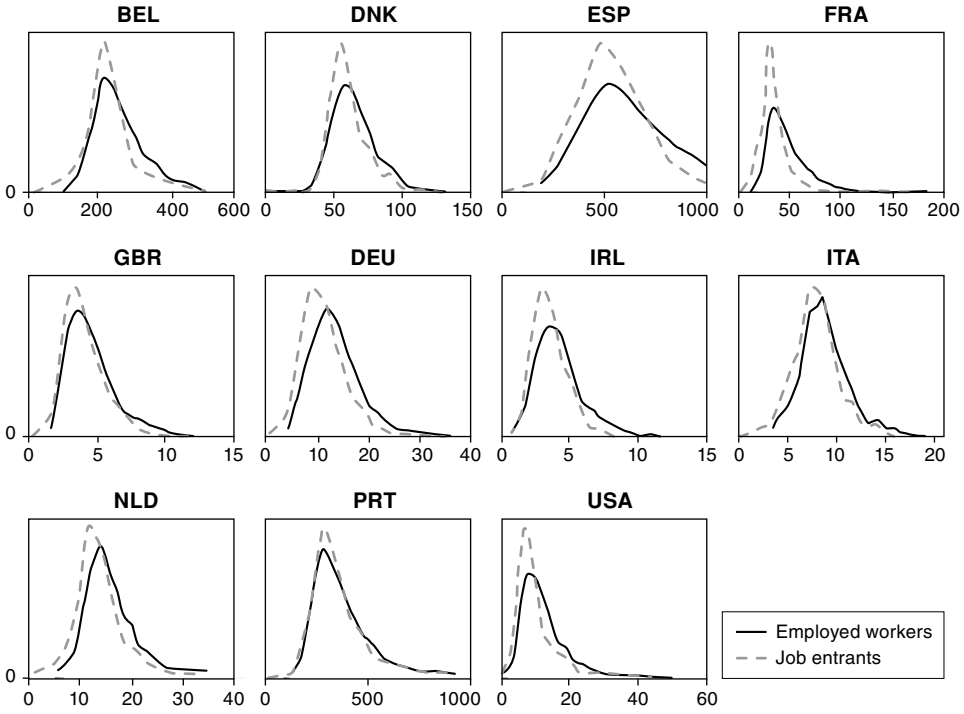


Figure 5.10
Density functions of hourly wages (in national currency) in 10 European countries (1994–1997) and in the United States (1993–1996).

Source: Jolivet et al. (2006).

offering them a reservation wage equal to their productivity, or $x = y_{\text{inf}}$. The integration of the above equation then entails:

$$\pi\left[y, w(y)\right] = \int_x^y \frac{\lambda_u(1+\kappa)}{(\lambda_u + q)\left[1 + \kappa\bar{\Gamma}(\xi)\right]^2} \mathrm{d}\xi$$

and thus, using the definition of profit, $\pi(y, w) = (y - w)\ell(w)$ and equation (5.53):

$$w(y) = y - \left[1 + \kappa\bar{\Gamma}(y)\right]^2 \int_x^y \frac{\mathrm{d}\xi}{\left[1 + \kappa\bar{\Gamma}(\xi)\right]^2} \tag{5.54}$$

This equation shows that, as in the previous model where jobs all have the same productivity, wages are lower than productivity. It also shows that wages rise with productivity whenever persons already in work receive offers, which corresponds to the case in which $\kappa = \lambda_e/q$ is strictly positive. For in the case where the rate at which job holders receive offers is zero, all wages are equal to the reservation wage x, itself equal to the instantaneous gain of job seekers, z.

Several contributions have estimated this model by inferring the distribution of the productivities of firms on the basis of the distribution of wages and transitions between jobs and unemployment (see Bontemps et al. 2000; Postel-Vinay and Robin, 2006; Eckstein and van den Berg, 2007). Bontemps et al. (2000), for example, estimate the cross-sectional distribution of wages, G, on French data, by a nonparametric method. Availing themselves of this distribution, and noting that the hazard rate of an episode of employment is equal to $q\left[1+\kappa\bar{H}(w)\right]$, with $\bar{H}(w)=(1+\kappa)G(w)/[1+\kappa G(w)]$, they can then estimate parameter κ by maximizing the likelihood of hazard rates on the labor market.[24] Finally, knowledge of the distribution H, of κ, and of wages w makes it possible to estimate the inverse wage function, meaning the distribution of productivities as a function of wages, which, in line with equation (5.54), takes the expression:

$$y(w)=w+\left[1+\kappa\bar{H}(w)\right]^2\int_x^w\frac{\mathrm{d}\xi}{\left[1+\kappa\bar{H}(\xi)\right]^2}$$

Comparison of the distribution of productivities thus obtained with that directly estimated on the basis of firm data shows that the wage-posting model predicts a distribution of the productivities of firms with an excessively high proportion of high-productivity firms (the "tail" of the density function is too thick). This result comes from the fact that the wage-posting model artificially limits competition among firms, since they are barred from making counteroffers. This hypothesis, which limits the market power of wage earners, entails that in order to explain high wages productivities must be higher than those observed in the data.

4.2.5 Sequential Auctions and Bargaining

The wage-posting model assumes a limited degree of competition among firms, to the extent that an employer whose employee receives an offer higher than his current wage cannot, by hypothesis, make a counteroffer. Postel-Vinay and Robin (2002), Dey and Flinn (2005), and Cahuc et al. (2006) have set forth models of sequential auction and bargaining where firms can make such counteroffers. We will see that these models make it possible to explain the distribution of wages better than models of wage posting do.

To make the equilibrium job search model more empirically pertinent, Postel-Vinay and Robin (2002) elaborate a model in which a firm can make a counteroffer when its worker receives an offer from a competing firm. In this setting, wage earners have more market power than in the wage-posting model because they can drive employers into direct competition. Postel-Vinay and Robin assume that firms make take-it-or-leave-it offers to workers. The productivities of firms and the situation of each worker are observed by all.

Hence, when a firm encounters a job seeker, the firm offers him his reservation wage, which he accepts, as in the wage-posting model. When two firms are in contact with the same worker, they simultaneously make him take-it-or-leave-it offers. Such is the case when a job holder paid at wage w in a firm of productivity y (with $w \leq y$) receives

[24] The maximum likelihood method within the framework of duration models is presented in section 3.2.

an offer from a firm of productivity y'. A range of outcomes are possible, depending on the amount of wage w and the productivities y and y'.

1. If $y < y'$ the competing firm, of productivity y', offers the lowest wage possible that allows it to attract the worker, while the firm of productivity y offers the highest wage possible that is compatible with nonnegative profits, to try to retain the worker. The firm of productivity y thus offers a wage equal to y. At equilibrium the worker is hired by the firm with the highest productivity, where she obtains a wage that makes her indifferent between staying with the type y firm at wage y or receiving wage $\omega(y, y')$ in the type y' firm. If we denote by $V_e(w, y)$ the expected discounted utility of a worker paid at wage w in a type y firm, the wage $\omega(y, y')$ obtained in the new type y' firm verifies $V_e[\omega(y, y'), y'] = V_e(y, y)$.

2. When $y' \leq y$, two possibilities must be considered, depending on whether the firm currently employing the worker is compelled to make a counteroffer in order to retain her.
   a. If $V_e(w, y) > V_e(y', y')$, the firm of productivity y' can offer at most wage y', but the expected utility of a worker who accepted that offer would be inferior to what she would get by staying with her current employer. So she stays with the productivity y firm and keeps wage w.
   b. If $V_e(w, y) < V_e(y', y')$, the productivity y firm can retain its employee by offering her a wage $\omega(y', y) > w$ such that $V_e[\omega(y, y'), y] = V_e(y', y')$. Thus the worker obtains a wage rise without changing firms.

Thus, in the sequential auction model, the offers of competing employers permit wage earners to obtain wage rises without changing firms. Furthermore, in this model workers can change firms while accepting drops in their wage. A worker paid at wage w can move from a productivity y firm to a productivity $y' > y$ firm where the wage is $\omega(y, y') < w$ since the higher productivity of the new firm may allow her to obtain larger wage rises in the future. With this in mind, the sequential auction model permits us to explain why a nonnegligible percentage of wage earners (from 20% to 35% depending on the country in the data of Jolivet et al., 2006) do change jobs with no interval of unemployment, but they do so at a reduced wage.

Cahuc et al. (2006) have introduced bargaining into the sequential auction model by making the assumption that workers can counter the offers of firms with propositions of their own. Bargaining gives workers the possibility to obtain higher wages than they can in the base model of sequential auction.[25] Estimation of the model of Cahuc et al. (2006) on French data for the period 1993–2000 indicates that low-skilled workers have very weak bargaining power, not significantly differing from zero, whereas more highly skilled workers have positive bargaining power.

[25] Bargaining models are presented in chapter 7.

By giving more market power to wage earners than the wage-posting model gives, sequential auction models with bargaining allow us to better replicate the empirical distribution of wages than the wage-posting model. These models also generate a wider dispersion of wages than the wage-posting model, to the extent that wages are heterogeneous within each firm. Papp (2013) has shown that the model of Cahuc et al. (2006) predicts values of the mean-min wage ratio that are compatible with empirically derived values, which lie between 1.5 and 2. Overall then, on-the-job search models with sequential auction and bargaining have the capacity to produce good empirical predictions.

An operational description of the labor market would also require that parameters λ_u, λ_e, and q describing a worker's transitions between different possible states be made endogenous. In particular, the job offers arrival rates depend on the number of vacant jobs and the number of job seekers—quantities that derive from the behavior of firms and the way in which wages are set. The job destruction rate q is in all likelihood influenced by variations in productivity and by the way wages are set. The matching models, which we develop in chapter 9, partly fill these gaps.

5 SUMMARY AND CONCLUSION

- Job search theory assumes that individuals know only the distribution of wages existing in the economy and that they must search in order to encounter employers who will make them definite wage offers. The optimal strategy for a job seeker consists of accepting any wage offer higher than his or her *reservation wage*. The latter depends on the set of parameters affecting the labor market, in particular the job destruction rate, the arrival rate of job offers, and unemployment insurance benefits.
- To get unemployment insurance benefits, one must in general have worked previously and contributed to an unemployment insurance fund for a specified period. One is then eligible for unemployment insurance benefits. A rise in the level of benefits increases the duration of unemployment for eligible job seekers but diminishes that of ineligible job seekers.
- Empirical studies of the determinants of the exit rates from unemployment generally utilize duration models, which explain the amount of time passed in a certain state—for example the length of unemployment spells—as a function of institutional data and the characteristics of a sample of individuals followed over a certain period. The estimation of these models poses problems linked in particular to the specification of the functions defining the exit rates from unemployment and the existence of censored data.
- Empirical studies show that the reservation wage and the average length of an unemployment spell are sensitive to the amount of unemployment insurance benefits. The elasticity of the duration of unemployment with respect to the replacement ratio varies between 0.4 and 1.6. An increase in the potential duration of benefits increases the effective duration of an unemployment spell by around 20% of the increase in the potential duration. Much of the effect exerted by an increase in the period of benefit payment is due to easing of liquidity constraints.

- Unemployment insurance influences the arrival rates of job offers by its effect upon the intensity and efficiency of the job search carried out by job seekers receiving benefits.
- Providing that the data are disaggregated to a sufficient degree, unobserved individual heterogeneity explains the core of interindustry wage differences. Yet the share of wage differences explained by the heterogeneity of firms remains substantial. In France, for example, the average of the differences in wage paid to an identical worker employed in two different firms lies in the 20% to 30% range but does not exceed the 2% to 3% range from one industry to another.
- Because it integrates the strategic behavior of firms, the equilibrium search model is characterized by an endogenous distribution of wages. It offers the advantage of explaining the wage-setting process and thus making possible the analysis of the overall effects of economic policy. The equilibrium search model also highlights the importance of taking into account on-the-job search with sequential auctions and bargaining when it comes to explaining the wage distribution. This model explains why the wage of an individual employee can increase or decrease when she moves from one job to another, why wages rise, on average, as workers gain experience, and why large firms pay higher wages than small firms to identical workers.

6 RELATED TOPICS IN THE BOOK

- Chapter 1, section 1: The reservation wage and the choice between consumption and leisure
- Chapter 3, section 2: Compensating wage differentials and the hedonic theory of wages
- Chapter 8, section 2: Theories of discrimination
- Chapter 9, section 3: The matching model
- Chapter 10, section 1: Technological progress and unemployment
- Chapter 12, section 2.2.2: Minimum wage in labor market with frictions
- Chapter 13, section 1: Unemployment insurance
- Chapter 14, section 2.1: Manpower placement services
- Chapter 14, section 3: Evaluation of labor market policies
- Chapter 14, section 4.2: Job search assistance and monitoring

7 FURTHER READINGS

Eckstein, Z., & van den Berg, G. (2007). Empirical labor search: A survey. *Journal of Econometrics*, *136*(2), 531–564.

Krueger, A., & Mueller, A. (2012). The lot of the unemployed: A time use perspective. *Journal of the European Economic Association*, *10*(4), 765–794.

Lalive, R., van Ours, J., & Zweimüller, J. (2006). How changes in financial incentives affect the duration of unemployment. *Review of Economic Studies*, *73*, 1009–1038.

Mortensen, D., & Pissarides, C. (1999). New developments in models of search in the labor market. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 3B, chap. 39). Amsterdam: Elsevier Science.

Postel-Vinay, F., & Robin, J.-M. (2006). Microeconometric search-matching models and matched employer-employee data. In R. Blundell, W. Newey, & T. Persson (Eds.), *Advances in economics and econometrics theory and applications*, Ninth World Congress (pp. 279–310). Cambridge, U.K.: Cambridge University Press.

Tatsiramos, K., & van Ours, J. (2012). Labor market effects of unemployment insurance design. IZA discussion paper no. 6950, forthcoming in *Journal of Economic Surveys*.

van den Berg, G. (2001). Duration models: Specification, identification and multiple durations. In J. Heckman & E. Leamer (Eds.), *Handbook of econometrics* (vol. 5, chap. 55). Amsterdam: Elsevier Science.

REFERENCES

Abbring, J., & van den Berg, G. (2003). The identifiability of the mixed proportional hazards competing risks model. *Journal of the Royal Statistical Society*, Series B, *65*, 701–710.

Abbring, J., van den Berg, G., & van Ours, J. (2005). The effect of unemployment insurance sanctions on the transition rate from unemployment to employment. *Economic Journal*, *115*, 602–630.

Abowd, J. M., Kramarz, F., Lengermann, P., McKinney, K., & Roux, S. (2013). Persistent inter-industry wage differences: Rent sharing and opportunity costs. *IZA Journal of Labor Economics*, *1*(7).

Abowd, J. M., Kramarz, F., Lengermann, P., & Roux, S. (2003). Interindustry and firm-size wage differentials in the United States and France (Working Paper). Cornell University.

Abowd, J., Kramarz, F., & Margolis, D. (1999). High wage workers and high wage firms. *Econometrica*, *67*, 251–334.

Abowd, J., & Zellner, A. (1985). Estimating gross labor force flows. *Journal of Business and Economic Statistics*, *3*, 254–283.

Albrecht, J., & Axell, B. (1984). An equilibrium model of search employment. *Journal of Political Economy*, *92*, 824–840.

Atkinson, A., & Mickelwright, J. (1991). Unemployment compensation and labor market transitions: A critical review. *Journal of Economic Literature*, *29*, 1679–1727.

Belzil, C. (2001). Unemployment insurance and subsequent job duration: Job matching versus unobserved heterogeneity. *Journal of Applied Econometrics*, *16*, 619–636.

Blundell, R., & Costa Dias, M. (2009). Alternative approaches to evaluation in empirical microeconomics. *Journal of Human Resources*, *44*(3), 565–640.

Bontemps, C. (1998). *Modèles de recherche d'emploi d'équilibre*. Thèse pour le doctorat en sciences économiques. Université de Paris I.

Bontemps, C., Robin, J.-M., & van den Berg, G. (2000). An empirical job search model with search on the job and heterogeneous workers and firms. *International Economic Review*, *41*(2), 305–358.

Boone, J., Fredriksson, P., Holmlund, B., & van Ours, J. (2007). Optimal unemployment insurance with monitoring and sanctions. *Economic Journal*, *117*, 399–421.

Boone, J., Sadrieh, A., & van Ours, J. (2009). Experiments on unemployment benefit sanctions and job search behavior. *European Economic Review*, *53*(8), 937–951.

Boone, J., & van Ours, J. (2006). Modelling financial incentives to get the unemployed back to work. *Journal of Institutional and Theoretical Economics*, *162*, 227–252.

Burdett, K., & Mortensen, D. (1998). Wage differentials, employer size, and unemployment. *International Economic Review*, *39*, 257–273.

Burgess, P., & Kingston, J. (1976). The impact of unemployment insurance benefits on reemployment success. *Industrial and Labor Relations Review*, *30*, 25–31.

Cahuc, P., & Le Barbanchon, T. (2010). Labor market policy evaluation in equilibrium: Some lessons of the job search and matching model. *Labour Economics*, *17*, 196–205.

Cahuc, P., Postel-Vinay, F., & Robin, J.-M. (2006). Wage bargaining with on-the-job search: Theory and evidence. *Econometrica*, *74*(2), 323–364.

Card, D., Chetty, R., & Weber, A. (2007a). Cash-on-hand and competing models of intertemporal behavior: New evidence from the labor market. *Quarterly Journal of Economics*, *122*, 1511–1560.

Card, D., Chetty, R., & Weber, A. (2007b). The spike at benefit exhaustion: Leaving the unemployment system or starting a new job?. *American Economic Review*, *97*, 113–118.

Chetty, R. (2008). Moral hazard versus liquidity and optimal unemployment insurance. *Journal of Political Economy*, *116*(2), 173–234.

Centeno, M. (2004). The match quality gains from unemployment insurance. *Journal of Human Resources*, *34*, 839–863.

Centeno, M., & Novo, A. (2009). Reemployment wages and UI liquidity effect: A regression discontinuity approach. *Portuguese Economic Journal*, *8*, 45–52.

Cox, D. (1975). Partial likelihood. *Biometrika*, *62*(2), 269–276.

Danforth, J. (1979). On the role of consumption and decreasing absolute risk aversion in the theory of job search. In S. Lippman & J. McCall (Eds.), *Studies in the economics of search* (pp. 109–131). New York, NY: Elsevier.

Devine, T. (1988). Arrival versus acceptance: The source of variation in reemployment rates across demographic groups (Working Paper). Pennsylvania State University.

Devine, T., & Kiefer, N. (1991). *Empirical labor economics: The search approach*. New York, NY: Oxford University Press.

Dey, M., & Flinn, C. (2005). An equilibrium model of health insurance provision and wage determination. *Econometrica*, *73*(2), 571–627.

Diamond, P. (1971). A model of price adjustment. *Journal of Economic Theory*, *3*, 156–168.

Dickens, W., & Katz, L. F. (1987). Inter-industry wage differences and industry characteristics. In K. Lang & J. Leonard (Eds.), *Unemployment and the structure of labor markets* (pp. 48–89). New York, NY: Basil Blackwell.

Dormont, B., Fougère, D., & Prieto, A. (2001). The effect of the time profile of unemployment insurance benefits on exit from unemployment (CREST Working Paper). Paris, www.crest.fr.

Eckstein, Z., & van den Berg, G. (2007). Empirical labor search: A survey. *Journal of Econometrics*, *136*(2), 531–564.

Ehrenberg, R., & Oaxaca, R. (1976). Unemployment insurance, duration of unemployment, and subsequent wage gain. *American Economic Review*, *66*, 754–766.

Gautier, P., Muller, P., van der Klaauw, B., Rosholm, M., & Svarer, M. (2012). Estimating equilibrium effects of job search assistance (IZA Discussion Paper No. 6768).

Gibbons, R., & Katz, L. (1992). Does unmeasured ability explain inter-industry wage differentials. *Review of Economic Studies*, *59*, 515–535.

Goux, D., & Maurin, E. (1999). Persistence of interindustry wage differentials: A reexamination using matched worker-firm panel data. *Journal of Labor Economics*, *17*, 492–533.

Graversen, B., & van Ours, J. (2008). How to help unemployed find jobs quickly: Experimental evidence from a mandatory activation program. *Journal of Public Economics*, *92*, 2020–2035.

Hairault, J.-O., Langot, F., & Sopraseuth, T. (2010). Distance to retirement and older workers' employment: The case for delaying the retirement age. *Journal of the European Economic Association*, *8*(5), 1034–1076.

Holzer, H. (1986). Reservation wages and their labor market effects for black and white male youth. *Journal of Human Resources*, *21*, 157–177.

Hornstein, A., Krusell, P., & Violante, G. (2007). Frictional wage dispersion in search models: A quantitative assessment (NBER Working Paper No. 13674).

Hornstein, A., Krusell, P., & Violante, G. (2011). Frictional wage dispersion in search models: A quantitative assessment. *American Economic Review*, *101*, 2873–2898.

Hunt, J., (1995). The effect of unemployment compensation on unemployment duration in Germany. *Journal of Labor Economics*, *13*, 88–120.

Jolivet, G., Postel-Vinay, F., & Robin, J.-M. (2006). The empirical content of the job search model: Labor mobility and wage distributions in Europe and the US. *European Economic Review*, *50*, 877–907.

Jones, S., & Riddell, C. (1999). The measurement of unemployment: An empirical approach. *Econometrica*, *67*, 142–167.

Kaplan, E., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, *53*, 457–481.

Katz, L., & Summers, L. (1989). Industry rents: Evidence and implications. *Brookings Papers on Economic Activity: Microeconomics* (pp. 209–275). Washington, DC: Brookings Institution Press.

Kiefer, N. M. (1988). Economic duration data and hazard functions. *Journal of Economic Literature*, *26*(2), 646–679.

Kiefer, N., & Neumann, G. (1993). Wage dispersion with homogeneity: The empirical equilibrium search model. In Bunzel et al. (Eds.), *Panel data and labour market analysis*. Amsterdam: North-Holland.

Krueger, A., & Mueller, A. (2010). Job search and unemployment insurance: New evidence from time use data. *Journal of Public Economics*, *94*(3–4), 298–307.

Krueger, A., & Mueller, A. (2011). Job search, emotional well-being and job finding in a period of mass unemployment: Evidence from high-frequency longitudinal data. *Brookings Papers on Economic Activity*, *42*(1), 1–81.

Krueger, A., & Mueller, A. (2012). The lot of the unemployed: A time use perspective. *Journal of the European Economic Association*, *10*(4), 765–794.

Krueger, A., & Summers, L. (1988). Efficiency wages and the inter-industry wage structure. *Econometrica*, *56*, 259–293.

Kuhn, P., & Mansour, H. (2011). Is Internet job search still ineffective? (IZA Discussion Paper No. 5955).

Kuhn, P., & Skuterud, M. (2004). Internet job search and unemployment durations. *American Economic Review*, *94*(1), 218–232.

Lalive, R., van Ours, J., & Zweimüller, J. (2005). The effect of benefit sanctions on the duration of unemployment. *Journal of the European Economic Association*, *3*(6), 1386–1417.

Lalive, R., van Ours, J., & Zweimüller, J. (2006). How changes in financial incentives affect the duration of unemployment. *Review of Economic Studies*, *73*, 1009–1038.

Lammers, M. (2012). The effects of savings on reservation wages and search effort (Nestspar Discussion Paper, 01/2012-023).

Lancaster, T. (1979). Econometric methods for the duration of unemployment. *Econometrica*, *47*(4), 939–956.

Lentz, R., & Tranaes, T. (2005). Job search and savings: Wealth effects and duration dependence. *Journal of Labor Economics*, *23*, 467–489.

Lynch, L. (1983). Job search and youth unemployment. *Oxford Economic Papers*, *35*, 271–282.

McCall, J. (1970). Economics of information and job search. *Quarterly Journal of Economics*, *84*, 113–126.

Meyer, B. (1990). Unemployment insurance and unemployment spells. *Econometrica*, *58*(4), 757–782.

Micklewright, J., & Nagy, G. (2010). The effect of monitoring unemployment insurance recipients on unemployment duration: Evidence from a field experiment. *Labour Economics*, *17*, 180–187.

Moorthy, V. (1989). Unemployment in Canada and the United States: The role of unemployment insurance benefits. *Federal Reserve Bank of New York Quarterly Review*, Winter, 48–60.

Mortensen, D. (1970). Job search, the duration of unemployment, and the Phillips curve. *American Economic Review*, *60*, 505–517.

Mortensen, D. (1986). Job search and labor market analysis. In O. Ashenfelter & R. Layard (Eds.), *Handbook of labor economics* (vol. II, pp. 849–919). Amsterdam: Elsevier Science.

Mortensen, D., & Pissarides, C. (1999). New developments in models of search in the labor market. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 3B, chap. 39). Amsterdam: Elsevier Science.

Murphy, K., & Topel, R. (1987). The evolution of unemployment in the United States: 1968–1985. In *NBER Macroeconomics Annual 1987* (vol. 2, pp. 11–68). National Bureau of Economic Research.

OECD. (1994). *The OECD jobs study*. Paris: OECD Publishing.

Papp, T. (2013). Frictional wage dispersion with Bertrand competition: An assessment. *Review of Economic Dynamics*, *16*(3), 540–552.

Pellizzari, M. (2006). Unemployment duration and the interactions between unemployment insurance and social assistance. *Labour Economics*, *13*(6), 773–798.

Postel-Vinay, F., & Robin, J.-M. (2002). The distribution of earnings in an equilibrium search model with state-dependent offers and counter-offers. *International Economic Review*, *43*(4), 989–1016.

Postel-Vinay, F., & Robin, J.-M. (2006). Microeconometric search-matching models and matched employer-employee data. In R. Blundell, W. Newey, & T. Persson (Eds.). *Advances in economics and econometrics theory and applications*, Ninth World Congress (pp. 279–310). Cambridge, U.K.: Cambridge University Press.

Slichter, G. (1950). Notes on the structure of wages. *Review of Economics and Statistics*, *32*, 80–91.

Stigler, G. (1961). The economics of information. *Journal of Political Economy*, *69*, 213–225.

Stigler, G. (1962). Information in the labor market. *Journal of Political Economy*, *70*, 94–105.

Tatsiramos, K., & van Ours, J. (2012). Labor market effects of unemployment insurance design. IZA discussion paper no. 6950, forthcoming in *Journal of Economic Surveys*.

Topel, R., & Ward, M. (1992). Job mobility and the careers of young men. *Quarterly Journal of Economics*, *107*(2), 439–479.

van den Berg, G. (1990). Nonstationarity in job search theory. *Review of Economic Studies*, *57*, 255–277.

van den Berg, G. (1999). Empirical inference with equilibrium search models of the labour market. *Economic Journal*, *109*, 283–306.

van den Berg, G. (2001). Duration models: Specification, identification and multiple durations. In J. Heckman & E. Leamer (Eds.), *Handbook of econometrics* (vol. 5, chap. 55). Amsterdam: Elsevier Science.

van den Berg, G., & Ridder, G. (1998). An empirical equilibrium search model of the labor market. *Econometrica*, *66*, 1183–1221.

van den Berg, G., & van der Klaauw, B. (2006). Counseling and monitoring of unemployed workers: Theory and evidence from a controlled social experiment. *International Economic Review*, *47*, 895–936.

van den Berg, G., van der Klaauw, B., & van Ours, J. (2001). Punitive sanctions and the transition from welfare to work (Working Paper). Tinbergen Institute.

van den Berg, G., & Vikström, J. (2009). Monitoring job offer decisions, punishments, exit to work, and job quality (IZA Discussion Papers No. 4325).

van den Berg, G., & van Ours, J. (1994). Unemployment dynamics and duration dependence in France, the Netherlands and the UK. *Economic Journal*, *104*, 432–443.

van Ours, J., & Vodopivec, M. (2006). How shortening the potential duration of unemployment benefits entitlement affects the duration of unemployment: Evidence from a natural experiment. *Journal of Labor Economics*, *24*, 351–378.

Venn, D. (2012). Eligibility criteria for unemployment benefits: Quantitative indicators for OECD and EU countries (OECD Social, Employment and Migration Working Paper No. 131).

Wolpin, K. (1987). Estimating a structural job search model: The transition from school to work. *Econometrica*, *55*, 801–818.

CHAPTER 6

Contracts, Risk-Sharing, and Incentive

In this chapter we will:

- Probe the reasons that firms and workers engage in long-term relationships
- See how the trade-off between insurance and incentive acts upon the remuneration rule for labor
- Investigate why firms make use of hierarchical promotions and internal markets
- Learn what the deferred payment mechanism is
- Discover what the efficiency wage theory has to tell us
- Look at how social preferences interact with incentives

INTRODUCTION

Within firms, those who manage human resources evidently have a toolkit of varied measures at their disposal (Oyer and Schaefer, 2011). The strategic variables currently used to optimize the return to labor include promotion, bonuses, profit sharing, status distinctions, quality circles, investment in training, and dismissal (see Lazear, 2011). Such a toolkit leads us to ask what form an optimal remuneration rule for labor would take. A priori, a remuneration rule ought to be based on the complete array of information available to both sides—primarily the results of the employees' activity and observation of the environment in which this activity takes place. The theory of contracts explains how technology and the preferences of actors both influence the choice of strategies for managing human resources (for complete presentations, see Salanié, 1997; Malcomson, 1999; Prendergast, 1999; and Bolton and Dewatripont, 2005). To be more precise, this theory analyzes how contractual relations allow two different types of problem to be managed: the *uncertainty of the environment* and the *private nature of certain information* concerning the activities and the performance of workers.

The wage relationship is often a long-term one, which takes concrete form when a "labor contract" is signed. Curiously, this type of contract very often specifies only rights and duties of a purely *formal* nature, without always linking remuneration explicitly

to performance. Simon (1951) had already noted this essential difference between an ordinary contract of sale and purchase and a labor contract governing a hierarchical relationship. Above all, a labor contract betokens a relationship of *subordination*, meaning that an employer and an employee have agreed that the latter will exercise his profession under the authority of the former. It may also set out the length of time this agreement will last and the amount of remuneration to be paid. This amount very often depends on criteria like seniority that do not, at first sight, appear to have much to do with individual performance, but Doeringer and Piore (1971) have drawn attention to the fact that large firms set up "internal markets" that function according to a logic very different from that of a competitive market of the kind described in chapter 3, where the remuneration of workers hinges on their productivity. A priori, though, it would seem to be more efficient to pay an employee according to the tasks that she effectively carries out, in other words, to pay her a wage corresponding to her output—the system known as "piece rate." In reality the modes of remuneration vary widely. Freeman and Rogers (1999) estimate that only around 45% of workers in the private sector in the United States in 1998 were receiving a remuneration that partly depended on their own performance or on that of their firm, through some type of collective profit sharing or employee stock ownership plan. These authors highlight the fact that during the 1990s the way workers were paid in the United States tended to shift toward remuneration linked to performance (which for that matter has been a contributory cause of widening income inequality; see Lemieux et al., 2009). So the pay employees receive is made up of some combination (the weighting varies) of time wages, piecework, stock ownership, and collective profit sharing. The recent survey of Bloom and Van Reenen (2011) confirms the figures of Freeman and Rogers. They estimate that the different studies on this topic situate the percentage of employees in the United States covered by one form or another of performance pay between 40% and 50% by the 2000s. Our aim in this chapter is to show that problems of incentive and risk-sharing play a determining role in how these components are weighted.

In section 1, key concepts relating to the labor contract are defined, with emphasis on the distinction between a *verifiable* element and one that is merely *observable*. What makes an element merely observable is that there could, in the nature of things, be no impartial judge who could verify any observation that might be made of it. Section 2 concentrates on contractual relationships when the actions of wage earners are verifiable by an impartial judge and the economic environment is uncertain; in this situation, the labor contract proves useful as an efficient way to share risk. Section 3 analyzes the labor relationship when the actions of wage earners are not verifiable but the result of them is. This context permits us to understand incentive problems and the linkage between a worker's pay and the results of her activity; in particular, it specifies how punishments, rewards, and the tailoring of individual remunerations come into play as means of incentive. Section 4 deals with incentive problems in informational structures in which neither actions nor results are verifiable. Such is generally the case for workers performing complex tasks. In this setting, there is no point in making a contract that stipulates a remuneration based on performance, since the latter cannot be verified by a third party should a dispute arise. As we will see, the impossibility of verifying the actions and performances of agents explains two elements frequently encountered in systems of remuneration: wages that rise with seniority and systems of internal

promotion.[1] Finally, section 5 attacks the question of motivations other than financial. A great deal of research carried out over more than 20 years highlights the importance of "social preferences" in the behavior of agents placed in situations of exchange with other agents (see the survey of Rebitzer and Taylor, 2011). This research establishes that most individuals attach importance to equity and reciprocity and are sensitive to social norms as well as to their own self-image and the image they wish to project to others. Purely financial incentives may miss their mark if they fail to take into account the social preferences of the agents at whom they are targeted.

1 THE LABOR CONTRACT

The features of labor contracts depend, to a large extent, on whether the results of an employee's activity can be observed and taken into account. These results can only appear explicitly in the contract if they are verifiable. If they are not, the work relationship is governed by implicit and self-enforcing clauses.

1.1 Explicit and Implicit Clauses

To set up a system of remuneration based on observed results is to presume that the latter can be established beyond dispute. This is why the terms used in drawing up the contractual document properly speaking are called *verifiable* elements. Under this heading are grouped all the parameters capable of being objectively assessed by an impartial court. For example, if the contract specifies the exact amount of wage to be paid, the task of checking to see whether this amount was indeed paid (by examining bank accounts, pay slips, and so on) can actually be assigned to a third party. The notion of a verifiable clause contains the idea that should a dispute arise, one of the parties would be able to supply *proof*, in the juridical sense of the term, sufficient to settle the matter.

As a general rule, there exist numerous parameters that could never be assessed with sufficient precision by an impartial tribunal. Such phenomena, opaque to third-party scrutiny, are "unverifiable." The results of collective or individual activities usually fall into this category because it is difficult to furnish real proof that what was accomplished fell short of what was intended. Hence parameters of this type will not appear in the contractual document. Note that the possibility of verifying the values of the parameters of a contract is not really tied to the possibility of observing these values. In fact the shared observation of parameters has only limited importance whenever no third party can certify what has been observed.

We can now state precisely the definition of *explicit* and *implicit* clauses (Carmichael, 1989). Analysis of the verifiable character of the wording of a contract

[1] Firms and their wage earners sometimes enter into contracts the purpose of which is to protect, or to make possible, certain investments, for example investments in training. These contracts pose specific problems having to do with the fact that one of the parties could capture all the benefits from the investment without necessarily having to bear the costs. This question, known as the "hold up" problem, is dealt with in chapters 7 and 14.

allows us, in theory, to place it in one or the other of these categories. All the clauses of an *explicit contract* being verifiable, they will appear in black and white in the text of the agreement, as will the penalties arising from their violation. For that matter, the existence of these penalties ensures that in the great majority of cases explicit contracts are respected. The case is different, however, with the clauses of an *implicit contract*. Since they are not verifiable, there is no reason why they should appear in any written document, which amounts to saying that in this situation, there is no contract in the juridical sense of the term.

1.2 Complete and Incomplete Contracts

The theoretical literature also adopts the terms *complete* and *incomplete* contract to distinguish between explicit and implicit contracts (Hart and Holmström, 1987). By definition, a contract is complete when it is possible, *at the moment of signing*, to foresee all the circumstances that could arise while it is in effect and to set out verifiable clauses for each of them. A complete contract thus comes to the same thing as an explicit contract. Conversely, an incomplete contract does not take some of these eventual circumstances into account. It is curtailed in this way for several reasons. First, the possible circumstances might simply be too numerous, and some of them highly improbable. The "production" costs of the contract (legal advice, preliminary study, the actual drafting, etc.) would outweigh the benefits to be derived. Second, certain circumstances cannot be verified, in which case there is no point in including them in the contract. Finally, a contract that aimed to use all the available information in an optimal manner might lead to clauses or rules of application that would outstrip the cognitive capacity of one of the partners. It would then be necessary to adopt simplified rules, and in that case one might also take the view that the contract was incomplete. In sum, there is no real difference between the definition of an incomplete contract and that of an implicit contract. The notion of "unverifiable clause" encompasses all the reasons for which a contract may be incomplete (for work on the links between the incompleteness of contracts and the assumption of rational behavior, see Hart and Moore, 1999, and Maskin and Tirole, 1999).

The impossibility of having a third party verify individual performance has at least two important consequences when an employee and an employer wish to enter into a long-term relationship. In the first place, it becomes pointless to describe in minute detail the tasks the employee will be expected to carry out. In reality, the labor contract most often takes the form of a relationship of subordination that simply acknowledges the employer's authority and sets out a specified amount of remuneration. It is not generally possible to know in advance what services will be supplied in return for the wage. In the second place, if this relationship is extended, that means the two parties have a mutual interest. The contract is then said to be *self-enforcing*. As the celebrated expression (apparently coined by Okun, 1981) goes, an implicit contract then takes the form of an "invisible handshake."

Having set out the various possible categories of contract, our next step will be to highlight the main properties of optimal contracts. These will necessarily differ according to the explicit or implicit nature of the contract. In studying these problems, the so-called agency model supplies a framework both inclusive and rigorous and has gradually come to dominate the literature.

1.3 The Agency Model

The agency model—also called the principal–agent model—analyzes the problems arising from the working out of contracts between two actors: the principal and the agent. In labor economics, the principal is the employer and the agent the employee. Confining the analysis to just two protagonists at this stage makes it possible to highlight a number of instructive traits, as the reader will see. More sophisticated models study the interactions among a larger number of actors (see Salanié, 1997; Bolton and Dewatripont, 2005).

The agency model assumes that the principal proposes a contract, which the agent can either accept or refuse. This reductive hypothesis allows the bargaining problem to be disposed of rapidly and lets us focus on analyzing the way the structure of information influences the characteristics of contracts (the theory of bargaining is set out in chapter 7). It is important to note that such an assumption makes no commitment as to whether labor market competition is perfect or imperfect. The only thing determined by the nature of labor market competition is the level of satisfaction the employer must offer the worker for the contract to be acceptable. For example, if the market is perfectly competitive, free entry entails zero profit, and the principal will necessarily have to offer a level of satisfaction that procures him zero profit; otherwise, the worker will turn to another employer.

The information available to each party and the degree to which it can be verified influence the properties of contracts offered by the employer. Here we can set out two textbook cases: in the first, the employee's effort is observed by both parties and is verifiable. Though the effort can be verified, the employee's output might be affected by contingencies unforeseen at the time the contract was signed. So both sides are faced with a problem of *risk-sharing*. The contract proposed by the employer then sets out the optimal division of risk and maximizes his expected profit. In the second case, the employee's effort is not verifiable, and the employer is faced with a problem of *moral hazard*. He must propose a contract that gives the worker an incentive to supply maximum effort at minimal cost.

As these observations show, the aim of labor contracts is to manage two types of problem: that of risk-sharing and that of incentive. We will study them in that order in the following sections.

2 RISK-SHARING

Risk-sharing between employers and workers has already been mentioned above to account for the rigidity of real wages. Empirical studies do in fact show that the real wage fluctuates less than production, employment, or hours worked, and it is clearly procyclical (see Abraham and Haltiwanger, 1995, and chapter 9, section 6, below).

These stylized facts do not fit well with a purely competitive determination of wages when the only contracts in existence are those made in a "spot market." Such contracts define the level of transactions in all foreseeable situations but include no provision for insurance. In the model of perfect competition laid out in chapter 3 (in which the labor market is represented as functioning solely on the basis of spot market contracts), variations in productivity lead to proportional variations in the wage. To grasp

this, let us take the case of an agent coming into such a market; this agent's preferences are represented by a quasi-concave utility function $U(C,L)$ where C and L designate respectively the agent's consumption of goods and her leisure. Here we assume that consumption of goods is identical to the agent's remuneration W, and that her leisure is equal to the difference between total endowment of time L_0, the duration of which is normalized to 1, and hours of work, denoted by h.

We will further assume that the production y of the agent depends on hours worked h and on a random variable ε according to function $f(h,\varepsilon)$ increasing in both its arguments and such that, on one hand, marginal productivity is decreasing $f_{hh} \leq 0$ and on the other, marginal productivity is strictly increasing with shock ε, which amounts to hypothesizing $f_{h\varepsilon} > 0$. Under these conditions, the profit Π of an employer is defined by the equality $\Pi = f(h,\varepsilon) - W$, and the zero profit condition then entails $W = f(h,\varepsilon)$.

The determination of work schedules and remuneration is represented in figure 6.1 for two values of ε, denoted ε_1 and $\varepsilon_2 > \varepsilon_1$, in the hours-wage plane. Each worker is able to force each employer to bid against the others and, therefore, to choose a combination of work schedule and remuneration that maximizes her utility subject to the zero profit condition. In graphic terms, remuneration and work schedule are determined, in every state of nature, by the tangent point between curve $f(h,\varepsilon)$ and an indifference curve. In figure 6.1, we observe that variations in real wages are greater than variations in hours worked if the elasticity of the labor supply is weak with respect to wages. Empirical studies have regularly found that the labor supply is weakly elastic to the real wage (see chapter 1). So the model of a perfectly competitive spot market predicts that productivity shocks lead to greater variations in wage than in work schedules, something that empirical observation contradicts (see Rogerson and Shimer, 2011).

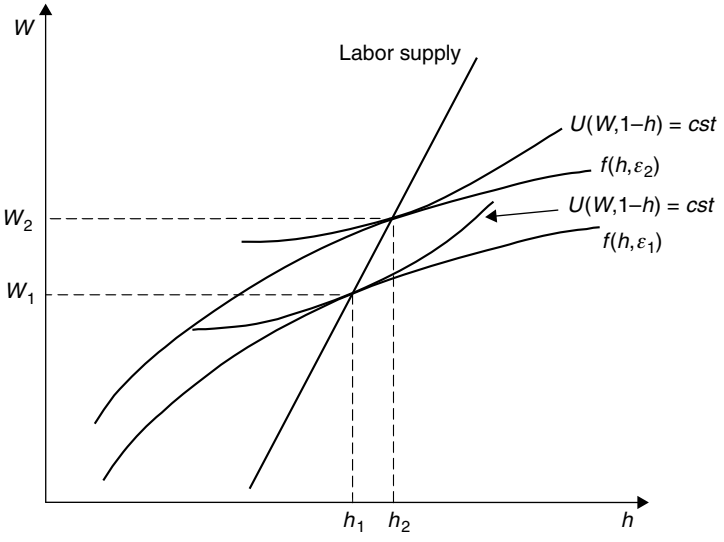


FIGURE 6.1
Wages and hours worked in a perfectly competitive spot market.

Note: *cst* stands for "constant."

These limitations of the spot market model suggest that the demand for insurance may play a role in determining wages. The earliest models in this field came from Baily (1974), Gordon (1974), and Azariadis (1975). They are set in an environment in which the performance of workers is verifiable. These models explain, in a highly satisfactory manner, the rigidity of real wages, but prove to be of little use in understanding underemployment and unemployment. Subsequent research, which we will present later, has explored the consequences of the absence of verifiability of individual performance. Although this research has helped economists to understand certain characteristics of labor contracts, it has not allowed them to establish that the insurance motive constitutes an important source of persistent unemployment (for a simple presentation of the main results achieved regarding insurance contracts, see Azariadis and Stiglitz, 1983; for more comprehensive overviews, see Rosen, 1985; Malcomson, 1999).

2.1 Symmetric or Verifiable Information

The first studies of the consequences of the demand for insurance assumed that individual performance was verifiable and showed how risk-sharing between an employer, who can diversify his assets on the financial markets, and an employee, whose access to such markets is limited, damps down the fluctuations in real wages. This approach also agrees with the most detailed empirical characteristics of wage formation: the well-established facts that the real wage of an employee is strongly correlated with the lowest rate of unemployment registered since the time she began her current job, and that it depends hardly at all on the current unemployment rate or the rate that existed at the time she was hired (Beaudry and DiNardo, 1991). We will see how, by taking labor mobility and insurance mechanisms into account, we can explain facts of this kind.

2.1.1 An Individual Insurance Contract Model

In what follows, we work with a model of an individual contract much like the models used by Green and Kahn (1983), Chari (1983), and Cooper (1983). It is a principal–agent model, in which the employer proposes a contract that the employee can only accept or reject.

Preferences and Technology

We retain the hypotheses adopted already in looking at a competitive spot market: the agent's preferences are represented by a quasi-concave utility function $U(C, L)$ where C and L designate respectively the agent's consumption of goods and her leisure. Here we assume that consumption of goods is identical to the agent's remuneration W and that her leisure is equal to the difference between total endowment of time L_0, the duration of which is normalized to 1, and hours of work, denoted h. The production y of the agent depends on hours worked h and on a random variable ε according to function $f(h, \varepsilon)$ increasing in both its arguments and satisfying $f_{hh} \leq 0$ and $f_{h\varepsilon} > 0$. The random variable ε is a continuous variable defined over the interval $\mathcal{E} = [\varepsilon^-, \varepsilon^+]$, the density of which is denoted $g(\varepsilon)$. The profit Π of the principal is defined by the equality $\Pi = f(h, \varepsilon) - W$. Finally, we do not a priori exclude the possibility that the principal may be risk-averse, and we will denote the utility function of the von Neuman–Morgenstern type representing his preferences by $v(\Pi)$, with $v' > 0$ and $v'' \leq 0$.

In this subsection, the observation of the random variable ε will be assumed to be *verifiable*.[2] Under this assumption, the literature on contracts habitually speaks of *symmetric* information to signify that the principal and agent have access to the same information and that neither one can manipulate it because it is verifiable by a court. An insurance contract $\mathcal{A} = \{W(\varepsilon), h(\varepsilon)\}$ then specifies ex ante, in other words, *before* knowing of the advent of the shock, the remuneration $W(\varepsilon)$ to be received by the agent and the hours of work $h(\varepsilon)$ that she must supply, whatever value of ε may be observed. An insurance contract is a *contingent* contract that takes into account all possible states of nature.

The Principal's Problem and the First-Order Conditions

The principal chooses a contract $\mathcal{A}$ that maximizes his expected utility and that offers the agent an earnings prospect at least equal to what she could obtain elsewhere. Let $\overline{U}$ be the expected utility corresponding to external opportunities, and $\Pi(\varepsilon)$ the profit $f[h(\varepsilon), \varepsilon] - W(\varepsilon)$ when the value of the shock is equal to ε; the optimal contract is the solution of the problem:

$$\max_{\mathcal{A}} \mathbb{E}v\left[\Pi(\varepsilon)\right] \tag{6.1}$$

Subject to the *participation constraint*:

$$\mathbb{E}U\left[W(\varepsilon), 1 - h(\varepsilon)\right] \geq \overline{U} \tag{6.2}$$

Let λ be the multiplier associated with this constraint; the Lagrangian of the principal's problem is written:

$$\mathcal{L} = \mathbb{E}v\left\{f\left[h(\varepsilon), \varepsilon\right] - W(\varepsilon)\right\} + \lambda\left\{\mathbb{E}U\left[W(\varepsilon), 1 - h(\varepsilon)\right] - \overline{U}\right\}$$

The first-order conditions are obtained by setting the derivatives of this Lagrangian to zero with respect to $h(\varepsilon)$ and $W(\varepsilon)$ for all values of ε. Let U_C and U_L be the partial derivatives of function $U(C, L)$ and let f_h be the marginal productivity of hours worked. We thus have:

$$\frac{\partial \mathcal{L}}{\partial h(\varepsilon)} = g(\varepsilon)\left\{f_h\left[h(\varepsilon), \varepsilon\right] v'\left[\Pi(\varepsilon)\right] - \lambda U_L\left[W(\varepsilon), 1 - h(\varepsilon)\right]\right\} = 0, \quad \forall \varepsilon \in \mathcal{E}$$

$$\frac{\partial \mathcal{L}}{\partial W(\varepsilon)} = g(\varepsilon)\left\{-v'\left[\Pi(\varepsilon)\right] + \lambda U_C\left[W(\varepsilon), 1 - h(\varepsilon)\right]\right\} = 0, \quad \forall \varepsilon \in \mathcal{E}$$

[2] Note that this hypothesis is satisfied if we assume that function f, hours h, and performance y are verifiable, since $y = f(h, \varepsilon)$.

If we eliminate the multiplier λ between these two equations, we see that the optimal contract is characterized by the following system:

$$\frac{U_L[W(\varepsilon), 1-h(\varepsilon)]}{U_C[W(\varepsilon), 1-h(\varepsilon)]} = f_h[h(\varepsilon), \varepsilon], \quad \forall \varepsilon \in \mathcal{E} \tag{6.3}$$

$$\lambda U_C[W(\varepsilon), 1-h(\varepsilon)] = v'[\Pi(\varepsilon)], \quad \forall \varepsilon \in \mathcal{E} \tag{6.4}$$

Relation (6.3) shows that the marginal rate of substitution between consumption and leisure is equal to the marginal productivity of labor. So the insurance contract yields a *Pareto efficient* allocation and can be described as a *first-best* contract. Relation (6.4) determines optimal risk-sharing; it entails:

$$\frac{U_C[W(\varepsilon), 1-h(\varepsilon)]}{U_C[W(\theta), 1-h(\theta)]} = \frac{v'[\Pi(\varepsilon)]}{v'[\Pi(\theta)]}, \quad \forall (\varepsilon, \theta) \in \mathcal{E}^2$$

What we have here is the Arrow-Borch condition, well known in insurance theory (see, for example, Laffont, 1989) and according to which the sharing of risk is optimal when the marginal rate of substitution of a gain—measured by the marginal utility of consumption—in state ε for a gain in state θ, is the same for the principal and for the agent.

2.1.2 THE PROPERTIES OF THE OPTIMAL CONTRACT

Let us first take the most common case, in which the principal is supposedly risk neutral ($v'' = 0$) because of his opportunity to diversify risk in a perfect financial market. As the reader can ascertain, differentiating the system (6.3) and (6.4) with respect to ε leads to the following comparative statics properties:

$$\left(\frac{U_{CL}^2 - U_{CC}U_{LL}}{U_C U_{CC}} - f_{hh}\right)\frac{\mathrm{d}h}{\mathrm{d}\varepsilon} = f_{h\varepsilon} \quad \text{and} \quad \frac{\mathrm{d}W}{\mathrm{d}\varepsilon} = \frac{U_{CL}}{U_{CC}}\frac{\mathrm{d}h}{\mathrm{d}\varepsilon} \tag{6.5}$$

It is evident as well that the first-best contract prescribes a wage independent of states of nature if the marginal utility of consumption is independent of hours worked ($U_{CL} = 0$). The functioning of a labor market with insurance contracts is thus very different from that of a spot market, in which the wage is highly sensitive to variations in productivity for empirically relevant values of labor supply elasticity. This was the result obtained by the early work of Baily (1974), Gordon (1974), and Azariadis (1975). It suggests that an employer who has low aversion to risk has a tendency to insure his employees by paying them a remuneration little dependent on the economic trend.

If we assume that the utility function is concave—which implies $U_{CC} < 0$ and $(U_{CL})^2 - U_{CC}U_{LL} < 0$—and that $f_{h\varepsilon} > 0$, it is clear that hours worked is an increasing function of the level ε of the random factor. This conclusion fits well with empirical observations, according to which hours worked rise when the economic trend turns up. However, Rosen (1985) and Malcomson (1999), among others, have pointed out that equations (6.3) and (6.4), which describe the optimal contract, also have some unconvincing implications: the direction in which remuneration W varies depends on the sign

of U_{CL}, which is not a priori determined, so the model does not succeed in reproducing the procyclicity of wages unambiguously. It is easy to verify, moreover, that the utility of the agent diminishes with ε if leisure is a normal good. Since $\mathrm{d}U = U_C \mathrm{d}W - U_L \mathrm{d}h$, we find with the help of relations (6.3) and (6.5) that the derivative $\mathrm{d}U/\mathrm{d}\varepsilon$ is of the sign of $(U_{CC}U_L - U_{CL}U_C)$. This quantity is negative if leisure is a normal good (see chapter 1, appendix 7.2). This is because hours worked increase. Hence, in adopting the usual hypothesis that leisure is a normal good, the model predicts that the agent's satisfaction *diminishes* when productivity increases (and even that her remuneration falls if $U_{CL} > 0$, which is also the prevalent hypothesis).

For the remuneration to be increasing unambiguously with ε, it would be necessary to adopt more restrictive hypotheses, for example, that the principal displays risk aversion ($v'' < 0$) and that hours worked take only two values, $h > 0$, and 0 (which amounts to supposing that the individual labor supply is inelastic). Under these hypotheses, differentiating the risk-sharing relation (6.4) implies $\mathrm{d}W/\mathrm{d}\varepsilon > 0$. But Malcomson (1999) points out that this relation also implies that remuneration and profit always vary in the same direction, something that is not verified for certain categories of workers. Finally, it should be noted that if the principal is risk neutral ($v'' = 0$) and hours worked still take no more than two values, relation (6.4) implies a *constant* wage that does not depend on productivity ε. The principal insures the agent perfectly against fluctuations in her income, which does not fit well with the procyclicity of the real wage.

2.1.3 Insurance and Labor Mobility

According to the foregoing model, wages depend solely on conditions prevailing in the labor market at the time the contract is signed (conditions summed up by the parameter $\overline{U}$ representing the expected utility offered by external opportunities at that time). Real wages are not, therefore, correlated with the state of the labor market during the period covered by the contract, which clashes with the conclusions of Beaudry and DiNardo (1991, 1995), according to which the real wage is significantly correlated with the lowest rate of unemployment recorded from the time the contract began. Beaudry and DiNardo take the view that the model yields this bad prediction because of an implicit and quite groundless hypothesis—that the cost of mobility is prohibitive once a contract is signed. In what follows, we construct a model excluding this hypothesis and, as we will show, it really does match the stylized facts better. This model is a simple, stationary version of the models of Harris and Holmström (1982) and Beaudry and DiNardo (1991). Unlike these authors, we assume that the distribution of shocks is stationary and that shocks are not autocorrelated.

A Model with Labor Mobility

We illustrate the effect of taking labor mobility into account in a simplified model in which individuals, with lifetimes of infinite length, discount the future at the rate $\delta \in (0, 1)$. We assume that length of time worked h is a variable that can take only one value if agents decide to work. The instantaneous utility $U(W, 1 - h)$ of an employee is then denoted by $U(W)$ and, without any loss of generality, we assume that the agent's production per unit of time, $f(h, \varepsilon)$, is simply equal to ε. At the beginning of each unit of time, productivity ε takes a value obtained by a random draw from a distribution $G(\varepsilon)$ assumed to be stationary. For the sake of simplicity, the employer is assumed to be risk

neutral. In this context, we can verify that the previous model, with no labor mobility, entails a constant wage independent of productivity ε. As we will see, such is not the case when mobility is taken into account.

To introduce labor mobility simply, we assume that when the state of nature ε comes about, the agent has the opportunity to quit the firm she is with and work h hours externally, which in that period procures for her the gain $U[\overline{W}(\varepsilon)]$. In this expression, $\overline{W}(\varepsilon)$ designates the outside wage, which is assumed to be increasing with ε; this conveys the notion that an upturn in the economic trend makes itself felt throughout the economy. Opportunities outside the contract then offer the agent an expected discounted present value $\overline{V}(\varepsilon) = U[\overline{W}(\varepsilon)] + \delta\mathbb{E}\overline{V}(\theta)$. In consequence, the expected discounted present value obtained in a firm offering a contract $\mathcal{A} = \{W(\varepsilon)\}$, amounts to $V(\varepsilon) = U[W(\varepsilon)] + \delta\mathbb{E}\max\left[V(\theta), \overline{V}(\theta)\right]$. Labor mobility forces the employer to offer a contract satisfying a *participation* constraint, which is written $V(\varepsilon) \geqslant \overline{V}(\varepsilon), \forall\varepsilon$. Let $\overline{U} = \mathbb{E}U[\overline{W}(\varepsilon)]$; for any contract satisfying the participation constraints, the definitions of $\overline{V}(\varepsilon)$ and $V(\varepsilon)$ imply the equalities $\mathbb{E}V(\varepsilon) = \mathbb{E}U[W(\varepsilon)]/(1-\delta)$ and $\mathbb{E}\overline{V}(\varepsilon) = \overline{U}/(1-\delta)$. In consequence, the participation constraints $V(\varepsilon) \geqslant \overline{V}(\varepsilon)$ are written:[3]

$$U[W(\varepsilon)] + \frac{\delta}{1-\delta}\mathbb{E}U[W(\theta)] \geq U[\overline{W}(\varepsilon)] + \frac{\delta}{1-\delta}\overline{U}, \quad \forall\varepsilon \tag{6.6}$$

The left side of this inequality represents the agent's expected utility if the state of nature ε occurs when the contract $\mathcal{A}$ applies, while the right side represents the expected utility that she would get by quitting the firm where contract $\mathcal{A}$ is in force. If (6.6) is satisfied, the agent never has an interest in leaving her firm, whatever the state of nature that occurs may be. Taking the expectation of both sides of inequality (6.6), we observe that the "global" participation constraint, that is, $\mathbb{E}U[W(\theta)] \geq \overline{U}$, is satisfied if inequality (6.6) is satisfied for all ε.

The principal, henceforth assumed to be risk neutral, chooses a contract that maximizes his expected gains, $\mathbb{E}[\varepsilon - W(\varepsilon)]/(1-\delta)$, taking into account participation constraints (6.6) for all possible values of ε. Let $\lambda(\varepsilon)$ be the multiplier associated with constraint (6.6) when state ε occurs. The Lagrangian of the principal's problem is defined by:

$$\mathcal{L} = \int \frac{[\varepsilon - W(\varepsilon)]}{1-\delta} g(\varepsilon)\mathrm{d}\varepsilon + \int \lambda(\varepsilon)\left\{U[W(\varepsilon)] + \frac{\delta}{1-\delta}\int U[W(\theta)]g(\theta)\mathrm{d}\theta - U[\overline{W}(\varepsilon)] - \frac{\delta}{1-\delta}\overline{U}\right\} g(\varepsilon)\mathrm{d}\varepsilon$$

The first-order conditions are found by setting the derivatives of this Lagrangian to zero with respect to $W(\varepsilon)$. We thus get:

$$\frac{\partial\mathcal{L}}{\partial W(\varepsilon)} = g(\varepsilon)\left\{-\frac{1}{1-\delta} + \lambda(\varepsilon)U'[W(\varepsilon)] + \frac{\delta}{1-\delta}U'[W(\varepsilon)]E\lambda(\theta)\right\} = 0$$

[3] For the contract to be self-enforcing, it would also have to include the possibility that the principal could break it in certain states of nature. We examine the consequences of this eventuality below.

The optimal contract is thus characterized by the following equality:

$$[(1-\delta)\lambda(\varepsilon) + \delta\mathbb{E}\lambda(\theta)]\, U'[W(\varepsilon)] = 1 \tag{6.7}$$

This equation differs from equation (6.4) describing risk-sharing in the model without mobility, which prescribed a constant wage with a risk-neutral principal and a production function $f(h,\varepsilon)$ additively separable with respect to h and ε. Here, wage $W(\varepsilon)$ depends on the state of nature ε through multiplier $\lambda(\varepsilon)$, which is not a priori a constant.

Properties of Contractual Wages

We can set out certain characteristics of contractual wages in detail by considering the set Λ^+ of states of nature for which the participation constraints are necessarily binding. Formally, this set is defined by $\Lambda^+ = \{\varepsilon \,|\, \lambda(\varepsilon) > 0\}$. Let us assume that this set is not empty and consider two states ε_1 and ε_2 which belong to this set and are such that $\varepsilon_1 > \varepsilon_2$. For these values ε_1 and ε_2, the constraints (6.6) are equalities. If we subtract these equalities side by side, it becomes evident that $U[W(\varepsilon_1)] - U[W(\varepsilon_2)]$ is equal to $U\left[\overline{W}(\varepsilon_1)\right] - U\left[\overline{W}(\varepsilon_2)\right]$. Now the last expression is positive since $\varepsilon_1 > \varepsilon_2$ and outside wages $\overline{W}(\varepsilon)$ are increasing with ε. We can state, therefore, that optimal wages $W(\varepsilon)$ are likewise increasing with ε over the set Λ^+. Since the agent's risk aversion dictates $U'' < 0$, it results, following risk-sharing relation (6.7), that the multipliers $\lambda(\varepsilon)$ are likewise increasing with ε over the set Λ^+. Let ε_λ then be the smallest value of ε for which we have $\lambda(\varepsilon) > 0$. The previous line of reasoning proves that the set Λ^+ is also characterized by $\Lambda^+ = \{\varepsilon \,|\, \varepsilon \geq \varepsilon_\lambda\}$. Conversely, we can deduce that we have $\lambda(\varepsilon) = 0$ for all $\varepsilon < \varepsilon_\lambda$.

The first-order condition (6.7) then shows that the contractual wage is constant for all $\varepsilon < \varepsilon_\lambda$. Conversely, when $\varepsilon \geq \varepsilon_\lambda$, the participation constraint is binding and the contractual wage $W(\varepsilon)$ is defined by the equality:

$$U[W(\varepsilon)] = U[\overline{W}(\varepsilon)] - \frac{\delta}{1-\delta}\left\{\mathbb{E}U[W(\theta)] - \overline{U}\right\}$$

Since $\mathbb{E}U[W(\theta)] \geq \overline{U}$, we observe that the contractual wage $W(\varepsilon)$ is less than the outside wage $\overline{W}(\varepsilon)$ for $\varepsilon \geq \varepsilon_\lambda$. To summarize, the participation constraints are binding in the "good" states of nature ($\varepsilon \geq \varepsilon_\lambda$) with wages that are increasing but less than outside wages, while in the "bad" states of nature ($\varepsilon < \varepsilon_\lambda$) the wage is constant and the participation constraints are not necessarily binding. These properties[4] of contractual wages are shown in figure 6.2.

Under these same hypotheses, the model without mobility entailed a constant wage in all states of nature. This wage ought then to be correlated solely with the state of the economy at the moment the contract is signed, which contradicts the results of Beaudry and DiNardo (1991). On the other hand, if it is possible for employees to leave their firms, the model shows that the contractual wage is no longer a constant and that it rises when the economic trend turns up. This conclusion does agree with that of Beaudry and DiNardo, which brings out a positive correlation between the contractual

[4]The horizontal part of the profile of contractual wages necessarily intersects with the curve representing the outside wage; otherwise, the contract would offer a gain inferior to outside opportunities.

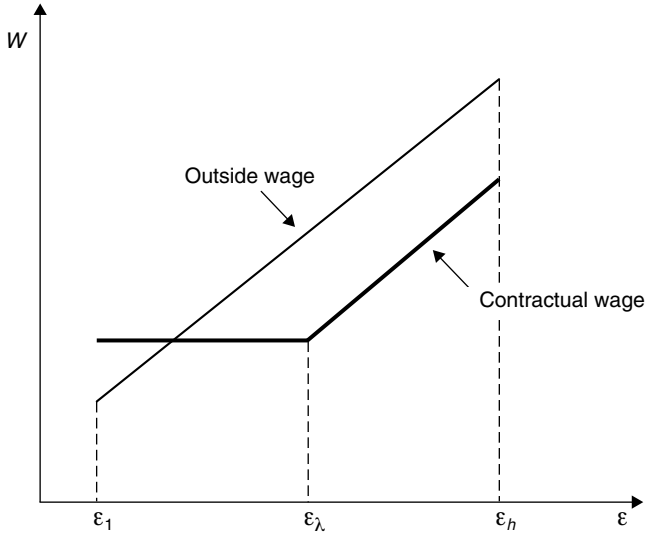


Figure 6.2
The wage contract with labor mobility.

wage and the weakest unemployment rate (assumed to appear for a productivity situated in "good" states of nature) since the beginning of the contract. It should be noted that if we had assumed that the principal was also able to break the contract, the contractual wage would not have been completely rigid downwards. There would have been "very bad" states of nature for which the principal's participation constraint would have been binding, which would have entailed a wage flexibility downwards for these states; and this would have been a better fit with the fluctuations in the real wage over the cycle (see Thomas and Worral, 1988, for a model that takes this possibility into account). Finally, we must also note that in the model with labor mobility, fluctuations in the contractual wage are damped down in comparison to those in the outside wage (see figure 6.2). This result also agrees with the observed fluctuations in the real wage over the course of a cycle, which do appear to be damped in comparison to labor productivity (in our model, the outside wage can be likened to a competitive spot wage perfectly correlated with the marginal productivity of labor). Taking mobility into account in a model with symmetric information thus gives a better explanation of certain stylized facts. It now remains to examine the effects of information asymmetry.

2.2 Asymmetric or Unverifiable Information

We come back to the static model without labor mobility; the assumption will now be that the observed values of the random factor ε are not verifiable. The literature on insurance contracts most often employs the term *asymmetric information* to describe this situation. This means that one of the actors—for our purposes, the principal—observes the true values of the shocks. The main thing to remember, though, is that this is a situation in which it is impossible, or very costly, to have the actually occurring values of the random variable ε verified by an impartial third party. From this perspective,

the terminology adopted by the literature using the agency model certainly has greater clarity. It uses the term *hidden information* to indicate that performances are unverifiable (see Salanié, 1997). With this hypothesis, a contract can no longer be simply a series of values of remuneration and effort indexed to future values of ε, for the principal will sometimes have an interest in claiming that the value of ε that applies is not the one that actually occurs. The "revelation principle" of Myerson (1979) makes it possible nonetheless to arrive at a characterization of optimal contracts.

2.2.1 The Revelation Principle

The revelation principle states that to each contract one can link another contract that is incentive compatible and that entails the same allocation. The revelation principle is worth our attention because it limits the search for optimal contracts to the set of contracts for which the principal does declare the true state of nature.

The Incentive-Compatible Contract

To show that to any contract $A = \{W(\varepsilon), h(\varepsilon)\}$ one can link another contract that is incentive compatible and that entails the same allocation, we start by showing that one can link an incentive-compatible contract to contract A. We then proceed to show that this incentive compatible contract yields the same allocation as contract A.

Let us imagine that the agent and the principal have agreed on a contract $\mathcal{A} = \{W(\varepsilon), h(\varepsilon)\}$; when the principal observes the value ε of the random shock, her interest is to declare that she has observed the state of nature $m(\varepsilon)$ which, given this contract $\mathcal{A}$, procures her the greatest possible profit. Formally, the state $m(\varepsilon)$ is defined by the equality:

$$m(\varepsilon) = \arg\max_{\theta} \{f[h(\theta), \varepsilon] - W(\theta)\} \tag{6.8}$$

Let us now consider contract $\widehat{\mathcal{A}} = \{\widehat{W}(\varepsilon), \hat{h}(\varepsilon)\}$ where $\widehat{W}(\varepsilon) = W[m(\varepsilon)]$ and $\widehat{h}(\varepsilon) = h[m(\varepsilon)]$. Compared to contract $\mathcal{A}$, contract $\widehat{\mathcal{A}}$ presents the advantage of being *incentive compatible*; in other words, if it is in force, the principal always has an interest in revealing the *true* state of nature. To demonstrate this result, let us suppose that $\widehat{\mathcal{A}}$ is in force and the principal declares that she has observed the state of nature θ, whereas the true state is ε. The profit attained by adopting this attitude is defined by the identity:

$$f\left[\widehat{h}(\theta), \varepsilon\right] - \widehat{W}(\theta) \equiv f\{h[m(\theta)], \varepsilon\} - W[m(\theta)] \tag{6.9}$$

Now, using definition (6.8) of signal $m(.)$, the right-hand side of this last equality satisfies:

$$f\{h[m(\theta)], \varepsilon\} - W[m(\theta)] \leq \max_{s} \{f[h(s), \varepsilon] - W(s)\} \equiv f[\widehat{h}(\varepsilon), \varepsilon] - \widehat{W}(\varepsilon) \tag{6.10}$$

Finally, relations (6.9) and (6.10) entail:

$$f[\widehat{h}(\theta), \varepsilon] - \widehat{W}(\theta) \leq f[\widehat{h}(\varepsilon), \varepsilon] - \widehat{W}(\varepsilon)$$

This inequality signifies that the principal makes less profit by "lying," that is, by announcing θ, than she does by revealing the true state ε of nature. So the principal does indeed have an interest in revealing the true state of nature when contract $\widehat{\mathcal{A}}$ is in force.

The revelation principle also entails that contracts $\mathcal{A}$ and $\widehat{\mathcal{A}}$ lead to the *same* allocation of resources. If ε comes about and contract $\mathcal{A}$ is in force, the principal declares that state $m(\varepsilon)$ has come about, which means that the agent must work $h[m(\varepsilon)]$ in exchange for compensation $W[m(\varepsilon)]$. On the other hand, if it is contract $\widehat{\mathcal{A}}$ that is in force, the principal announces the true state ε of nature since $\widehat{\mathcal{A}}$ is incentive compatible. The agent must then work $\widehat{h}(\varepsilon) \equiv h[m(\varepsilon)]$ and receives wage $\widehat{W}(\varepsilon) \equiv W[m(\varepsilon)]$, so the allocation of resources is identical under contracts $\mathcal{A}$ and $\widehat{\mathcal{A}}$. Since it is possible to link any contract to an incentive-compatible contract that leads to the same allocation of resources, the revelation principle implies that the search for the optimal contract can be confined to the set of incentive-compatible contracts. In practice, the optimal contract is the solution of problem (6.1) of maximization of expected profit, given the participation constraint (6.2) and adding the *incentive-compatible constraints*:

$$f[h(\varepsilon), \varepsilon] - W(\varepsilon) \geq f[h(\theta), \varepsilon] - W(\theta), \quad \forall(\varepsilon, \theta) \tag{6.11}$$

The direct solution of this optimization problem is generally complex (see Rosen, 1985, and Salanié, 1997). But an astute observation made by Cooper (1983) gives us a very simple way to find out which incentive-compatible constraints will be binding.

A Method for Finding the Second-Best Contract

Let $\mathcal{A}_1 = \{W_1(\varepsilon), h_1(\varepsilon)\}$ be the first-best contract with symmetric information defined by the first-order conditions (6.3) and (6.4) and let us imagine that this contract is in force in a situation of asymmetric information. When state ε appears, the principal announces that it is state $m_1(\varepsilon)$, the solution of problem (6.8), that has occurred. It is not difficult to specify the properties of state $m_1(\varepsilon)$ according to the form of the utility function $U(C, L)$ of the agent. For that purpose, let $\Pi(\varepsilon, \theta)$ be the profit $f[h_1(\theta), \varepsilon] - W_1(\theta)$ that comes to the principal when, with the first-best contract $\mathcal{A}_1$ in force, she announces that she has observed state θ whereas in reality it is state ε that has come about. Taking into account comparative statics relations (6.5) and the risk-sharing condition (6.3) satisfied by the first-best contract, we find that the partial derivative Π_θ of profit $\Pi(\varepsilon, \theta)$ with respect to θ, satisfies the equalities:

$$\Pi_\theta = h_1'(\theta) f_h[h_1(\theta), \varepsilon] - W_1'(\theta) = \left(\frac{U_L U_{CC} - U_{CL} U_C}{U_{CC}} \right)_{(h_1(\theta), W_1(\theta))} h_1'(\theta) \tag{6.12}$$

In the first place, this equation shows that $m_1(\varepsilon) = \varepsilon$ for all utility functions such that $U_L U_{CC} - U_{CL} U_C = 0$ at every point. In that case, leisure demand is then independent of income (see chapter 1, appendix 2). This property is satisfied, for example, if the agent's utility function takes the form $U[W + \phi(1 - h)]$ with $U' > 0$, $U'' \leq 0$, $\phi' > 0$, and $\phi'' \leq 0$ (Azariadis, 1983). Under these hypotheses, the optimal contract with asymmetric information—also called the second-best contract—is no different from the first-rank contract. Moreover, it was established in chapter 1, appendix 7.2 that leisure is a normal good if and only if $U_{CL} U_C - U_L U_{CC} > 0$. If we accept this standard hypothesis, relation (6.12) shows that profit $\Pi(\varepsilon, \theta)$ is increasing with θ. The signal $m_1(\varepsilon)$ is thus equal

to the upper bound ε^+ of the set of possible values of the random variable ε. Conversely, if $U_{CL}U_C - U_L U_{CC} < 0$, leisure is an inferior good, profit $\Pi(\varepsilon, \theta)$ becomes a decreasing function of θ, and state $m_1(\varepsilon)$ coincides with the lower bound ε^- of the possible values of ε. In conclusion, the results of this analysis are summed up in the following manner:

$$m_1(\varepsilon) = \begin{cases} \varepsilon & \text{if there is no income effect} \\ \varepsilon^+ & \text{if leisure is a normal good} \\ \varepsilon^- & \text{if leisure is an inferior good} \end{cases} \tag{6.13}$$

Thus, asymmetric information is not a source of any inefficiency when the demand for leisure is independent of income. Conversely, when the demand for leisure depends on income, the firm will most often have an interest in sending out misleading messages if the first-best contract applies. This contract is thus not incentive compatible and, in the definition of the optimal second-best contract, incentive-compatible constraints (6.11) corresponding to states in which the principal would have lied if the first-rank contract had been in force will be binding.

2.2.2 An Example with Two States of Nature

We illustrate the revelation principle in a simple model with only two states of nature. Assuming that leisure is a normal good, the conclusions agree better with the stylized facts than the conclusions that issue from an equivalent model with symmetric information.

The Principal's Problem

With the help of response $m_1(\varepsilon)$ described by (6.13), it is possible to find out the properties of the second-best contract in a model with only two states of nature ε^+ and ε^-, equiprobable and such that $\varepsilon^+ > \varepsilon^-$. To make the exposition simpler, we will assume as well that the principal is risk neutral and that the production function takes the multiplicative form $f(h, \varepsilon) = \varepsilon h$. The optimal contract maximizes the principal's expected profit subject to the participation and incentive-compatible constraints; hence assuming that each state has the same probability of occurring, it is the solution of the problem:

$$\max_{(h^i, W^i)_{i=+,-}} \left[\frac{1}{2}\left(\varepsilon^+ h^+ - W^+\right) + \frac{1}{2}\left(\varepsilon^- h^- - W^-\right) \right]$$

subject to constraints:

$$\frac{1}{2}U(W^+, 1 - h^+) + \frac{1}{2}U(W^-, 1 - h^-) \geq \overline{U} \tag{6.14}$$

$$\varepsilon^+ h^+ - W^+ \geq \varepsilon^+ h^- - W^- \tag{6.15}$$

$$\varepsilon^- h^- - W^- \geq \varepsilon^- h^+ - W^+ \tag{6.16}$$

Let $\lambda \geq 0$, $\mu_1 \geq 0$, and $\mu_2 \geq 0$ be the Kuhn and Tucker multipliers respectively associated to the participation constraint (6.14) and the incentive-compatible

constraints (6.15) and (6.16); the first-order conditions of the principal's problem are found by setting the derivatives of the Lagrangian—which the reader may write out in full if he or she wishes—to zero with respect to variables W^i and h^i for $i = +, -$. The result is:

$$\frac{\partial \mathcal{L}}{\partial W^+} = -\frac{1}{2} + \frac{\lambda}{2} U_C(W^+, 1 - h^+) - \mu_1 + \mu_2 = 0 \tag{6.17}$$

$$\frac{\partial \mathcal{L}}{\partial W^-} = -\frac{1}{2} + \frac{\lambda}{2} U_C(W^-, 1 - h^-) + \mu_1 - \mu_2 = 0 \tag{6.18}$$

$$\frac{\partial \mathcal{L}}{\partial h^+} = \left(\frac{1}{2} + \mu_1\right) \varepsilon^+ - \frac{\lambda}{2} U_L(W^+, 1 - h^+) - \mu_2 \varepsilon^- = 0 \tag{6.19}$$

$$\frac{\partial \mathcal{L}}{\partial h^-} = \left(\frac{1}{2} + \mu_2\right) \varepsilon^- - \frac{\lambda}{2} U_L(W^-, 1 - h^-) - \mu_1 \varepsilon^+ = 0 \tag{6.20}$$

The Optimal Contract When Leisure Is a Normal Good

If we add up the equalities (6.17) and (6.18), we can easily verify that $\lambda > 0$; the participation constraint (6.14) is thus binding. Taking leisure to be a normal good, as usual, we know from rule (6.13) that the principal has an interest in overestimating the true state of nature when the first-best contract is in force. In other words, the principal would lie if the "bad" state of nature ε^- came about. It results that constraint (6.15) is not binding; hence $\mu_1 = 0$, and that constraint (6.16) is saturated; hence $\mu_2 \geq 0$. In addition, condition (6.19) entails:

$$\frac{\lambda}{2} U_L(W^+, 1 - h^+) = \frac{\varepsilon^+}{2} - \mu_2 \varepsilon^- > \left(\frac{1}{2} - \mu_2\right) \varepsilon^+ \tag{6.21}$$

Noting once again that relation (6.17) gives $\lambda U_C(W^+, 1 - h^+) = 1 - 2\mu_2$, we arrive at the following inequality:

$$\frac{U_L(W^+, 1 - h^+)}{U_C(W^+, 1 - h^+)} > \varepsilon^+ \tag{6.22}$$

Since $\mu_1 = 0$, conditions (6.20) and (6.18) respectively entail $\lambda U_L(W^-, 1 - h^-) = (1 + 2\mu_2)\varepsilon^-$ and $\lambda U_C(W^-, 1 - h^-) = 1 + 2\mu_2$. Eliminating the positive quantity $1 + 2\mu_2$ between (6.21) and (6.22), we get:

$$\frac{U_L(W^-, 1 - h^-)}{U_C(W^-, 1 - h^-)} = \varepsilon^- \tag{6.23}$$

In the first place, we can show, using the incentive-compatible constraints, that wages and hours vary in the same direction. Thus, the binding constraint (6.16) entails $W^+ - W^- = \varepsilon^-(h^+ - h^-)$, whereas constraint (6.15), which is not binding, entails $\varepsilon^+(h^+ - h^-) \geq W^+ - W^-$. In consequence, we have $(\varepsilon^+ - \varepsilon^-)(h^+ - h^-) \geq 0$. Since $\varepsilon^+ > \varepsilon^-$, we deduce $h^+ > h^-$ and so $W^+ > W^-$. Differently to the case in which information was symmetric, the wage now rises unambiguously when the economic trend turns up. This property is a direct consequence of the hypothesis of asymmetric information.

The values of ε being unverifiable, the agent knows that the principal has no interest in declaring that the good state of nature ε^+ has appeared if such a declaration leads to a higher wage. Hart (1983) has shown in a much more general model that the principal has an incentive to reveal the true state of nature if a heavy work schedule is linked to a high wage.

Conditions (6.22) and (6.23) also indicate that the inefficiency due to asymmetric information only manifests itself in the good state of nature. In this state, the marginal rate of substitution between consumption and leisure surpasses marginal productivity, whereas these quantities must be equal—see (6.3)—in the first-best contract. With a few calculations, we can show that this inefficiency leads to higher remuneration and longer hours of work in the second-best contract when the good state of nature occurs if leisure is a normal good, which is the empirically relevant assumption (see chapter 1). This result suggests that asymmetric information helps to increase employment. The absence of verifiability of workers' performance does not therefore help to explain underemployment. Malcomson (1999) notes that this result might however explain the fact that in many contracts the firm has the right to demand that its employees supply a certain volume of overtime hours. Overall, the model with asymmetric information, although a disappointment when it comes to explaining underemployment, does come to conclusions that fit better with the stylized facts than does the model in which information is assumed to be symmetric, that is, verifiable.

Overall, taking risk-sharing by employers and employees into account fits well with certain empirical characteristics of wages and hours, like the low variability of wages and the procyclicity of hours and compensation. We will now proceed to show that the labor contract also helps us to solve incentive problems when the employee's effort is not verifiable. This dimension of the wage relationship allows us to gain an understanding of a number of empirical elements concerning wage formation.

3 INCENTIVE IN THE PRESENCE OF VERIFIABLE RESULTS

To this point we have assumed that hours worked were perfectly verifiable and could therefore be written into the labor contract explicitly. But hours worked must not be confused with the "effort" made by an employee in carrying out his tasks. In practice, it is possible in most circumstances to check very easily that an employee is present at his place of work at the set times, but it is much harder to assess the intensity of his effort, although the latter determines the speed, precision, and quality with which tasks are carried out. For this reason, much thought has been devoted to the study of labor relations when workers' effort is not verifiable. The employer is faced with an incentive problem: that of drafting a contract that will impel the worker to furnish the maximum of effort at the least cost.

In this section, we focus on situations in which effort is not verifiable, but the results of an agent's activity are. The case in which neither the effort nor the results are verifiable will be analyzed in the following section. We begin by showing, using the agency model with hidden action, how the absence of verifiability of effort keeps the employer from correctly insuring her employees against fluctuations in activity. The hunt for incentive mechanisms does, indeed, lead employers, in certain circumstances

on which the theory of contracts sheds light, to offer remunerations tied to collective or individual results when it would have been in her interest to offer constant remuneration, independent of results, if effort were verifiable. We will then see that the relationship between result and remuneration can take different forms, ranging from incentive pay to promotion based on hierarchical rules, the logic of which is closely similar to that governing sports tournaments.

3.1 The Principal–Agent Model with Hidden Action

The situation analyzed by the agency model with hidden action is schematically comparable to that of a gold prospector or a salesperson. When the owner of a gold property, or the manager of a firm, employs persons of this type, she anticipates remunerating them on the basis of their results, that is, on the basis of how much gold is found or the volume of sales. In this context, a mediocre result does not necessarily reflect a feeble *effort* on the part of the gold prospector or the salesperson. The fact is that in many circumstances the result in question also reflects general conditions, independent of the will of the actors, in which their activity takes place. The quality of the gold property worked over by the prospector or the demand for the product sold by the salesperson falls into this category. There is generally an element of risk in the individual's activity, against which he wishes to be insured. But a complete insurance, providing remuneration independent of the result, and thus of the effort made, is highly likely to provide little incentive. The agency model shows how the rules of remuneration give rise to a trade-off between the need for insurance and incentive. Finally, in cases where the employer receives information from sources other than direct observation of individual performance (the performance of a team, for example, or reports made by supervisors), the question of what indicators to use in regulating remuneration arises, as does that of the efficiency of rules based solely on verifiable data.

In the agency model with hidden action, the principal—or the employer—is confronted with a problem of moral hazard, inasmuch as she does not know, a priori and with certainty, what actions the agent—or employee—has undertaken to achieve the observed results. In this context, the basic agency model shows that the remuneration rule chosen by the principal depends on the results of the agent's activity and will arrive at a compromise between the motives of insurance and incentive.

The canonical agency model with hidden action focuses on the behavior of a principal and an agent whose decisions unfold in the following sequence: (1) the principal offers a contract; (2) the agent accepts or refuses the contract; (3) if the agent turns it down, the protagonists go their separate ways, but if the agent accepts it, he then supplies an effort; (4) a random event that affects the result of the agent's effort occurs; (5) the principal and the agent observe the result; (6) the principal remunerates the agent according to the terms of the contract. The optimal decisions can be found through backward induction, so we must first define the behavior of the agent who has accepted a contract, then determine the choice of the principal, who anticipates the agent's decisions.

3.1.1 The Canonical Agency Model

We begin by describing the behaviors of agents and the principal, and we will show that the main property of the optimal remuneration rule is a trade-off between risk and insurance.

The Agent's Behavior

In order to study the decisions of the two parties, we will consider a very simple static model. Thus, we assume that the utility function describing the agent's preferences takes the exponential form:

$$U[W - C(e)] = -\exp\{-a[W - C(e)]\} \tag{6.24}$$

In this expression, the variables W and e designate respectively the remuneration received by the agent and the effort he has expended in the production process. The function $C(e)$ represents the cost linked to the supplying of effort e. To simplify the calculations, we will adopt the quadratic representation $C(e) = ce^2/2$, $c > 0$, but all the results of this section remain true on the assumption that the cost function is strictly convex. Finally, readers are reminded that the parameter $a > 0$ is the index of absolute risk aversion, equal to $-U''/U'$ (see for example Mas-Colell et al., 1995, chapter 6). The utility function chosen, which is of the CARA (Constant Absolute Risk Aversion) type, thus entails a constant index of absolute risk aversion. The choice of a hypothesis of this kind makes the exposition of the agency model a great deal simpler, while the conclusions reached extend, in essence, to more general environments (see Salanié, 1997; Macho-Stadler and Perez-Castrillo, 2001; and Bolton and Dewatripont, 2005). When necessary, we will make clear which results flow specifically from this hypothesis.

When the agent supplies effort e, he allows the principal to reap the benefit of production $y = e + \varepsilon$, where ε is a normal random variable with zero mean and standard error σ. This random variable represents factors that are beyond the agent's control—things like health, the weather, and luck. The presence of a random variable prevents an impartial third party from knowing exactly the effort e of the agent by observing his production y. The effort thus cannot be verified, but the production can. The principal is then in a position to construct a remuneration rule based on observation of the production achieved. To simplify, we assume that the principal adopts the linear rule $W = w + by$, where w represents a fixed wage independent of the performance of the agent and b is a piece rate on production y (it can be shown that the optimal remuneration rule is indeed linear, with the hypotheses of constant index of absolute risk aversion and a normal random variable; see Holmström and Milgrom, 1987). If we assume that the agent has to make his decisions before knowing the realization of the random variable ε but with knowledge of the remuneration rule proposed by the principal, he chooses a level of effort that maximizes his expected utility. Since $W = w + b(e + \varepsilon)$, the definition (6.24) of the agent's preferences shows that this expected utility is then equal to $-\exp\{-a[w + be - C(e)]\}\mathbb{E}[\exp(-ab\varepsilon)]$. And since the random variable ε follows a normal distribution with zero mean and standard variation σ, the random variable $\exp(-ab\varepsilon)$ follows a log-normal distribution, the mean of which[5] is equal to

[5] At this point the reader may wish to refer to mathematical appendix C, section 3.3, at the end of the book, which establishes the main properties of normal and log-normal distributions. Here we simply note that the probability density of a random variable X following the normal distribution $\mathcal{N}(m,\sigma)$ is given by $f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x-m)^2}{2\sigma^2}\right]$. The random variable $\exp(X)$ then follows a log-normal distribution with mean $\exp[m + (\sigma^2/2)]$.

$\exp(a^2b^2\sigma^2/2)$. In sum, the utility expected by the agent is written:

$$\mathbb{E}U = -\exp\left\{-a\left[w + be - C(e) - \frac{ab^2\sigma^2}{2}\right]\right\} \tag{6.25}$$

The maximization of expected utility implies that the level of effort e^* chosen by the agent is such that $C'(e^*) = b$, or $e^* = b/c$. This equality portrays the *incentive* properties of the remuneration rule. The agent's effort evidently does not depend on the fixed part w of this rule but increases as the relationship between remuneration and performance, measured by the parameter b, rises in intensity.

The Principal's Behavior

To set the value of b, the principal reckons on level e^* of effort by the agent, since the contract is signed before the agent starts work. The relationship between the remuneration rule and the level of effort, $C'(e^*) = b$, is imposed on the principal and is called the *incentive-compatible constraint*. The principal must also take into account the *participation constraint*, which indicates the conditions under which the agent accepts the contract. We will assume that the agent can always attain an expected utility $\overline{U}$ outside the contractual relationship. Hence the participation constraint is written $\mathbb{E}U \geq \overline{U}$ with $\overline{U} < 0$. The principal, assumed to be risk neutral, chooses w and b in such a way as to maximize her expected profit, given this participation constraint and knowing that the agent's effort is equal to e^*. In these circumstances, the agent's production is given by $y = e^* + \varepsilon$ and his remuneration amounts to $W = w + b(e^* + \varepsilon)$. Since the random variable ε has zero mean, the profit expected by the principal, $\mathbb{E}(y - W)$, is equal to $(1 - b)e^* - w$. In the end, the principal's problem comes down to the following optimization problem:

$$\max_{\{w,b\}} [(1 - b)e^* - w] \quad \text{subject to} \quad C'(e^*) = b \quad \text{and} \quad \mathbb{E}U \geq \overline{U} \tag{6.26}$$

Let us set $\bar{x} = -\ln(-\overline{U})/a$; taking the logarithms of the opposites of the two sides of the participation constraint $\mathbb{E}U \geq \overline{U}$, we find that this constraint takes the form:

$$w + be^* - C(e^*) - \frac{ab^2\sigma^2}{2} \geq \bar{x} \tag{6.27}$$

The problem (6.26) can be simply solved by noting that the effort e^* defined by the incentive constraint is independent of the fixed part of the remuneration. Let us suppose that the principal has settled on the value of parameter b; it is clearly in her interest to select w in such a way as to bind the agents's participation constraint, since w does not affect e^*. Carrying the value of w thus obtained into the principal's problem, we observe that the optimal value of parameter b is the solution of the following problem:

$$\max_b \left[e^* - C(e^*) - \frac{ab^2\sigma^2}{2} - \bar{x}\right] \quad \text{subject to constraint } C'(e^*) = b$$

The Trade-off Between Risk and Incentive

The quadratic form $ce^2/2$ of the cost function allows us to define explicitly the optimal value of b, as follows:

$$b^* = \frac{1}{1 + ac\sigma^2} \tag{6.28}$$

This simple formula perfectly illustrates the trade-off between the motives of incentive and insurance. At the optimum, positive effort $e^* = b^*/c$ results from a positive value of b^*, since in this case the remuneration varies with the level of production. The negative linkage between b^* and σ reflects the trade-off between risk and incentive. The higher the risk (σ is large), the higher the insurance motive and the lower the incentive motive (b^* is small). At the limit, parameter b^* goes to zero when the variance of ε is infinite. In this case, production is no longer linked to effort, and the incentive motive vanishes. We also see that b^* diminishes with the degree of absolute risk aversion a. In other words, the more risk averse an agent is, the less marked the relationship between the result and the remuneration becomes. On the other hand, it is easy to verify that the more the fixed part of the remuneration grows in importance, the stronger risk aversion is. The optimal value w^* of the fixed part of the remuneration is found by bringing the value of b^* defined by (6.28) into the participation constraint (6.27) written in the form of an equality, or:

$$w^* = \bar{x} - \frac{1 - ac\sigma^2}{2c\,(1 + ac\sigma^2)^2} \tag{6.29}$$

We can also observe that parameter b^* decreases with measure c of the disutility of effort. Thus, an agent for whom the disagreeability of effort has less weight than it does for someone else will be more attracted to a compensation rule that privileges payment by results. When agents are heterogeneous according to characteristic c and when employers do not observe this characteristic, then employers may increase the relative importance of the variable part of the remuneration as compared to the fixed part in order to attract agents who are more tolerant of effort or, to put it another way, ones who are more efficient (see Lazear, 1986 and 2000, for models built around this mode of selection).

First-Best Optimum and Second-Best Optimum

It is important to point out that the nonverifiable character of effort and the variability of remuneration mean that the contract arrived at produces an allocation that is a *second-best optimum*. This means that it would have been possible to find a contract that improved the outcome for at least one of the partners, with no detriment to that of the other, if effort were verifiable. The fact is, given a contract prescribing variable remuneration, that any other contract which allotted the average of the remuneration prescribed by the earlier contract to the employee under all states of nature would provide the employer with the same expected profit. On the other hand, it would clearly improve the situation of the agent, since he is not risk neutral. So the absence of complete insurance proves to be inefficient. When effort is not verifiable, the only possible incentive mechanisms necessarily link remuneration to production (the only verifiable variable) and so there cannot be total insurance.

So as better to grasp the consequences of this situation of moral hazard, let us suppose that effort e is verifiable and that the contract stipulates a remuneration that always takes the form $W = w + by$. When effort is verifiable, it is as though the principal had the ability to decide how much effort the agent was making. The principal's problem then consists of maximizing her expected profit with respect to (b, w, e) subject to the worker's participation constraint (6.27) only. The expected production and remuneration being respectively equal to e and $w + be$, the problem defining the so-called first-best contract is written:

$$\max_{\{w,b,e\}} [e - (w + be)] \quad \text{subject to } w + be - C(e) - \frac{ab^2\sigma^2}{2} \geq \bar{x}$$

We see that the participation constraint is binding and that the optimal values, denoted (b^o, w^o, e^o), are defined by:

$$C'(e^o) = 1, \quad b^o = 0, \quad w^o = \bar{x} + C(e^o)$$

The first-best allocation corresponds to a pure insurance contract, in which the employer insures the worker totally against the hazards of production by giving him remuneration $w^o = \bar{x} + C(e^o)$, independent of production. We may also note that effort in the first-best contract, defined by $C'(e^o) = 1$, is greater than effort in the second-best contract defined by the equation (6.28) where $C'(e^*) = b^* < 1$, given that the employee is averse to risk. So the first-best contract entails a higher level of production.

It is worth noting that the level of effort e^o in the first-best contract is attained even if effort is unverifiable when the agent is risk neutral ($a = 0$). In this case, equation (6.28) shows that the agent has no need to be insured, and the principal has an interest in offering a remuneration such that all the production goes to the agent ($b^* = 1$). In this context, the first- and second-best allocations coincide. Thus the incentive motive exists independently of risk-sharing, and a value of piece rate b strictly lower than unity, entailing a fall in production, is, in a sense, the price to pay for solving the problem of moral hazard that besets a principal facing a risk-averse agent.

3.1.2 EMPIRICAL ILLUSTRATIONS

The agency model we have just illustrated predicts that the efforts of workers depend positively on the financial incentives offered to them and that the optimal remuneration rule yields a trade-off between risk and insurance. The first prediction receives stronger confirmation from empirical research than the second.

On the Trade-off Between Risk and Incentive

Prendergast (2002) draws up a comprehensive assessment of empirical research focused on the trade-off between risk and incentive. He groups the results according to occupational categories labeled executives, sharecroppers, and franchisees. For the executives, the evidence is inconclusive. For sharecropping, research tends to show that the fraction of output sharecroppers keep is increasing with the noisiness of the financial returns, which inverts the result predicted by the agency model. Similarly, the research tends to vindicate a positive relation between franchising (in which remuneration is closely tied

to performance) and the risk incurred, which also contradicts the notion of a trade-off between risk and incentive. Overall, Prendergast estimates that what we know empirically suggests a relation between risk and incentive that is the inverse of what the canonical agency model predicts. This does not mean that the trade-off between risk and incentive does not exist; rather it may be the case that other factors are at work and that these factors dominate this trade-off.

Prendergast thus suggests that the riskier the situation is, the greater the marginal return to effort, and that that might give rise to an apparently increasing relation between the extent of risk and the incentive motive. To grasp this result, let us assume that production is tied to effort by relation $y = \gamma e + \varepsilon$ where γ represents the marginal return to effort. If we revert to the calculations of the canonical agency model, it is easy to verify that the optimal values of effort and of parameter b are given by:

$$e^* = \gamma \frac{b^*}{c}; \quad b^* = \frac{1}{1 + ac\left(\frac{\sigma}{\gamma}\right)^2}$$

If we assume that the marginal return to effort increases with risk, that γ increases with σ, it is perfectly possible that b^* becomes an increasing function of σ, so that the greater the risk, the greater would be the incentive motive. Prendergast (2002) shows that this situation may come about in a model where firms have an interest in delegating more decision-making power to their employees when the latter are better informed than their managers. More delegation then entails a greater marginal return to effort in the presence of increased risk, which induces an increasing relation between the incentive motive and the risk.

On the Power of Incentives: The Example of Autoglass Installers

Lazear (2000) studied the evolution of compensation schemes within Safety Glass Corporation, a large autoglass installer in the United States; his observations clearly illustrate the main lessons of the basic agency model. Until January 1994, glass installers were paid an hourly wage rate, which did not vary in any direct way with the number of windshields or windows that were installed. During 1994 and 1995, following a change in management, this firm moved gradually from a system of fixed hourly wages to a piece-rate system. Rather than being paid for the number of hours that they worked, installers were paid for the number of glass units that they installed. But a guaranteed minimum wage was also part of the new system. Hence workers who looked with diffidence upon the introduction of this new system (because they were less motivated or less productive than others) did not automatically have an interest in quitting the firm, since they could continue to earn the minimum wage.

Lazear had access to very precise data that specified the monthly production of each worker, and he followed more than 3,000 workers over a 19-month period. Since the piece-rate system was phased in over 19 months, many workers were employed under both regimes. Thus, data on individual output are available for most installers both during the hourly-wage period and during the piece-rate pay period. Lazear has estimated that the switch from hourly to piece-rate pay led to a 44% increase in output per worker. It would however be unsafe to jump to the conclusion that this increase was due solely to this new system of remuneration. It is perfectly possible that the new

system exerted a *selection effect* on the workforce at Safety Glass. Since Lazear's data made it possible to control for individual effects linked to each worker's particular ability, he could estimate this selection effect. He finds that approximately half of the 44% difference in productivity reflects an incentive effect. The other half therefore reflects a selection effect. Specifically, by controlling for seniority Lazear was able to establish that persons hired after 1 January 1995, meaning they had experienced only the piece-rate system, had log productivity 0.24 greater than those hired under the old regime. Hence the shift from a fixed-wage system to a piece-rate system did exert a selection effect. Sorting occurred primarily through the hiring process and, to a lesser extent, from a reduction in quits among the highest output workers or from an increase in quits among the least productive workers (who were protected by the guaranteed minimum wage). In general terms, the goal of inducing self-selection is one of the leading explanations for the adoption of performance-based pay in organizations (see section 3.2.2 and Oyer and Schaefer, 2011).

The agency model also predicts that changing the remuneration rule ought to lead to wide variation in individual performance. Lazear does indeed observe that the variance of individual production reached the level of 2.53 under the new system of performance pay, whereas it had been only 2.02 under the fixed-wage system.

On the Power of Incentives: The Example of Tree Planters in British Columbia

Another frequently cited example is the compensation schemes of tree planters in British Columbia, studied by Paarsch and Shearer (1999) and Shearer (2004) through a field experiment in which nine workers were randomly selected and then randomly assigned to be paid using piece rates or a fixed wage (the sample size was small because it was hard to convince the employers to accept an experimental test of their remuneration policies). Each worker was observed for 60 days. It was found that piece rates led to a 20% increase in individual-level productivity—a result in striking proximity to what Lazear found for the workers at Safety Glass. Shearer (2004) also found that the standard deviation of output across workers was wider under piece rates and that, in this system of remuneration, the unit cost was 13% lower than under fixed wages.

These studies therefore suggest that financial incentives do influence the behavior and the performance of workers in the way predicted by the theory of incentive (other examples that tend to confirm this suggestion are reviewed in Lazear and Oyer, 2010). The very thorough survey of Bloom and Van Reenen (2011) arrives at the conclusion that the available empirical evidence shows that, as a general rule, pay-for-performance incentives are associated with improvements in organizational performance. It should nevertheless be pointed out that financial incentives can have counterproductive effects to which we will return in section 5 of this chapter on social preferences.

3.2 Should Remuneration Always Be Individualized?

To this point we have assumed that the principal could only make the remuneration of an agent depend on that person's individual production. But even in cases as simple as that of the gold prospector or the salesperson, there is no reason why the sharing rule need depend exclusively on individual production, if there are other verifiable variables, the utilization of which would make it possible to work out more efficient contracts.

3.2.1 THE AGENCY MODEL WITH TWO SIGNALS

In a very general way, we may suppose that the principal observes not just individual production y but a signal $\tilde{\varepsilon}$ independent of the agent's level of effort yet capable of being correlated with the random variable ε. Like production, the signal $\tilde{\varepsilon}$ is not only observable but also verifiable. For example, weather conditions do not depend on how hard a farm worker exerts himself, but they do very often affect the harvest and are verifiable. More generally, signal $\tilde{\varepsilon}$ may concern macroeconomic variables or the observation of the production of other agents or of the "team" to which the agent in question belongs (see Holmström, 1982, for a meticulous analysis of the problem of moral hazard in teams). It is in the principal's interest to make use of this signal when the efforts of agents combine in more or less complex ways in the production process.

In this setup, the agent's compensation rule may depend on the observation of his individual production y and that of signal $\tilde{\varepsilon}$. A (linear) compensation rule thus takes the form $W = w + by - \tilde{b}\tilde{\varepsilon}$. The definition (6.24) of the agent's preferences shows that his expected utility is now equal to:

$$-\exp\{-a[w + be - C(e)]\}\,\mathbb{E}\left\{\exp[-a(b\varepsilon - \tilde{b}\tilde{\varepsilon})]\right\}$$

Let us assume, in order to simplify, that the random variable $\tilde{\varepsilon}$ is normally distributed with zero mean and standard error σ, and let ρ be the correlation coefficient between the variables ε and $\tilde{\varepsilon}$. We thus have $cov(\varepsilon, \tilde{\varepsilon}) = \rho\sigma^2$. In these conditions the random variable $-a(b\varepsilon - \tilde{b}\tilde{\varepsilon})$ follows a normally distributed law with zero mean and variance $a^2\sigma^2(b^2 + \tilde{b}^2 - 2\rho b\tilde{b})$, and the random variable[6] $\exp[-a(b\varepsilon - \tilde{b}\tilde{\varepsilon})]$ has a log-normal distribution with mean $a^2\sigma^2(b^2 + \tilde{b}^2 - 2\rho b\tilde{b})/2$. The expected utility of the agent is now written:

$$\mathbb{E}U = -\exp\left\{-a\left[w + be - C(e) - \frac{a\sigma^2}{2}\left(b^2 + \tilde{b}^2 - 2\rho b\tilde{b}\right)\right]\right\}$$

We observe that optimal effort is always characterized by the equality $C'(e^*) = b$. The mean of the random variable $\tilde{\varepsilon}$ being zero, the principal's expected profit is again equal to $(1 - b)e^* - w$ and in consequence, the optimal compensation rule is again the solution of the problem (6.26). Taking the logarithms of the opposites of both sides of the participation constraint $\mathbb{E}U \geq \overline{U}$, we find that the latter now takes the following form:

$$w + be - C(e) - \frac{a\sigma^2}{2}\left(b^2 + \tilde{b}^2 - 2\rho b\tilde{b}\right) \geq \bar{x} \qquad \textbf{(6.30)}$$

3.2.2 THE OPTIMAL COMPENSATION RULE

As before, the principal has an interest in choosing w in such a way as to bind the participation constraint. If we bring the value of w thus obtained into the principal's problem, we see that the optimal values of parameters b and $\tilde{b}$ solve:

$$\max_{\{b,\tilde{b}\}}\left[e^* - C(e^*) - \frac{a\sigma^2}{2}\left(b^2 + \tilde{b}^2 - 2\rho b\tilde{b}\right) - \bar{x}\right] \quad \text{subject to } C'(e^*) = b$$

[6] Mathematical appendix C, section 3.3, points out that if $X \rightsquigarrow \mathcal{N}(0, \sigma_X)$, then $\exp(X)$ has a log-normal distribution with mean $\exp(\sigma_X^2/2)$.

As $C(e) = ce^2/2$, we find, after simple calculations, that the optimal values b^* and $\tilde{b}^*$ are defined by:

$$b^* = \frac{1}{1 + ac\sigma^2(1-\rho^2)} \quad \text{and} \quad \tilde{b}^* = \rho b^* \tag{6.31}$$

If variables ε and $\tilde{\varepsilon}$ are independent, the correlation coefficient ρ is equal to zero and the indexation coefficient $\tilde{b}^*$ is null. The observation of $\tilde{\varepsilon}$ then has no informative value. Conversely, if variables ε and $\tilde{\varepsilon}$ are not independent, the optimal remuneration rule takes into account all the information available. The optimal value of the fixed part of the remuneration is obtained by using the participation constraint (6.30) written in the form of an equality and definitions (6.31) of b^* and $\tilde{b}^*$, or:

$$w^* = \bar{x} - \frac{1 - ac\sigma^2(1-\rho^2)}{2c\left[1 + ac\sigma^2(1-\rho^2)\right]^2} \tag{6.32}$$

We see that total remuneration, $W^* = w^* + b^* y - \tilde{b}^* \tilde{\varepsilon}$, *falls* when $\tilde{\varepsilon}$ increases for a given value of y. This result flows from the fact that the principal knows that a high value of production is less the consequence of a special effort on the part of the agent than it is of an exogenous rise in the random variable $\tilde{\varepsilon}$. An interesting case is that in which $\tilde{\varepsilon}$ becomes an indicator of the activity of others employed in the firm, or even of the activity in other firms in an analogous environment. If the principal cannot "filter out" the contribution of other workers, or the general market trend, to the agent's production, then it is not optimal to make the remuneration of an individual depend solely on production. This justifies schemes in which a part of the remuneration depends on an indicator relative to the performance of others in the same firm or the economic trend in a particular sector. For example, profit-sharing rules adopted in certain firms frequently appear to reflect thinking of this kind (Cahuc and Dormont, 1997). Along the same lines, Gibbons and Murphy (1990) have observed that there might be grounds for penalizing the managers of a firm when that firm's share price does not rise as fast as the average index of the stock market.

3.3 Some Reasons That Performance Pay May Be Inefficient

Two major sources of inefficiency in compensation schemes based on verifiable observations are the multiplicity of the tasks that go to make up the content of the work done by any individual and the fact that an agent's activities are generally observed by his supervisors, whose objectives do not necessarily coincide completely with those of the principal. That being the case, an employee may have an interest in focusing part of his effort on activities likely to catch the supervisor's approving eye.

3.3.1 Multitasking

In what has gone before, an employee's remuneration was based on the putatively verifiable observation of a *single* scalar deemed to represent the agent's production. This approach eliminates much of the difficulty arising from *multitasking*—the fact that the productive activities of most individuals have many dimensions. Given the reality of multitasking, the principal may be tempted to base an agent's remuneration on the only

verifiable observations available. But in doing so, the principal actually gives the agent an incentive to put more effort into precisely the kind of actions that do give rise to verifiable observations but that may not necessarily bring the principal the most benefit. The history of the former USSR abounds in anecdotes about projects selected according to the number of nails used or their weight (both verifiable quantities). Multitasking is one of the reasons firms usually adopt implicit contracts. Brown (1990), for example, shows that the frequency of implicit contracts rises, and that of piece rate diminishes, with an indicator of how many different tasks agents are assigned.

The remuneration of school teachers well illustrates the question of multitasking. If their pay depends too much on the measurable test results of their pupils, school teachers may have an incentive to focus their efforts on improving test results, to the detriment of tasks harder to measure, like imparting confidence to pupils, teaching them to work together in groups, and enhancing their noncognitive capacities. On this account, some school administrators prefer not to use pupils' test scores to determine teachers' pay (Oyer and Schaefer, 2011).

When possible, firms sometimes choose to index the remuneration of agents to *global* indicators strongly correlated with the objectives of the principal. Payment for managers in the form of stock options is a practice that is spreading precisely because it makes it possible to align the interests of managers (i.e., agents) with those of shareholders (here regarded as the principal). Likewise, all the players on a soccer team generally receive the same bonus when their team wins: if the center forward were paid for the number of goals he scored, he would probably have a tendency to try to score goals too often, instead of passing the ball to teammates in a better position to do so (on questions of multitasking see Holmström and Milgrom, 1991; the summary of Prendergast, 1999; and Oyer and Schaefer, 2011).

3.3.2 Supervision and Rent-Seeking

In firms above a certain size, it is not the principal who observes the performance of agents. This activity is delegated to supervisors whose precise role is to report what they observe to the principal. But supervisors are themselves agents, and their objectives do not necessarily coincide with those of the principal. For example, it is sometimes observed that in order to avoid friction with the people with whom they have to work every day, supervisors tend to write reports in which bad performances are made to look better than they are, thus minimizing the degree of difference among the employees they supervise (see for example Murphy and Cleveland, 1991).

Another, and surely more important, problem is known in the literature as *rent-seeking*. It is caused by the fact that agents may derive a comparative advantage from devoting a part of their efforts to actions that will impress supervisors so that the latter will write favorable reports about them, instead of devoting all their efforts to tasks that are the most beneficial to the principal (on this, see Milgrom, 1988, and Tirole, 1992). In France, for example, teachers are hired through competitions, and in some of these the members of the judging panel are well-known personalities. There is a tendency for the candidates to espouse the opinions of these personalities, or at any rate to demonstrate that they are acquainted with them (which bears a corresponding cost in time), so as to make a favorable impression. Prendergast (1993b, 1999) points out that rent-seeking most often makes its appearance in situations in which it would be extremely hard to find any objective yardstick by which to measure production, and our example of hiring

competitions for teachers falls into this category to some extent. Prendergast (1993b) shows that these situations breed yes-men, whose purpose is simply to avoid standing out from the crowd.

A Model with Rent-Seeking

The inefficiency generated by rent-seeking comes to light naturally if we marginally change the basic agency model. Let us now suppose that the agent is able to exercise two types of effort. He can put effort e into activities that are directly productive, in which case his production y is again given by $y = e + \varepsilon$. But the agent can also put out effort α, which (for simplicity) has no productive value but allows him to impress the supervisor favorably. In doing so, the agent knows that the supervisor who observes y will write a report stating that the agent performed $y + \alpha$. Since the agent's remuneration depends only on the supervisor's report (the principal receives no other information), it can be written $W = w + b(y + \alpha)$. Let us assume that this agent's preferences can be described by the exponential function $U = -\exp\{-a[W - C(e) - K(\alpha)]\}$ in which the disutilities linked to efforts e and α are represented by the quadratic functions $C(e) = ce^2/2$ and $K(\alpha) = k\alpha^2/2$; reasoning identical to that followed in the basic agency model arrives at the following expression of expected utility:

$$\mathbb{E}U = -\exp\left\{-a\left[w + b(e+\alpha) - C(e) - K(\alpha) - \frac{ab^2\sigma^2}{2}\right]\right\}$$

The reader will see that the levels of effort e^* and α^* chosen by the agent are such that $C'(e^*) = K'(\alpha^*) = b$. In this simple model, the agent equates the marginal costs of the two types of effort to the piece rate. Let $\overline{U}$ again be the reservation utility of the agent, and let us posit $\bar{x} = -\ln(-\overline{U})/a$; taking the logarithms of the opposites of the two sides of the participation constraint $\mathbb{E}U \geq \overline{U}$, we find that the latter constraint comes down to the inequality:

$$w + b(e+\alpha) - C(e) - K(\alpha) - \frac{ab^2\sigma^2}{2} \geq \bar{x}$$

Since the agent's production is given by $y = e^* + \varepsilon$ and his remuneration amounts to $W = w + b(e^* + \alpha^* + \varepsilon)$, the principal's expected profit, or $\mathbb{E}(y - W)$, is equal to $(1-b)e^* - b\alpha^* - w$. The principal then decides on her remuneration rule by maximizing her expected profit subject to incentive and participation constraints, or:

$$\max_{\{w,b\}}[(1-b)e^* - b\alpha^* - w] \quad \text{subject to } C'(e^*) = K'(\alpha^*) = b \quad \text{and} \quad \mathbb{E}U \geq \overline{U}$$

Since the principal always has an interest in choosing the fixed part w of the compensation scheme in such a way as to bind the agent's participation constraint, we can carry the value of w thus obtained into the expected profit. We then see that the optimal value of parameter b solves:

$$\max_{b}\left[e^* - C(e^*) - K(\alpha^*) - \frac{ab^2\sigma^2}{2} - \bar{x}\right] \quad \text{subject to } C'(e^*) = K'(\alpha^*) = b$$

The Inefficiency of Performance Pay

Given the quadratic cost functions, we easily find that the optimal remuneration rule b^* is characterized by:

$$b^* = \frac{1}{1 + \frac{c}{k} + ac\sigma^2} \quad \text{with } e^* = \frac{b^*}{c} \quad \text{and} \quad \alpha^* = \frac{b^*}{k}$$

These equalities show that the principal takes "rent-seeking" activity into account by reducing the piece rate b^*. The more profitable rent-seeking is to the agent, that is, the weaker parameter k is, the less is paid for performance. In other words, rent-seeking weakens the variable part of total remuneration and strengthens the fixed part; this increases the inefficiency bred by the moral hazard problem. Hence the first-best optimum is not reached with risk-neutral agents ($a = 0$), since even in this case $b^* < 1$ obtains.

This model is fairly rudimentary in that it says nothing about the remuneration of supervisors, who in the model have the passive role of simply transmitting the results they observe to the principal, with no a priori effect on what they themselves are paid. In actuality, supervisors may also be subject to a system of performance-based remuneration that could modify their behavior. The study of Bandiera et al. (2009), based on a controlled experiment concerning changes in managerial incentives in an English fruit-picking company, show that this is indeed the case. In this experiment, workers are paid by a piece-rate system (tied to the weight of the fruit gathered), but the managers of teams of pickers are remunerated either with a fixed wage or by a bonus proportional to the overall quantity of fruit harvested by their team. Bandiera et al. observed that when managers are paid a fixed wage, they have a tendency to show favor to workers with whom they feel "socially connected." In their study, social connections are measured by having the same country of origin, living in the same neighborhood, or even whether the manager and worker arrived at the company at the same time. Bandiera et al. find that when managers are receiving a fixed wage, socially connected workers benefit from more managerial attention, which has the effect of increasing their productivity and thus winning them larger pay packets, since workers are paid on a piece-rate basis throughout this experiment. The attitude of managers changes radically when their remuneration switches from a fixed wage to bonuses based on the overall output. Bandiera et al. then observe that there is no longer any correlation between the productivity of workers and their social connections to their managers. This example illustrates yet again the influence of incentives when results are adequately verifiable and shows that the quality of the incentives put in place at higher organizational levels can have an effect on performances at lower levels.

Conversely, this model also highlights the danger that may arise from basing remuneration solely on verifiable results. When an employee's performance is evaluated solely on the basis of *verifiable* data, there is a strong risk of provoking an inefficient allocation of that agent's efforts because he will begin to focus his efforts exclusively on activities that will pay off, given the criterion being used to evaluate performance. In the preceding model, this criterion is simply performance *reported*, not observed, by supervisors; but the model actually applies to other situations too. For example, when it comes to the problem of multitasking, variable α can be interpreted as a particular effort intended to boost a specific indicator, upon which the agent's remuneration is in

part based. Prendergast (1999) adduces a number of situations, from doctors paid on a fee basis to educational institutions rewarded for the number of degrees they grant, in which a system of "objective" compensation is a cause of inefficiency.

As we will see in the following section, one remedy for these detrimental outcomes, which surface when neither effort nor performance can be verified, lies in constructing systems of promotion based on the relative performance of agents and/or grounding long-term relationships on implicit contracts; the latter are sometimes called incomplete contracts, or informal relationships.

4 INCENTIVE IN THE ABSENCE OF VERIFIABLE RESULTS

In this section we assume that both the effort made and the results achieved by an agent are unverifiable. If we look again at the static agency model under these hypotheses, a double problem of moral hazard emerges, since the employee can no longer a priori trust her employer when the latter promises to pay a high wage in exchange for good performance: if remuneration increases with observed production, the employer always has an interest in declaring that he has observed the lowest level of production, so as to pay the lowest wage possible. This possible difficulty is met by invoking the notion of *reputation*: a firm could not behave in this way because its employees would inevitably quit and would spread the news that the firm was behaving in this way; the firm would then have greater difficulty in recruiting new workers (the models of Bull, 1983, 1987, take up this idea). Another approach takes the view that when a relationship lasts for more than one period, that means the two parties have a mutual interest in it. The contract that binds them is thus *implicit* and *self-enforcing*.

We will see that this last approach does allow us to understand several important features of wage relationships and the functioning of the labor market in the absence of verifiability of results (see Chiappori et al., 1994, for more details). In the first place, the occurrence of double moral hazard in this context explains the use of promotions, following a hierarchical logic that is very different from the logic that links remuneration directly to performance or productivity. The double moral hazard also accounts for the existence of compensation rules based on seniority, which are frequently observed in firms. Finally, the inefficiency induced by the double moral hazard may, in certain circumstances that the efficiency wage model of Shapiro and Stiglitz (1984) illustrates, be the source of involuntary unemployment.

4.1 Promotions and Tournaments

Following the seminal work of Doeringer and Piore (1971), many studies have highlighted compensation rules specific to large firms and known as the *internal market*. Large firms appear to adopt rules that are seemingly quite unconnected with the outside, and supposedly competitive, world. Among other things, these rules define the systems of promotion, the positions, and the wages that go with them. Wages seem to follow a hierarchical logic, largely independent of the productivity of labor. A wage raise generally goes along with a *promotion*, when the agent changes position in the

hierarchy. In some large firms the salary of the CEO is more than three times higher than that of the vice presidents. A gap that large would seem to indicate that the internal market of a firm is a structure that allows a solution to certain problems of incentive. "Tournament" theory makes it possible to explain some of the properties of internal markets by linking wages to the hierarchical grades at which agents arrive according to their relative performance.

4.1.1 A Tournament Model

Tournament theory starts with the idea that the principal creates competition among his agents by, on one hand, promising them prizes specified *in advance*, and on the other, making it clear to them that the awarding of these prizes will depend not on the *absolute* level of an individual's production but on the place that this level occupies *relative* to that of the other competitors. The model of reference is that of Lazear and Rosen (1981), but here we use a slightly more general one, close to that of Malcomson (1984), which has the advantage of fitting better with the foregoing analyses of optimal remuneration rules.

The Rules of the Game

In analyzing the properties of a promotion system, one ought to use an explicitly dynamic model. But we prefer to avoid the excessive analytical complication to which that option leads and will therefore make do with a static model: a large firm in which a given number N of employees each produce a quantity $y = e + \varepsilon$ of goods, where ε is a normally distributed random variable with zero mean and standard error σ, proper to the individual in question. To simplify the notation, we do not index individuals, and we assume that random variables affecting individual production are independent. The N employees receive a given fixed wage w_0 and they all aspire to promotion, in which case they will receive wage $w_0 + b$, $b > 0$. The purpose of this simple formalization is to make it clear that the hierarchical structure of the firm rests on a given grid of remunerations in which wages change discontinuously, and only with promotion. The principal chooses the number L of those promoted and the value b of the "bonus" that comes with promotion. The tournament unfolds according to an extremely simple rule: the principal announces that he will offer a promotion to the L persons who have performed best. We will solve the principal's problem and show that there is no difference between choosing the number of those promoted L and the value of the bonus b, or the minimal level of production $\bar{y}$ that qualifies an agent for promotion and the bonus b.

On the Value of Promotions When Individual Effort Is Unverifiable

When an individual's production cannot be impartially assessed by a third party, the advantage of the tournament in comparison with other kinds of incentive is that it only contains verifiable clauses. The number L of those to be promoted and the wage $w_0 + b$ that each will receive are known before the competition begins, and an impartial tribunal can easily determine whether the prescribed promotions have in fact taken place and whether every employee has been paid according to the agreed wage scale. Moreover, the firm has no a priori interest in lying about the possible finishing order, since in any case it pays the same wage bill $w_0N + bL$, which is likewise known beforehand. At most, the firm's management might favor "pet" candidates, but it cannot change the number of promotions or the value of the total wage bill. In this sense, promotions

constitute a simple way for the employer to commit himself to pay the bonuses he has promised, since the value of all the bonuses is verifiable.

Thus we see that promotions and internal markets in general allow the clauses of a contract to be made *explicit*. Just as in a tournament, the rules, the different stages of the game, and the rewards are made perfectly clear at the outset, and are verifiable. The wages corresponding to each grade in the hierarchy are totally uncoupled from the productivity of labor, and it is the number of promotions and the wage gap between the different rungs that, if correctly calibrated, constitute an optimal incentive scheme. In other words, your superior does not earn twice as much as you because she is twice as productive but because that fact will give you reason to put plenty of effort into your current assignment, in the hope of climbing the rungs of the hierarchy.

The Behavior of the Agent

To simplify the analysis, and in order to concentrate solely on the characteristics of the internal market, we will suppose that all agents are risk neutral. More precisely, the utility function of an agent is simply written $U = W - C(e)$, where the cost of effort is measured by the quadratic function $C(e) = ce^2/2$. Given the proposed compensation scheme, each agent knows that she will receive wage w_0 whatever her level of production may be and that she will, in addition, be entitled to bonus b only if her production is greater than $\bar{y}$, or $\varepsilon \geq \bar{y} - e$. Let $\Phi(.)$ be the cumulative distribution function of the random variable ε; this event will happen with a probability equal to $[1 - \Phi(\bar{y} - e)]$, when an employee supplies effort e. Her expected utility is then written:

$$\mathbb{E}U = w_0 + b\,[1 - \Phi(\bar{y} - e)] - C(e) \tag{6.33}$$

Knowing b and $\bar{y}$, every agent chooses the level of effort e^* that maximizes her expected utility. Let $\phi = \Phi'$ be the probability density function of the disturbance ε; we then find that e^* is the solution to:

$$b\phi(\bar{y} - e^*) = C'(e^*) \tag{6.34}$$

It is easy to verify that with a normally distributed disturbance, relation (6.34) defines a unique value of effort $e^* = e^*(b, \bar{y})$ increasing with bonus b, but with a direction of variation which is ambiguous with $\bar{y}$ (this direction depends on the sign of $\bar{y} - e^*$). We can also verify that the second-order conditions dictate $b\phi' + C'' > 0$.

The Behavior of the Principal

If we assume, for the sake of simplicity, that the total production of the firm is the sum of individual productions, the expected profit per capita is:

$$\mathbb{E}\pi = e^* - w_0 - b\,[1 - \Phi(\bar{y} - e^*)] \tag{6.35}$$

The principal determines b and $\bar{y}$ in such a way as to maximize this profit per capita, taking into account the incentive constraint (6.34) and the participation constraint $\mathbb{E}U \geq \overline{U}$, where $\overline{U}$ again designates an exogenous level of utility accessible outside the firm. This problem is simple to solve if we limit ourselves at the outset to the

values of variables b and $\bar{y}$ which make the participation constraint binding. In this case, relation (6.33) shows that b and $\bar{y}$ always verify:

$$w_0 + b\,[1 - \Phi(\bar{y} - e^*)] = \overline{U} + C(e^*) \tag{6.36}$$

If we carry this equality into definition (6.35) of expected profit per capita, we get $\mathbb{E}\pi = e^* - C(e^*) - \overline{U}$. The maximization of this expression yields:

$$C'(e^*) = 1 \Leftrightarrow e^* = \frac{1}{c} \tag{6.37}$$

With the help of relations (6.34) and (6.36), we see that this level of effort can be attained by choosing a production norm $\bar{y}$ and a bonus b satisfying:

$$b\phi(\bar{y} - e^*) = 1 \tag{6.38}$$

Equations (6.38) and (6.36) define the optimal values of $\bar{y}$ and of b, given the value of e^* yielded by (6.37). Since all the workers whose individual production surpasses $\bar{y}$ are promoted, the number of promotions is defined by $L = N\,[1 - \Phi(\bar{y} - e^*)]$. It therefore makes no difference to the principal whether he proposes a contract stipulating the bonus and the minimal value of production that will trigger a promotion or the bonus and the number of promotions that will be made.

Note that the system of promotions through the ranks of a preestablished hierarchy provides each competitor at the outset with an average gain equal to what she could achieve otherwise, that is, $\overline{U}$. But in the aftermath, the winners of the tournament—those promoted—obtain a level of utility greater than that of the losers. If the latter remain with the firm, that is because they still have hope of being promoted in an upcoming tournament. This point could be taken into account in an explicitly dynamic model in which workers participate in a number of successive tournaments (see Meyer, 1992).

Increasing Risk

The system of relations (6.38) and (6.36) also furnishes some interesting details about the effects of increased uncertainty. This eventuality can be schematically likened to an increase in the complexity of the organization, which makes individual supervision more random. In this interpretation, the standard error σ must be an increasing function of the size N of the firm.

The consequence of increased uncertainty can be analyzed by approximating the solution defined by equations (6.38) and (6.36). Let us assume that the gap between $\bar{y}$ and e^* is not too large. Since $\Phi(0) = 1/2$, and since the probability density of a normal variable satisfies $\phi(0) = 1/\sigma\sqrt{2\Pi}$ and $\phi'(0) = 0$, a first-order expansion around the mean gives:

$$\phi(\bar{y} - e^*) \simeq \phi(0) = 1/\sigma\sqrt{2\Pi} \quad \text{and} \quad \Phi(\bar{y} - e^*) \simeq \frac{1}{2} + (\bar{y} - e^*)\phi(0)$$

Relations (6.38) and (6.36) then entail:

$$b \simeq \sigma\sqrt{2\Pi} \quad \text{and} \quad [1 - \Phi(\bar{y} - e^*)] \simeq \frac{C(e^*) + \overline{U} - w_0}{\sigma\sqrt{2\Pi}}$$

Readers will see that an increase in uncertainty, here deemed equivalent to a rise in σ, amplifies the wage gap and reduces the proportion $(1 - \Phi)$ of promotions. Hence, there ought to be few promotions in organizationally complex firms, in which the assessment of individual performances is imprecise, or in those in which "chance" plays a significant role—but the promotions that do occur ought to be accompanied by a strong increase in remuneration. To the extent that the standard error σ increases with the size N of the firm, this model also predicts that the level of compensation should increase with the number of individuals who aspire to a promotion. Note that these results have been reached on the assumption that agents are risk neutral. Aversion to risk, on the other hand, would have the effect of reducing the gaps between the various grades of the hierarchy. Examination of wage policies and promotion rules in certain large firms confirms this prediction. But before presenting a few illustrations of these results, it will be instructive to reflect on the limitations of promotion based on performance.

Tournaments and Rent-Seeking

The tournament model formalizes and simplifies a system of promotions based on the respective performance of agents. But there exist many organizations, including certain large industrial firms, in which promotions are made essentially on the basis of seniority. It would seem that a hierarchy in which seniority is the preponderant factor must lead to an inefficient allocation of resources, since agents no longer have an incentive to make great effort. The seniority rule, like many other so-called bureaucratic rules, is partially explained by the fact that it makes it possible to avoid rent-seeking activity. The ground for this conclusion can easily be shown by crossing the rent-seeking model with the tournament model. To that end, let us suppose that each agent can put out respectively an effort α which does no more than impress the supervisor and an effort e which only increases her individual production, still given by $y = e + \varepsilon$. Let us also assume that an agent's promotion depends only on the performance $y + \alpha$ reported by her supervisor. As before, the principal chooses the number L of those to be promoted and the value b of the bonus corresponding to the promotion. But now the principal announces that he will offer a promotion to the L persons who have the best performance as *reported* by the supervisors (since, by hypothesis, the principal delegates the observation of results to supervisors).

Let $\bar{r}$ be the level of reported performance that triggers a promotion, and let $K(\alpha) = k\alpha^2/2$ be the cost linked to rent-seeking activity; the expected utility of the agent, who is promoted if her performance exceeds $\bar{r}$, is now written:

$$\mathbb{E}U = w_0 + b\,[1 - \Phi(\bar{r} - e - \alpha)] - C(e) - K(\alpha)$$

Knowing b and $\bar{r}$, the agent chooses levels of effort e^* and α^*, which maximize her expected utility. We thus have:

$$b\phi(\bar{r} - e^* - \alpha^*) = C'(e^*) = K'(\alpha^*) \qquad \textbf{(6.39)}$$

Relation (6.35) giving the expression of profit per capita here takes the form:

$$\mathbb{E}\pi = e^* - w_0 - b\,[1 - \Phi(\bar{r} - e^* - \alpha^*)]$$

Likewise relation (6.36) giving the values of b and $\bar{r}$, which make the participation constraint binding, is here written:

$$w_0 + b\,[1 - \Phi(\bar{r} - e^* - \alpha^*)] = \overline{U} + C(e^*) + K(\alpha^*) \tag{6.40}$$

Bringing this last equality into the expression of profit per capita, we find $E\pi = e^* - C(e^*) - K(\alpha^*)$. The maximization of this expression leads to levels of effort e^* and α^* characterized by:

$$e^* = \frac{1}{c(1 + \frac{c}{k})} < \frac{1}{c} \quad \text{and} \quad \alpha^* = \frac{1}{c + k} > 0 \tag{6.41}$$

Finally, the performance norm $\bar{r}$ and the bonus b are found by substituting these values of e^* and α^* in equations (6.39) and (6.40). From that we can deduce the number of promotions proposed by the principal, which is given by $L = N[1 - \Phi(\bar{r} - \alpha^*)]$.

Relations (6.41) show that effort e^* (or α^*) increases (or decreases) with the cost k of rent-seeking. This means that rent-seeking reduces the effort dedicated to production. If k is small with respect to c, rent-seeking activity pays off handsomely for the agent, while the firm's interest in staging the tournament—that is, a system of promotions based on performance—is lessened, since productive effort falls off. In practice, above a threshold of minimum verifiable effort, it may be in the firm's interest to abandon the system of promotion based on performance for a system based on seniority, which does not elicit rent-seeking activity and probably also makes it possible to save a portion of the supervision costs. That assumes, however, that a system based on seniority is capable of giving wage earners sufficient motivation through adequate sanctions or incentives if productivity drops too low.

More on Promotions

The models we have used in this part are very simple. They only illuminate a portion of the logic of promotions and would need to be extended in various directions. Some research stresses the notion that promotions send a signal about the quality of employees, making it possible to assign them to the tasks best suited to their abilities (Waldman, 1984; Sattinger, 1993). This would explain the importance of the wage gains that generally go along with promotions. Higher pay with promotion keeps workers whose good qualities would be signaled to other employers by their promotion from quitting the firm. Using a longitudinal data set that contains detailed information concerning the internal labor market history of a medium-sized firm in the financial services industry in the United States, DeVaro and Waldman (2012) find some support for the idea that promotions serve as a signal of worker ability. Promotions are also a way to give workers incentive to accumulate specific human capital (Carmichael, 1983; Prendergast, 1993a; Chang and Wang, 1996). Finally, promotions may also be explained by uncertainty about the efficiency of employees. Harris and Holmström (1982) consider a situation in which the quality of every employee is uncertain and is gradually revealed by her performance. If the worker is risk averse, a risk-neutral firm ought to have an interest in insuring her against this uncertainty by paying her a constant wage, dependent on her expected efficiency at the time of hiring. The most efficient workers, however, would then be given an incentive to look for other jobs, since other employers, observing their quality as

revealed by their past performance, would be ready to offer them higher wages. In consequence, the firm has an interest in offering limited insurance and in working out a system of promotion with a low starting wage and steeper wage rises as justified by performance.

The model of Gibbons and Waldman (1999a) takes up the learning process of Harris and Holmström (1982) and adds the acquisition of human capital and the assignment of employees to tasks adapted to their abilities. This model, which integrates several dimensions of an individual's career within a firm, reproduces well the main results of empirical studies of the subject, like, for example, significant wage rises accompanying promotion or the existence of "fast tracks" in which an individual who has been rapidly promoted to one grade in the hierarchy is then promoted rapidly to the next one (on the subject of careers, see the comprehensive panorama of Gibbons and Waldman, 1999b).

4.1.2 Empirical Illustrations

The predictions of the tournament model have often been tested in the realm of sport. As we would expect, studies show that golfers hit the ball more carefully, and racing drivers take greater risks, when the prizes offered are bigger (see Prendergast, 1999, who points out, however, that these studies are rather confirmations of the general principles of the theory of incentive than of the tournament model). In economics, the tournament model has relevance when applied to the properties of hierarchical structures and the wages linked to each grade.

The study carried out by Baker et al. (1994a, b) on a large American firm in the service sector, for which the data available covered the period 1969–1988, sharpens and confirms certain predictions of the preceding models. In the first place, this study shows that the relative weight of each grade in the hierarchy remains very stable. Whereas the firm tripled in size over the period in question, the rates of promotion from one grade of the hierarchy to another hardly varied at all. Second, figure VI in Baker et al. (1994a), reproduced in figure 6.3, indicates that the average wage corresponding to each grade increases at a rising rate as one moves up the hierarchy. This property accords with the size effect highlighted in the tournament model, according to which compensation increases with the number of individuals aspiring to promotion. In the firm in question, the number of employees decreased very gradually from level 1 up through level 4 (the four lowest levels) and, as we see, average wage growth is small in this part of the hierarchy. However, the relative size of each grade falls off very sharply between levels 5 and 8 (in 1980 there were 86 people in grade 5, 25 in grade 6, 4 in grade 7, and 1 in grade 8). As figure 6.3 shows, the more competitors there are in relation to the number of posts available in the next highest grade, the more steeply the average wage climbs.

Figure 6.3 also brings out the fact that the wage does not remain constant within each grade of the hierarchy. In other words, certain individuals (even the majority) see their wage rise without being promoted; this means that there are incentive mechanisms other than the tournament at work within each hierarchical grade.

The conclusions reached by Eriksson (1999) point in the same direction as those of Baker et al. (1994a, b). Using a sample group of 2,600 managers taken from 210 Danish firms who were followed from 1992 to 1995, Eriksson estimates that hierarchical grade explains 60% of the variation in wages. He also confirms certain predictions of the tournament model, finding that the "prize" awarded (i.e., bonus b in the theoretical model)

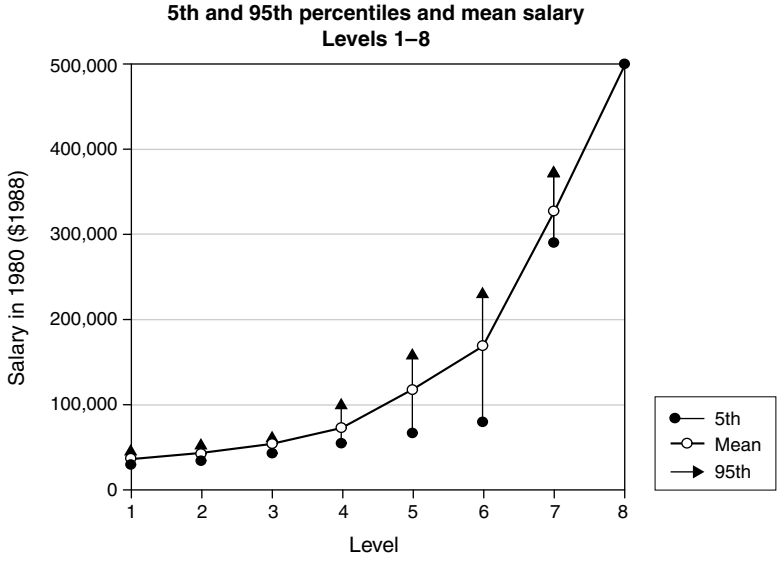


FIGURE 6.3
Remunerations and hierarchical grades.

Source: Baker et al. (1994a).

increases when the number of competitors rises. Further, he highlights a significant relationship between the variability of demand addressed to a sector and the dispersion of wages within that sector. This conclusion accords with the prediction of the theoretical model that bonus b increases with standard error σ characterizing the distribution of wages. Mention should also be made of McCue (1996), a study of persons employed in the state of Michigan that finds that *internal* mobility—that is, successive promotions within the same firm—explains around 15% of the wage rise for men over the life cycle.

More recently, DeVaro (2006) has estimated a structural model of promotion tournaments relying on the data of the Multi-City Study of Urban Inequality, which is a cross-sectional employer telephone survey of 3,510 establishments carried out between 1992 and 1995 in four metropolitan areas: Atlanta, Boston, Detroit, and Los Angeles. The results prove to be coherent with the implications of the theoretical model presented above. DeVaro finds that promotions are determined by relative performance, that the effort of a worker is increasing in the wage spread attached to promotions, and that the wage spread is increasing with the stochastic component of performance.

4.2 SENIORITY AND INCENTIVES

We will now take into explicit consideration the dynamic dimension of the wage relationship in a context where production is not verifiable; our purpose is to illustrate the interplay between seniority and incentives, which was emphasized in the pioneering

paper of Lazear (1979) and followed up in Lazear (2011). We can characterize optimal long-term contracts very simply using the "shirking" model, which we begin by laying out; the links among incentives, seniority, and human capital emerge clearly with the help of this model.

4.2.1 The Shirking Model

The shirking model assumes two levels of effort: the first strictly positive and again denoted by e, which gives rise to a disutility $C > 0$ and allows the agent to realize production $y_t > 0$ at date t; and the second with a value of 0, the disutility and the production of which are both normalized to 0 as well. Note that in this model it is *obligatory* for an agent furnishing level of effort e to achieve production y. So when the principal observes that production has taken the value 0, he can be sure that the employee has been shirking. If production were a verifiable magnitude, the principal could arrange this employee's remuneration as follows: the payment of a fixed wage w such that the participation constraint would be binding when the employee was not caught shirking, and a low wage (or even a penalty) w_1 when inspection found that she was not furnishing effort $e > 0$. But when individual production is not a verifiable magnitude, it becomes necessary to invent other remuneration rules.

The shirking model takes the dynamic dimension of the wage relationship into account and assumes that the principal proposes a contract $\{w_t;\, t = 0, 1, .., +\infty\}$ specifying the wage the employee will receive on each date. If the agent is caught shirking, that is, her work is inspected and she is found to be furnishing a null effort, she is paid to the end of that period and is fired. Note that the shirker receives her wage for that period even if she has not supplied any effort during it. This is an offshoot of the unverifiable character of production, which prevents the employer from proposing a remuneration based on results. In what follows, we will use the exogenous constant parameter $p \leq 1$ to designate the probability that the principal will inspect the agent's activity in each period. This less-than-perfect supervision ($p \neq 1$) is explained by the costs arising from checking up on the activities of employees—costs that are likely to be greater in a large firm. We will also assume that at each period the agent risks losing her job with an exogenous constant probability denoted by q.

The Behavior of the Principal

If $\delta \in [0, 1]$ designates the discount rate, the profit expected by the employer from the continuation of the contract after the t^{th} period, or Π_t, is written:

$$\Pi_t = y_t - w_t + \delta \left[(1 - q)\max(\Pi_{t+1}, \Pi^s_{t+1}) + q\bar{\Pi}_{t+1}\right] \tag{6.42}$$

In this expression, $\bar{\Pi}_{t+1}$ designates the profit expected when the contractual relationship winds up at the end of period t. This might, for example, be the profit expected in a "competitive" labor market or the profit derived from leaving the position empty. We shall assume that the employer considers this quantity as a parameter dependent on general macroeconomic conditions and outside his control. The term Π^s_{t+1} represents the expected profit of the principal if he decides to "cheat," in other words, to break the contract at the end of period t. Relation (6.42) is now easy to grasp. In the present

period, the principal obtains an instantaneous profit equal to $y_t - w_t$, but at the end of this period the job is destroyed with a probability q, in which case the employer expects the gain $\bar{\Pi}_{t+1}$. If the job is not destroyed, which happens with probability $1 - q$, the principal decides to respect the implicit contract when this attitude procures for him a gain Π_{t+1} superior to the gain Π^s_{t+1} that he would achieve by not abiding by the contract.

If, at period t, the employer decides to fire his employee by wrongly claiming that she has not supplied the required effort, his expected profit amounts to:

$$\Pi^s_t = y_t - w_t + \delta\bar{\Pi}_{t+1} \tag{6.43}$$

At each date, the employer respects the contract if doing so permits him to expect a profit greater than the one he would obtain by breaking the contract. For that, it is necessary and sufficient that the *employer's incentive constraint* $\Pi_t \geq \Pi^s_t,\ \forall t \geq 0$ be satisfied. Now, relations (6.42) and (6.43) entail:

$$\Pi_t - \Pi^s_t = \delta(1-q)\left[\max(\Pi_{t+1}, \Pi^s_{t+1}) - \bar{\Pi}_{t+1}\right], \quad \forall t \geq 0$$

We then easily verify that the incentive constraint $\Pi_t \geq \Pi^s_t,\ \forall t \geq 0$, is equivalent to condition $\Pi_{t+1} \geq \bar{\Pi}_{t+1}$, for all $t \geq 0$. Abiding by the contract also necessitates that the employer has no better alternative—a property that, as we have seen, characterizes the *participation constraint*. Since the gain expected by the principal at date t outside the contractual relationship amounts to $\bar{\Pi}_t$, the participation and incentive conditions finally come down to the inequalities $\Pi_t \geq \bar{\Pi}_t,\ \forall t \geq 0$.

The Behavior of the Agent

To focus our analysis more narrowly on the incentive problem, we will now assume that workers are risk neutral. That being the case, if an agent supplies effort $e > 0$ during the t^{th} period of the contract, she attains a level of utility equal to $w_t - C$ over the course of this period and, more generally, she expects an intertemporal level of utility V_t satisfying:

$$V_t = w_t - C + \delta\left[(1-q)\max(V_{t+1}, V^s_{t+1}) + q\overline{V}_{t+1}\right] \tag{6.44}$$

In this expression, V^s_{t+1} represents the expected utility of an agent who decides no longer to furnish effort e at period $t+1$; it is defined by relation (6.45) below. The term $\overline{V}_{t+1}$ designates the utility expected when the contractual relationship comes to an end after t periods. This corresponds to the utility expected from searching for a job. Relation (6.44) signifies that in the present period, the employee obtains instantaneous utility $w_t - C$, but that at the end of this period the probability is only $1 - q$ that the job will still be there. If it is, she decides to furnish effort $e > 0$ at date $t+1$ if doing so procures for her a utility V_{t+1} greater than the utility V^s_{t+1}, which she would get by not producing this effort. But if the job is destroyed, which happens with probability q, the employee then obtains a level of utility equal to $\overline{V}_{t+1}$.

When at the t^{th} period of the contract an employee shirks, she receives wage w_t but does not undergo the disutility C that comes with supplying effort e. As there is a

probability p of being monitored, in which case she will be fired, her expected utility is written:

$$V_t^s = w_t + (1-p)\delta\left[(1-q)\max(V_{t+1}, V_{t+1}^s) + q\overline{V}_{t+1}\right] + p\delta\overline{V}_{t+1} \tag{6.45}$$

An employer who wishes the agent to supply effort e at each period must find a way to make the *worker's incentive constraint* $V_t \geq V_t^s$ satisfy $\forall t \geq 0$. With the help of relations (6.44) and (6.45), we arrive at:

$$V_t - V_t^s = -C + p\delta\left[(1-q)\max(V_{t+1}, V_{t+1}^s) + q\bar{V}_{t+1}\right] - p\delta\overline{V}_{t+1}$$

We then easily verify that the incentive constraint $V_t \geq V_t^s$, $\forall t \geq 0$ is equivalent to condition:

$$V_{t+1} - \overline{V}_{t+1} \geq \frac{C}{p\delta(1-q)}, \quad \forall t \geq 0 \tag{6.46}$$

Rent and the Set of Feasible Contracts

At this stage it will be helpful to bring the notion of *rent* associated with the labor contract into sharper focus. In a general way, this term designates the difference between the gains procured by the contract and those that would flow from the best outside opportunity. In this case, for the agent the rent at date t is equal to $V_t - \overline{V}_t$, whereas for the principal, it amounts to $\Pi_t - \bar{\Pi}_t$. The incentive constraint (6.46) signifies in particular that in order to give an employee incentive to put out effort *today*, she must expect a strictly positive rent from doing so *tomorrow*. In this model, the incentive mechanism is forward looking and the wage w_t exerts no influence on the effort of period t. The incentive to furnish strong effort during this period comes from the prospect of the *future* gains specified by the contract, in other words the series of wages starting from date $t+1$. It is worth noting that, unlike future wages, the hiring wage plays no incentive role. The importance of this will emerge when we come to characterize the optimal contract.

Finally, in order for the employee to remain under contract at date t, it is also necessary that she not find a better alternative. This participation condition is given here by $V_t \geq \overline{V}_t$ for all $t \geq 0$. We immediately see that it is satisfied, *except at* $t = 0$, when the incentive constraint (6.46) is satisfied. The participation conditions thus dictate the only supplementary constraint $V_0 \geq \overline{V}_0$. In sum, the set P of levels of utility and profit attainable by using self-enforcing contracts is defined by:

$$P = \left\{ (\Pi_t, V_t) \,\middle|\, \Pi_t \geq \bar{\Pi}_t,\ V_{t+1} - \overline{V}_{t+1} \geq \frac{C}{p\delta(1-q)},\ V_0 \geq \overline{V}_0,\ \forall t \geq 0 \right\} \tag{6.47}$$

From now on we will simply refer to P as being the set of *feasible* contracts. The next step is to spell out the properties of optimal contracts. The characterization of optimal contracts is made a great deal easier by using the notion of surplus. We then see that the existence of a self-enforcing contract is equivalent to conditions which successive surpluses must satisfy and that the optimal contract does not offer any rent to the agent at the time of hiring.

Surplus and the Existence of a Self-Enforcing Contract

By definition, an optimal contract satisfies the incentive and participation constraints of the worker and the employer and maximizes, at every date, the expected profit of the principal. Let us first take a look at the conditions under which the incentive and participation constraints are satisfied. A useful notion in this context is that of *global surplus* at date t. Let S_t be the global surplus; it is equal by definition to the sum of the rents that the contract procures. We thus have:

$$S_t \equiv V_t - \overline{V}_t + \Pi_t - \bar{\Pi}_t, \quad \forall t \geq 0$$

Adding up relations (6.42) and (6.44), we get a difference equation that looks forward and that completely defines the series of surpluses. It is written:

$$S_t - \delta(1-q)S_{t+1} = y_t - C + \delta(\bar{\Pi}_{t+1} + \overline{V}_{t+1}) - (\overline{V}_t + \bar{\Pi}_t), \quad \forall t \geq 0 \tag{6.48}$$

We observe that wages do not appear in this equation. In consequence, the value of the surplus does not depend on the level of wages. This property follows from the hypothesis that principal and agent are both risk neutral, and it would not be verified with individuals who did present risk aversion. It makes possible a simple answer to the question of the existence of self-enforcing contracts. The right-hand side of relation (6.48) contains only variables considered as *exogenous* parameters by the partners to the contract. Consequently, the global surplus is also, at this stage, an exogenous parameter. Since by definition $\Pi_t - \bar{\Pi}_t = S_t - (V_t - \overline{V}_t)$ for all $t \geq 0$, the set P of feasible contracts described by relation (6.47) is also characterized in the following manner:

$$P = \left\{ V_t \,\middle|\, S_{t+1} \geq V_{t+1} - \overline{V}_{t+1} \geq \frac{C}{p\delta(1-q)}, S_0 \geq V_0 - \overline{V}_0 \geq 0, \forall t \geq 0 \right\} \tag{6.49}$$

This way of presenting the set of feasible contracts allows us to deal with the question of the existence of self-enforcing contracts easily. The fact is that for a contract of this type to exist, it is necessary and sufficient that the set P not be empty. Relation (6.49) shows that this condition is satisfied when the series of surpluses has well-defined lower bounds. To be precise, we have:

$$P \neq \emptyset \Longleftrightarrow S_0 \geq 0 \quad \text{and} \quad S_{t+1} \geq \frac{C}{p\delta(1-q)}, \quad \forall t \geq 0 \tag{6.50}$$

These inequalities show that an employer and a worker will agree on an implicit, self-enforcing contract when it offers them the opportunity to generate an overall nonnegative surplus over the *entire duration* of the contract, and strictly positive for every period $t \geq 1$. The initial period and the subsequent periods are different in kind because the incentive mechanism is forward looking. At the moment of hiring, it is sufficient that the surplus offered by the contract be simply positive, but at date $t \geq 1$, the surplus has to exceed quantity $C/p\delta(1-q)$, which is strictly positive, in order to give the agent incentive to supply effort e in all the periods subsequent to t.

In a world without moral hazard, a firm and a worker would have an interest in coming to terms when doing so allowed them to generate a nonzero surplus S_t at

every date. So moral hazard has the effect of restricting the set of feasible contracts, since conditions (6.50) show that it becomes necessary for surplus S_t to be greater than $C/p\delta(1-q)$ for every $t \geq 1$. Taking moral hazard into account thus induces a form of Pareto inefficiency, inasmuch as exchanges such that $0 \leq S_t \leq C/p\delta(1-q)$ at a date $t \geq 1$ (mutually advantageous ones, that is), will not be realized. This inefficiency can lead to the exclusion of some workers with low productivity from long-term contractual relationships.

Rent and the Optimal Contract

We can easily find the expression of the optimal contract if we remember that it is equivalent to setting values for w_t or V_t. Relation (6.44) shows, in fact, that there is a bijection between the series of wages and the series of intertemporal utilities. Formally, then, we can view the employer's decision variables as the employee's utility levels rather than wages. The definition of the global surplus entails that the expected profit be expressed in the form $\Pi_t = -V_t + (\overline{V}_t + \overline{\Pi}_t + S_t)$. Since the terms in parentheses in this equality are all exogenous parameters, the search for a self-enforcing contract maximizing profit Π_t at every date $t \geq 0$ is equivalent to minimizing intertemporal utility V_t over the set P of feasible contracts defined by (6.49). We then see that the optimal self-enforcing contract is characterized in the following manner:

$$\begin{aligned} &\text{(i) if } P \neq \emptyset, \text{ then } V_0 = \overline{V}_0 \quad \text{and} \quad V_{t+1} = \overline{V}_{t+1} + C/p\delta(1-q), \quad \forall t \geq 0 \\ &\text{(ii) if } P = \emptyset, \text{ no self-enforcing contract exists.} \end{aligned} \tag{6.51}$$

Thus, when the series of surpluses is such that the set P of feasible contracts is not empty, we have $V_0 = \overline{V}_0$ at date $t = 0$, which signifies that the optimal contract offers no rent to the worker at date $t = 0$. But at all subsequent periods, the agent obtains a gain V_{t+1} strictly superior to the external opportunity $\overline{V}_{t+1}$ of quantity $C/p\delta(1-q)$, which gives her an incentive, for one thing, to supply effort $e > 0$ and, for another, not to voluntarily quit the firm in which she is working. It should also be noted that the principal captures the entire surplus of the contractual relationship ($\Pi_0 - \overline{\Pi}_0 = S_0$) and never has an interest in breaking the implicit contract that ties him to the employee, precisely because this contract procures him more than the outside opportunity if the set P is not empty: we in fact have $\Pi_t - \overline{\Pi}_t = S_t - C/p\delta(1-q)$ for all $t \geq 1$.

These properties of the optimal contract suggest that the wage at the time of hiring plays a special role. We will now make this point clear by relating it to the role of seniority in the wage profile.

4.2.2 The Deferred Payment Mechanism

The accumulation of human capital and experience can cause wages to rise with seniority (see chapter 4, particularly sections 2.3 and 4.1). But an increasing link between wages and seniority also constitutes an incentive mechanism for newcomers to an enterprise. Such a "deferred payment" mechanism would indeed appear to be a response to the problem of incentivation which firms face in the shirking model set out above. Hence the problems raised by moral hazard may also lead firms to adopt the deferred payment system, with its increasing relationship between wages and seniority.

The Reasons That Wages Rise with Seniority

In the first place, the improvement in human capital that comes with the acquisition of knowledge and skill increases productivity, and this in itself is an explanation for the wage profile over the course of careers (Mincer, 1974; Becker, 1975; see chapter 4). Specifically, the accumulation of *general* human capital of the sort than can be put to use in a wide range of firms ought to lead to an increasing relationship between *experience* in the labor market and wages. Conversely, the accumulation of specific human capital of the sort that can only be put to use in one particular job may lead in certain circumstances to an increasing relationship between *seniority* in the firm and wages.

The existence of information problems may also explain an increasing relationship between experience or seniority and wages. In the equilibrium job search models laid out in chapter 5, workers knew the distribution of wages and had access to a limited number of job offers per unit of time. Within this framework, wages rise with experience, for the probability of receiving a job offer from a firm proposing a high wage rises the longer an individual has been present in the labor market. Better knowledge of an employee's characteristics, which makes it possible to assign him to tasks at which he is most efficient, also constitutes a reason for wages to rise with seniority (Jovanovic, 1979; MacDonald, 1982). Finally, problems of incentive contribute to the existence of an increasing relationship between wages and seniority. In this connection, Lazear (1979, 1981) has advanced the proposition that a system of "deferred payment," in which workers get low pay at the outset of their careers but a promise of generous remuneration towards the end of them, constitutes a simple and particularly efficient incentive mechanism.

We will demonstrate that the mechanism of deferred payment and the role of human capital in wage-earning careers are well illustrated by the shirking model presented above. Further, we will see that empirical investigation generally finds that experience and seniority do have a positive effect on wages, but it does not actually pinpoint the causes of this increasing relationship.

The Optimal Wage Profile in the Shirking Model

Let us return to the shirking model to show how the mechanism of deferred payment emerges naturally as a solution to the incentive problem facing the firm. With some simple calculations, relations (6.44) and (6.51) that define the optimal contract allow us to express optimal wages in the following manner:

$$w_0 = \overline{V}_0 - \delta \overline{V}_1 + C - \frac{C}{p} \quad \text{and} \quad w_t = \overline{V}_t - \delta \overline{V}_{t+1} + C + \frac{C}{p}\left[\frac{1}{\delta(1-q)} - 1\right] \tag{6.52}$$

We see that the series of optimal wages is linked quite simply to the levels of utility associated with outside opportunities. To highlight this linkage more tellingly, let us assume that at each period t these outside opportunities procure an instantaneous level of satisfaction $\bar{w}_t - C$, where $\bar{w}_t$ represents the "outside" wage. Thus we have:

$$\overline{V}_t = \sum_{i=0}^{+\infty} \delta^i (\bar{w}_{t+i} - C), \quad \forall t \geq 0 \tag{6.53}$$

If the human capital accumulated by an individual is of the *general* kind, he can expect ever larger external gains, and the series of $\bar{w}_t$ will be increasing (see for example Harris and Holmström, 1982). Conversely, if the human capital accumulated by an individual is of the *specific* kind, the effect on outside opportunities will be zero, and the series of $\bar{w}_t$ will not be increasing (Jovanovic, 1979, uses a model grounded on this hypothesis). It is possible to link the optimal wage profile to the series of outside wages $\bar{w}_t$. Relation (6.53) does in fact entail that $(\overline{V}_t - \delta\overline{V}_{t+1})$ is equal to $\bar{w}_t - C$. Equations (6.52) giving the optimal wages are then written:

$$w_0 = \bar{w}_0 - \frac{C}{p} \tag{6.54}$$

$$w_t = \bar{w}_t + \frac{C}{p}\left[\frac{1}{\delta(1-q)} - 1\right], \quad \forall t \geq 1 \tag{6.55}$$

These last two relations show that the contractual wage is inferior to the outside wage ($w_0 < \bar{w}_0$) at career onset but then overtakes it in the subsequent stages ($w_t > \bar{w}_t$). We also observe that starting from date $t = 1$, the wage increases (or diminishes) with the same frequency as outside opportunities. It is thus clear that only general human capital influences the observed wage, through its impact on outside opportunities. Specific human capital, on the other hand, which by definition has no influence on the outside wage, does not affect the observed wage. This result flows from the hypothesis that the worker has no bargaining power: the employer here unilaterally proposes a labor contract which binds the participation constraint of the worker. If the worker did have nonzero bargaining power, the participation constraint would not be binding and specific human capital would exert a positive influence on the wage.

In figure 6.4 we depict the properties of an optimal wage profile, on the hypothesis that levels of outside utility will be increasing. This figure brings out a particular form of the "deferred payment" mechanism, the theory of which was elaborated by Lazear (1979, 1981) in particular. In this mechanism, the workers with the most seniority in a firm would be paid at a rate that surpassed their marginal productivity, while the workers with less seniority would be paid at a rate falling short of their marginal productivity. This arrangement gives the workers hired most recently an incentive to furnish the efforts demanded of them in order to stay with the firm long enough to get the benefit of the wages reserved for "old hands." In our model, the mechanism of deferred payment takes an extreme form, since only the hiring wage w_0 is less than the competitive wage $\bar{w}_0$ which notionally reflects the agent's marginal productivity. Note that the discontinuity between the hiring wage and the subsequent wages would be attenuated if we were to take into account a certain heterogeneity in hires and in the amount of time needed for the firm to get a clear picture of the abilities of its workers. These elements would have the effect of spreading out the deferred payment mechanism more evenly over time, and so the wage profile would show a pattern of increase more like the empirical observations that we examine below.

It is also evident that the incentive constraints impose a steep slope on the wage profile mainly at career *onset*, after which it is rather the participation constraints that influence this profile. Now the levels of utility that an individual can expect outside

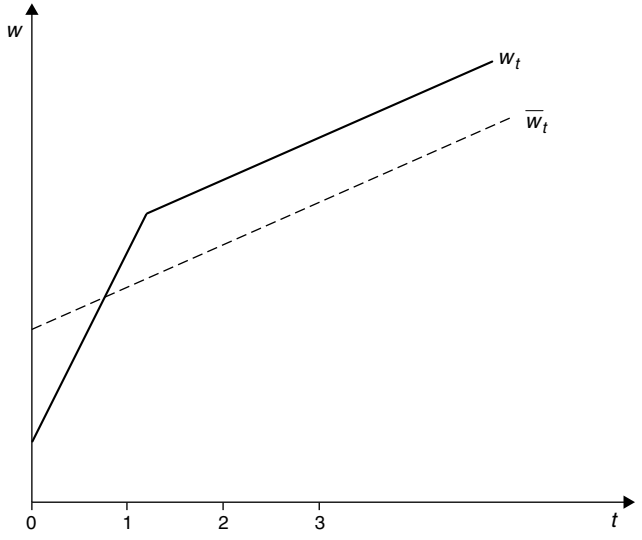


FIGURE 6.4
The optimal wage profile with general human capital and deferred payments.

his current job evolve with changes in his human capital. If this capital grows little or not at all over time, the wage profile ought to flatten out quickly; the converse is true if the value of human capital increases over time. In other words, the positive effect of the incentive constraints on wage growth should be felt mainly at the onset of an individual's career, and that of the accumulation of human capital at a later stage.

Empirical Elements Concerning the Mechanism of Deferred Payment
Some empirical studies suggest that certain firms do in fact adopt wage policies based on the deferred payment mechanism. Lazear and Moore (1984), for example, compare the incomes of independent workers with the incomes of workers who are a priori identical but carry out similar functions within a firm of which they are employees. Since the independent worker has no need to give herself incentive to make the necessary effort, her income profile ought to be identical to marginal productivity. Lazear and Moore find that the wage profile of employees is steeper than that of independent workers, which points to the conclusion that the mechanism of deferred payment is a way of giving employees incentive to make the desired efforts.

The work of Kotlikoff and Gokhale (1992) is even more convincing. Using data for the period 1969–1983 covering a sample of 300,000 employees of large North American firms specializing in sales, these authors achieve a reconstruction of the productivity profile of an employee from the time he enters a firm, based on the wage of new entrants. The idea is that when hiring new workers employers equate the expected value of a worker's compensation to the expected value of his productivity. Data showing how expected compensation varies with the age of hire then supplies indications about how productivity varies with age. Kotlikoff and Gokhale infer the age-productivity profiles

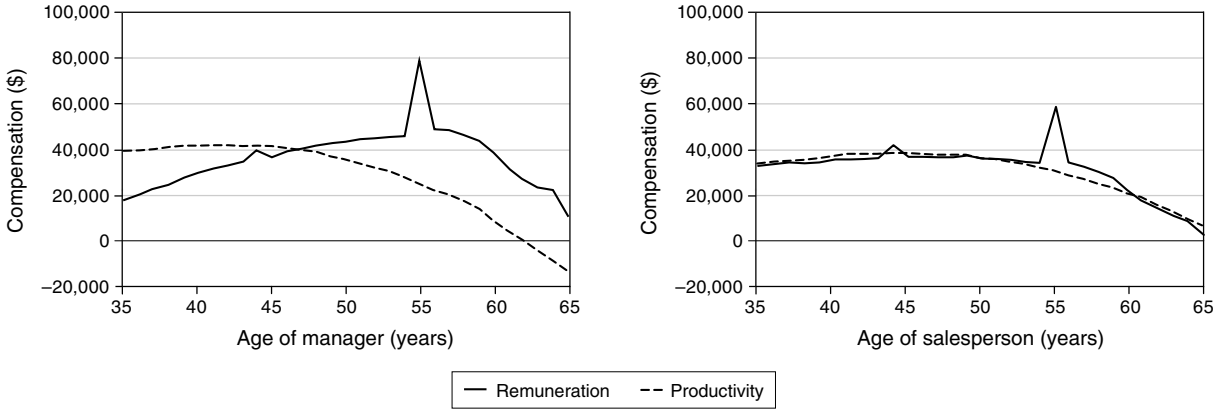


FIGURE 6.5
Profiles of remuneration and productivity.

Source: Kotlikoff and Gokhale (1992, figures 3 and 4).

by using data on the present expected value of earnings of new hires. Figure 6.5 shows the wage profiles and productivity of a manager and a salesperson between ages 35 and 65, that is, over a 30-year career with a firm.[7] The manager's wage profile conforms to the theory of deferred payment. It approximately lags marginal productivity over the course of the first 10 years, then overtakes productivity during the remaining 20 years. Conversely, the wage profile of a salesperson differs little from that of his productivity. This near overlap derives from the fact that the activity of a salesperson is *verifiable* (an impartial tribunal can determine an employee's sales volume by, for example, checking his sales records) and in consequence his wage can be largely based on the number of articles he has sold.

Another instructive feature of the shirking model—see relation (6.55)—is that, the hiring wage excepted, an employee's wage ought to increase when the frequency p of supervision declines. Now this frequency is likely an increasing function of the ratio of the number of supervisors checking on the performance of employees to the number of employees. So, all other things being equal, the deferred payment mechanism suggests that firms paying high wages are also those in which the supervisors/employees ratio is lowest. The study of Groshen and Krueger (1990) on hospitals in the United States does in fact come to this conclusion.

4.3 EFFICIENCY WAGE AND INVOLUNTARY UNEMPLOYMENT

The optimal wage profile in the shirking model works like a bonding mechanism: workers accept being paid less than their productivity at the onset of their careers, and higher pay later on. In certain cases the hiring wage can even be negative. This means that workers accept that they are making a deposit with the employer, who will pay them

[7] The peaks at around 55 years of age correspond to the payout of "retirement capital."

back later on. Labor contracts do not generally stipulate this covenant. Hence it is useful to analyze the way the labor market functions when employers cannot manipulate wage profiles for incentive purposes as much as they wish. From this point of view, Shapiro and Stiglitz (1984) have built a celebrated "efficiency wage" model, in which there exists *involuntary* unemployment at labor market equilibrium. This model illustrates how the labor market would function in a limit case in which employers had no choice in the matter of wage profiles, and on that account constitutes a very fragile explanation of unemployment.

4.3.1 The Model of Shapiro and Stiglitz

To suppress the bonding mechanism, Shapiro and Stiglitz (1984) adopt a stationary version of the shirking model in which firms are constrained to pay the same wage at all periods. The incentive mechanism is then based solely on the risk of unemployment.

A Bonding Mechanism

The optimal wage profile described by relations (6.54) and (6.55) is interpretable as an incentive mechanism in which there is a "bond" the agent is obliged to post at the time of hiring which will be gradually paid back to her. It is just as though the agent were to be paid a base wage equal to the outside wage $\bar{w}_t$ at every period $t \geq 0$, but had at the outset deposited a sum equal to C/p with the employer; this sum is gradually reimbursed starting at date $t \geq 1$ in the form of a bonus added to the base wage and amounting in each period to[8] $(C/p)\{[1/\delta(1-q)]-1\}$. Akerlof and Katz (1989) pointed out that when the shirking model is specified in this way, it is not possible to "smooth out" the profile of optimal wages, that is, to narrow the gap between the hiring wage w_0 and the subsequent wages w_t. To grasp this point clearly, let us suppose that the agent receives a wage $w_m > w_0$ over the initial period of the contract. The employer cannot, at certain periods $t \geq 1$, pay a remuneration less than the optimal wage w_t characterized by (6.55), for in the periods that he did so, the employee's incentive constraint (6.46) would no longer be satisfied.

Accordingly, the shirking model displays a bonding mechanism that cannot be substituted for a regularly increasing wage profile. Such a mechanism is not at all realistic in practice, and in any case the payment of a deposit does not exist except in the rarest of cases in the labor market. One of the reasons used to explain this absence is the imperfection of the financial markets (Shapiro and Stiglitz, 1985): workers supposedly suffer from liquidity constraints that prevent them from collecting the sums necessary to put down the deposit. Shapiro and Stiglitz (1984) base their theory of the efficiency wage on the practical impossibility of making this deposit mechanism work and show that it would lead to the emergence of involuntary unemployment.

A Particular Stationary Version of the Shirking Model

Shapiro and Stiglitz (1984) adopt a stationary version of the shirking model in which $\overline{V}_t = \overline{V}$, for all $t \geq 0$. Moreover, this version radically suppresses the bonding

[8] It is easy to verify that the deposit C/p is equal to the present value, discounted at rate $\delta(1-q)$, of the sum of the bonuses.

mechanism by assuming that the principal pays the *same* wage w at every period. With this hypothesis, the agent is given incentive to furnish effort $e > 0$ throughout the duration of the contract if and only if the principal pays him the wage defined by the right-hand equality of relation (6.52), or:

$$w = (1-\delta)\overline{V} + C + \frac{C}{p}\left[\frac{1}{\delta(1-q)} - 1\right] \tag{6.56}$$

In this case, the utility V_t expected by the agent is the same at each period t and can be denoted simply by V. A consequence of the hypothesis of wage stationarity made by Shapiro and Stiglitz is that $V_0 = V$. In consequence, relation (6.51) shows that the agent benefits from a *rent* over the *whole* of the duration of the contract such that $V_0 - \overline{V} = C/p\delta(1-q)$. If we take the view, as Shapiro and Stiglitz do, that $\overline{V}$ designates the expected utility of an unemployed person, it results that accepting a job offer procures a gain V_0 *strictly* superior to the gain $\overline{V}$ of an unemployed person. Unemployment is thus *involuntary* in nature, since anyone looking for work prefers to accept a job at the current wage w (which offers her an expected utility V_0) rather than remain unemployed (which offers an expected utility equal to $\overline{V}$).

Let z be the gains of an unemployed person at every period, and let $s \in [0,1]$ be the (endogenous) probability of returning to work at every period. In a stationary state, the intertemporal utility of an unemployed person $\overline{V}$ satisfies the following equation:

$$\overline{V} = z + \delta\left[sV + (1-s)\overline{V}\right]$$

Since $V = \overline{V} + C/p\delta(1-q)$, an unemployed person's expected utility is expressed as a function of the rate of return to work according to the formula:

$$(1-\delta)\overline{V} = z + \frac{sC}{p(1-q)}$$

If we carry this equality into the expression of the efficiency wage (6.56), we find a relationship between the wage paid to employees and the exit rate from unemployment, which takes the form:

$$w = z + C + \frac{C}{p}\left[\frac{1}{1-q}\left(s + \frac{1}{\delta}\right) - 1\right] \tag{6.57}$$

The exit rate from unemployment depends on the level L of overall employment. Relation (6.57) thus supplies a link between wages and employment that needs to be made explicit. To that end, let N be the (exogenous) size of the labor force; the level of unemployment is then equal to $N - L$. In a stationary state, the flow qL of entries into unemployment equals the flow of exits $s(N-L)$ out of it. Consequently, we have $s = qL/(N-L)$ and in carrying this value into (6.57), we do indeed find a relationship between the wage level and the employment level, written:

$$w = z + C + \frac{C}{p}\left[\frac{1}{1-q}\left(\frac{qL}{N-L} + \frac{1}{\delta}\right) - 1\right] \tag{6.58}$$

This relation, which is often called the *incentive curve* (*IC*), is represented in figure 6.6. It is increasing and possesses a vertical asymptote at point $L = N$. This property signifies that there is never full employment at equilibrium. This is easy to see: in a situation where there is no risk of lasting unemployment, an employee knows that in case of job loss, he will immediately find another one. He then has an interest in shirking, since doing so no longer threatens him with any loss. In this model, the fear of unemployment plays an incentive role only if unemployment lasts a certain length of time, for it is during this period that the agent suffers losses.

Labor Market Equilibrium

To close this model, we must still specify the behavior of firms. Like MacLeod and Malcomson (1998) and Malcomson (1999), we can take the view that the profit linked to outside opportunities, equal to $\bar{\Pi}$ in the stationary state, designates the expected profit of vacant jobs. If y represents the constant exogenous production of a worker, the gain Π expected from a filled job is given by the equality:

$$\Pi = y - w + \delta\left[(1-q)\Pi + q\bar{\Pi}\right] \tag{6.59}$$

Let us also assume, for the sake of simplicity, that information is perfect in the labor market, which thus operates without friction. Let the situation be one in which $s < 1$, an equilibrium in which there are more unemployed persons than vacant jobs. Vacant jobs are then immediately filled by the unemployed and so offer the same prospect of profit as jobs that are filled. With these hypotheses, we will have $\Pi = \bar{\Pi}$. Finally, let us assume that there is no barrier to entry into the labor market. Competition will then cause entrepreneurs to open up vacant jobs as long as the profit expected from such job creation surpasses the cost of installing new equipment. Let C_K be the exogenous, supposedly constant, value of this cost; entries into the market for goods will stop when the expected profit $\bar{\Pi}$ from a vacant job is exactly equal to C_K. At free entry equilibrium, we will thus have $\Pi = \bar{\Pi} = C_K$ and relation (6.59) defining the expected profit from a filled job entails that the equilibrium value w^* of the efficiency wage is given by $w^* = y - (1-\delta)C_K$. Carrying this equality into equation (6.58), which characterizes the efficiency wage, we find ultimately that equilibrium employment L^* is given by:

$$w^* = y - (1-\delta)C_K = z + C + \frac{C}{p}\left[\frac{1}{1-q}\left(\frac{qL^*}{N-L^*} + \frac{1}{\delta}\right) - 1\right]$$

Labor market equilibrium is thus situated at the intersection of the incentive curve (*IC*) and the horizontal line with ordinate $y - (1-\delta)C_K$; it is represented by point E in figure 6.6.[9] This equilibrium is characterized by involuntary unemployment linked to

[9]The existence of this equilibrium assumes, for one thing, that the exit rate from unemployment, equal to $qL^*/(N-L^*)$, is inferior to unity, and for another, that the horizontal line with ordinate w^* intersects the curve (*IC*); this occurs when the following condition is satisfied:

$$y - (1-\delta)C_K > z + C + \frac{C}{p}\left[\frac{1}{\delta(1-q)} - 1\right]$$

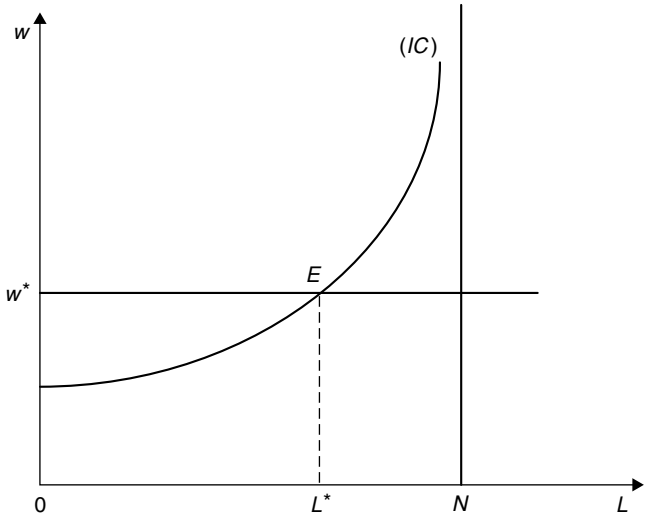


Figure 6.6
The equilibrium of the model of Shapiro and Stiglitz (1984).

a downward rigidity in the real wage. The $N - L^*$ unemployed persons would indeed all agree to work for a wage *less* than w^*, since the situation of an employed participant procures an expected utility superior to that of an unemployed one. But they would then have an interest in shirking, which dissuades employers from offering lower wages.

4.3.2 Final Remarks on Efficiency Wage Theory

The shirking model leads us to an equilibrium with involuntary underemployment, in which employees receive a rent that gives them incentive to supply an adequate level of effort. This model suffers, however, from a major theoretical weakness having to do with the fact that firms could think of other remuneration schemes more sophisticated than the payment of an unvarying wage (Yellen, 1994). These might include, for example, the payment of an award when an employee is found not to be shirking (MacLeod and Malcomson, 1989). More generally, we have shown that the shirking model with no restriction on individuals' strategies leads to a bonding mechanism that offers no rent to employees, so that there is no involuntary unemployment at labor market equilibrium (this objection to the efficiency wage theory is also known as the *bonding critique*; see Carmichael, 1985, 1989, and 1990). The result that there is no rent for the employee when the employer unilaterally decides on the clauses of the contract is not linked to the particular kinds of incentive mechanism that we have considered. Fundamentally, it illustrates a general principle of the theory of incentives, which is that a principal who has at her disposal a sufficiently wide range of strategies can always make the agent's participation constraint binding and thus appropriate the entire surplus flowing from the contractual relationship (see, for example, Kreps, 1990, p. 604).

The existence of a rent for the agent in a model with moral hazard is thus grounded on restrictions—that require explanation—on the strategic options of individuals.

The Financial Market and the Minimum Wage

Shapiro and Stiglitz (1985) made the objection that credit market imperfections rendered the bonding mechanism impracticable. It is not, however, certain that this argument carries the weight it may appear to at first sight, for it is grounded in an excessively strict interpretation of the shirking model. If we were to add certain hypotheses, like the assumption that workers are heterogeneous and that it therefore takes time for each one's aptitudes to reveal themselves, the shirking model would produce a compensation policy resembling a wage profile increasing in normal fashion with seniority. To rescue the efficiency wage theory as a possible explanation of involuntary unemployment, we need to find reasons to explain why firms cannot reasonably offer an increasing wage profile that would bind the participation constraint of workers.

One reason might be the existence of a legal miminum wage w_m exceeding the hiring wage w_0. We have shown that in this case firms pay wage w_m in the initial period, then wage w_t defined by (6.55) in the following periods. Each worker receives a rent equal to $w_m - w_0$ and unemployment becomes involuntary again. This situation is conceivable, but the reason for the rent, and thus involuntary unemployment, is not a necessity inherent in the incentive mechanism, in other words a problem of moral hazard. On the contrary, the reason is the existence of a wage floor making remuneration downwardly rigid. A purely competitive model would have come to the same qualitative conclusions about unemployment. More precisely: moral hazard entails the existence of an abrupt step up in the wage profile that would not have been necessary in a traditional supply and demand model. The equilibria are thus not a priori the same in the two types of model. But it is always a constraint on the downward flexibility of the real wage that enables us to explain the involuntary nature of unemployment.

Rent and Asymmetric Information

Beaudry (1994) and Arvan and Esfahani (1993) have advanced a justification for the existence of rents; it is based on the notion that workers who observe imperfectly the characteristics of the firm that hires them may doubt the credibility of undertakings given about remuneration profiles that will rise, or rewards that will be paid out. Let us suppose, for example, that there are two types of employers. With the "bad" ones, production y is low (independently of the efforts the workers make), and these employers have an interest in systematically discharging their employees after having promised them an increasing wage profile, or rewards. With the "good" employers, production y is high, and they can offer credible contracts. Within this setup, to offer remunerations that pay a rent constitutes a way for the "good" employers to signal their quality. The hypothesis of a double asymmetry of information grounded in moral hazard and adverse selection thus allows us to save the efficiency wage theory as the foundation of a form of involuntary unemployment. Moen and Rosen (2006) build a model of this type where workers are heterogeneous and where firms observe neither their efforts nor their productivities. These models do not, however, explain why firms choose to signal their characteristics by offering high wages, when they might, for example, spend more on advertising their products, which could turn out to cost less than letting rents go to their employees.

Wage Rigidity, Incentive, and Rent

The model of Shapiro and Stiglitz (1984) suggests that unemployed persons ought to offer their services at less than the current wage and that firms ought to refuse these offers. Some recent studies based on surveys of employers do in fact show that they are reluctant to lower wages, even in situations in which there is significant unemployment. According to Blinder and Choi (1990), Bewley (1995), and Campbell and Kamlani (1997), the company executives surveyed believe that wage reductions would be judged "unfair" by employees and would provoke increased turnover and reduced intensity of effort in response. Agell and Lundborg (1995) come to identical conclusions and also find that firms do not want to hire unemployed persons offering to work for less.

Fehr and Falk (1999) have studied the downward rigidity of wages within the framework of an experiment in which two groups—firms and workers—have the opportunity to agree on a wage contract through a mechanism of bilateral bidding. When the participants are forced to sign only *incomplete* contracts, in which the level of effort is not stipulated in advance, the experiment shows that firms refuse to bid wages down. On the other hand, when the actors have the opportunity to sign complete contracts, remunerations become markedly more flexible and approach their competitive values. These surveys and experiments reinforce the view that wage policies are, in the broad sense, driven by the need for incentive, but they give no particular indication as to the existence of *rents*. The downward rigidities highlighted by these empirical studies do not in the least contradict the general properties of self-enforcing contracts *without rent* developed in this section. For example, a deferred payment mechanism offering no rent over the whole course of the wage relationship is just as "rigid" as the unvarying wage in the model of Shapiro and Stiglitz (1984).

5 SOCIAL PREFERENCES

Many studies done in real life and laboratory settings suggest that the well-being of an individual may depend (positively or negatively) on the well-being of others, especially the well-being of those who make up her professional circle of acquaintances. All the models of incentivizing contract we have examined to this point leave out this aspect, on the assumption that the well-being of an individual depends solely on her income. If we do assume that an individual's preferences can also depend on others' well-being, then in very general terms we may describe her as being endowed with "social preferences." These may take different forms: she may be motivated by concern for fairness (i.e., equity) or reciprocity for example. It is also possible that some people derive satisfaction from aiding others, including their colleagues at work or their boss, without expecting financial reward. Are the lessons we have learned from the theory of incentivizing contracts set forth above modified substantially when we introduce the existence of such social preferences? That is the question we now address.

5.1 Gift Exchange, Reciprocity

Numerous empirical studies suggest that social norms like fairness or morality influence the formation of wages. For example, Bewley (1995) states, on the basis of a survey of

300 people involved in formulating wage policies (managers of companies, trade unionists, consultants, etc.): "My findings support none of the existing economic theories of wage rigidity, except those that emphasize the impact of pay cuts on morale" (Bewley, 1999, p. 460). The surveys of company managers by Blinder and Choi (1990) and Campbell and Kamlani (1997) find that they give the need for fairness top priority. This need can be taken into account in our basic model by making the preferences of workers depend on the average wage that prevails in the economy.

5.1.1 Social Norms and Wage Formation

The idea that individuals are particularly attached to a feeling of *equity* or *fairness* is reinforced by work in psychology (Argyle, 1991) and organization theory (Lawler, 1994). When applied to labor relations within a firm, this idea signifies that an employee expects that his effort will be rewarded by remuneration regarded as fair. On the other side, the employer takes for granted that in exchange for the wage paid, his employee will supply an effort regarded as fair. This concept is linked to the work of anthropologists in the tradition of Mauss (1923): it amounts to comparing a variety of exchange relationships, unfolding over a sufficiently extended period, to a sequence of *gifts* and *counter-gifts*. Akerlof's (1982) article brought it within the purview of economists. According to Akerlof, the employee's gift consists of exceeding prevailing work standards, in exchange for which the employer pays him a wage exceeding the so-called reference wage.

According to Akerlof (1982), the importance of fairness is enhanced by the fact, widely documented, that numerous employees do exceed prevailing work standards in their firms, yet at the same time those whose performance doesn't meet those standards are not systematically fired. Observations of this kind cannot be understood using the traditional neoclassical model. For Akerlof, the explanation has to be sought, in part, in the domain of sociology: an employee has a tendency to develop feelings for her firm and for the smaller group consisting of her colleagues. In these circumstances, an employee derives satisfaction from making a gift of extra effort to the firm, a satisfaction analogous to that which she would feel when offering an unusually valuable present to a friend or relative. In this case, the employer clearly has no interest in raising work standards. Likewise, if an employee takes satisfaction from the well-being of the coworkers in her group, the firm does not necessarily have an interest in getting rid of those who are less productive, or even in checking on them more closely than on the rest. So in their celebrated study of the behavior of American soldiers during World War II, Stouffer et al. (1949) observed that during training exercises, soldiers with greater physical capacities spontaneously helped out the weaker ones, without looking for any personal advantage. For those soldiers, it was probably a case of increasing their own satisfaction by raising that of the group as a whole.

5.1.2 An Illustrative Model

The consequences of fairness for wage formation and employment can be illustrated by assuming, following Akerlof (1982), that the preferences of workers are influenced by social norms. Let us consider a labor market with a continuum of identical workers, the measure of which is normalized to 1. Let ω be the average wage prevailing in the

economy; the preferences of a worker are represented by a utility function $u(R, e, \omega) = R[1 + \beta(e/\omega)] - (e^2/2)$, with $\beta \geq 0$. In this expression, e represents the level of effort that will be chosen by a worker if he is hired (the value of e is 0 for all those who do not participate in the labor market). The variable R designates income, equal to wage w if the worker is employed, and equal to the opportunity cost of labor, denoted θ, otherwise. Parameter θ is characterized by a cumulative distribution function $G(.)$ defined on the set of nonnegative real numbers. When β is strictly positive, this specification of preferences expresses the hypothesis that an individual takes more satisfaction from his effort, the higher his relative wage, w/ω, is. It fits well with the notion of fairness just discussed. Finally, we assume that individual production is simply equal to the level of effort e. The free entry condition entails zero profit, and thus a wage w is equal to individual production.

If preferences are unaffected by considerations of fairness ($\beta = 0$), and if the labor market is perfectly competitive, the level of effort maximizes $e - (e^2/2)$, which entails $e = 1$. The utility of employed workers is equal to $1/2$. All individuals with a characteristic θ less than $1/2$ decide to work, and total employment amounts to $G(1/2)$.

If we now assume that $\beta > 0$, we are in a position to show that considerations of fairness can lead employers to offer relatively high wages in order to take advantage of the process of "gift exchange." Under this hypothesis, each worker takes ω as given and chooses his level of effort by solving the following problem:

$$\max_e e[1 + \beta(e/\omega)] - (e^2/2)$$

Optimal effort $e(\omega)$ is thus equal to $[1 - 2(\beta/\omega)]^{-1}$. Since each worker chooses his level of effort as a function of the average wage, the equilibrium is necessarily symmetric. We thus have $e = \omega$ at equilibrium, and so effort and wage are characterized by the equalities:

$$e = w = 1 + 2\beta$$

This relation shows that social norms influence productivity and effort at equilibrium. Workers are given an incentive to make an extra effort, and they receive high wages in exchange. Employed workers obtain a utility equal to $\beta + (1/2)$, and employment rises to level $G[\beta + (1/2)]$, which is superior to that obtained in the absence of fairness considerations. Hence employment does depend on social norms too, and increases with the importance workers place on equity. This result is not, however, general; Akerlof (1982) and Akerlof and Yellen (1990) present examples in which fairness raises unemployment.

Notwithstanding that, this model does allow us to illustrate a very general result: the inefficiency of competitive equilibrium in the presence of social norms. For a given value of β, the optimal allocation is calculated by maximizing the sum of the utilities of workers present in the market in a symmetrical situation in which each worker supplies the same level of effort, or $e = \omega$. That amounts to maximizing the utility of every worker with $e = \omega$. We then obtain $e = 1 + \beta$, which corresponds to a level of effort increasing with the degree of consideration for fairness but inferior to that obtained at competitive equilibrium. The social norm is like an externality that compromises the efficiency of the competitive mechanism.

5.1.3 Empirical Confirmation

The approach proposed by Akerlof makes the representation of preferences richer by integrating an explicitly social dimension. For more than 20 years, numerous laboratory and field experiments have confirmed the importance of this dimension in exchange relations (see the comprehensive panorama of Charness and Kuhn, 2011, on laboratory experiments; and that of List and Rasul, 2011, on field experiments).

Laboratory Experiments

The clearest confirmation of the importance of fairness (or reciprocity) in exchanges was supplied by the results of the *ultimatum game*, with the first articles bearing on the topic appearing in the early 1980s. The simplest version of the ultimatum game is an experimental setup in which one person (the offerer, or proposer) is given a sum of money and chooses how much she is prepared to offer to another participant (the responder). If the responder accepts the offer, it is implemented. If he refuses the offer, both agents get nothing in the end. If a proposer endowed with x$ had nothing but her own interest in mind, she ought to make a proposal of 1$ to the other participant, and he ought to accept it. This experiment has been repeated countless times, and the outcome that prevails is very different. In fact, it turns out on average that the proposer makes an offer close to an equal division of the sum she controls and that the responder rejects the offers he judges too unbalanced. The ultimatum game thus tends to prove that in a bilateral exchange, individuals accord primary importance to fairness or reciprocity.

The gift-exchange game is a more sophisticated experimental setup than the ultimatum game, in that it tries to simulate a market situation. The conclusions that emerge are nevertheless in line with those of the ultimatum game. As Charness and Kuhn (2011) put it: "Probably no experimental game in the area of labor economics has had as much impact as the gift-exchange game, which tests the notion . . . that there is a positive relationship between wages and effort" (p. 281). The paper of reference on the gift-exchange game remains the laboratory experiment conducted by Fehr et al. (1993). The setup consists of mimicking a labor market, a priori competitive, with the subjects (students in this instance) separated into "firms" and "workers." There are more workers than firms, and a firm must be matched with just one worker. Basically, workers make offers regarding the effort they are prepared to expend (at a cost to themselves) and the firms respond with wage offers. As it is the worker who makes the first move, with a proposal of effort he will have to fulfill, he ought "logically" to anticipate that, whatever level of effort he declares, when the firm makes its move it will have no interest in offering him a wage higher than the minimum stipulated by the game, so that his best move will be to propose no more than the minimum of effort stipulated by the game. And the only "logical" outcome of the game would be the contract (e_0, w_0), where e_0 and w_0 designate respectively the minimum effort and the minimum wage stipulated by the game. But that is not the outcome that this experiment yields. In fact, the wage proposals are strongly and positively correlated to the effort proposals: the effort proposed is on average around four times higher than the minimum effort, and the wage proposed corresponds on average to about twice the minimum wage. Fehr et al. (1998) have replicated this laboratory experiment with Austrian soldiers and in a "noncompetitive" environment where there are as many firms as workers. The results are similar to Fehr et al. (1993).

A matchup between one firm and one worker is a highly artificial scenario. We might suppose that the gift-exchange relation would vanish or dwindle sharply once

the firm was employing a plurality of workers, if only because of the possibility of free riding. Maximiano et al. (2007) created an experimental setup that made it possible to compare a bilateral gift-exchange game like that of Fehr et al. (1993) with a multilateral exchange game where each firm employs four workers. In the latter case, the firm can gain a lot more than each worker it employs and also gain a lot more than the firm employing just one worker. Yet the results show that the efforts proposed are scarcely any lower than in the setting of bilateral gift exchange. Maximiano et al. suggest that in choosing their level of effort, agents within a firm are guided by the *intentions* of reciprocity they attribute to other agents.

Other experiments have tried to determine whether individuals react more strongly to intentions on the part of others that they think will increase their own well-being (positive reciprocity) than they do to intentions that they think will diminish their own well-being (negative reciprocity). To that end, setups are arranged in which the responder receives offers that may come from a proposer or may be generated randomly by a machine. Experimenters observe that responders do not react much differently to proposals that improve their situation, whether they are made by another person or the machine. Conversely, responders react very strongly (and negatively) to proposals that worsen their situation if the proposer is human but much less strongly if the proposal is generated by the machine. These experiments would seem to indicate that individuals are more sensitive to negative reciprocity than to positive reciprocity (see Offerman, 2002; Charness, 2004).

In the Field

Many laboratory experiments reveal the importance people attach to reciprocity in exchanges, but can we be certain that they give it the same importance in "real life"? Gneezy and List (2006) have supplied a partial answer to this question with the help of two field experiments. In the first, students were recruited to work on converting the catalogue of a university library to digital form: for six hours they had to enter information about books into a database. The recruiting was done through posters describing the task to be performed and stating that the wage would be $12 per hour, with no mention of the fact that this was an experiment. In the end the sample population came to 19 students. Ten of them were chosen at random to make up the control group, which was paid the advertised wage. The other nine, who made up the treatment group, were told just before their shift began that they would be paid $20 an hour instead of $12. In line with the gift-exchange hypothesis, the members of the treated group supplied markedly more effort than those in the control group (around 25% more) . . . during the first 90 minutes of their 6-hour shift! For the remaining 4$^1/_2$ hours, the effort levels were indistinguishable across the two groups.

In their second natural experiment, Gneezy and List invited students to take part in a door-to-door fundraising drive to support a research center at their university. Similarly to the library task, it was announced that the work would be done over a weekend and paid at a rate of $10 per hour, again without stating that this was an experiment. The overall sample size this time was 23 participants, 10 of whom were randomly assigned to the control group that was paid the specified sum. The 13 members of the treatment group were told, after being briefed on the task they were to perform, that they would be paid $20 per hour instead of the $10 that had been advertised. The results were very close to the results of the university library experiment. The members of the test group

collected much more money than those in the control group during the first three hours of their shift (almost 70% more), but after that there was no longer a significant gap.

The experiments of Gneezy and List thus raise doubts about how long the gift-exchange mechanism may persist in the real world of work. But we should note that their experiments only test the effects of *positive* reciprocity (in both cases, the members of the treatment group were surprised by *good* news). Their results are in line with the results of the laboratory experiments cited above, which showed that individuals react in relatively muted fashion to situations of positive reciprocity but in relatively vivid fashion to situations of negative reciprocity. This pattern is confirmed by the field experiment conducted by Kube et al. (2011), in which a university library sought student helpers to enter the titles of books into a catalogue. The advertised wage for this task was $15 per hour, which is what the members of the control group were paid. Kube et al. constructed two treatment groups. In the first, the participants were informed immediately before work began that in fact they would only be paid $10 per hour, while in the second they were told that in fact they were going to get $20 per hour. Kube et al. observed a negative gap of more than 20% between the average output of the control group and that of the underpaid test group. In contrast, they detected practically no difference in productivity between the control group and the overpaid test group. This experiment confirms that employees are measurably more sensitive to negative reciprocity than to positive reciprocity.

It also tends to confirm the results obtained by Bewley (1995) on the basis of a survey of 300 people involved in formulating wage policies (managers of companies, trade unionists, consultants, etc.). He reports that managers are very reluctant to lower wages during a recession, since they fear that employees may react by immediately reducing their level of effort and then persist in doing so after growth returns. Hence managers are well aware of the potentially worrying effects of negative reciprocity. A good example is supplied by Lee and Rupp (2007), who studied the impact of significant and permanent reductions in their salaries on the performance of pilots in seven U.S. airlines. The restructuring of the American air travel sector forced many companies to reduce their wage costs, by negotiation or fiat, at the start of the 2000s. The amounts at stake were obviously much larger than they were in the laboratory experiments—a senior pilot in a major company typically earned more than $200,000 per annum and the cuts ranged in size from 8% to 33%. Additionally, these subjects were facing a drop in their income over the long term. Lee and Rupp compare the performance of pilots, proxied by the percentage of on-time flights, just before and just after the announcement of the salary cuts, during a 40-day window (this procedure is called an "event study"). The period is sufficiently short to avoid picking up tendencies to arrive late that might have been present in some companies independently of the event studied. Using almost 1.4 million observations (daily route-level measures, in airline jargon), and controlling for weather conditions and local circumstances at each airport, the authors observe a significant drop in performance, with longer and more frequent delays in arrival, in companies that had reduced the salaries of their pilots: the share of arrivals delayed by 15 minutes or more, which happens on about 19% of flights on average, rose by three points. But this increase in lateness lasted no more than a week. Lee and Rupp also found that pilots did not reduce their output of effort in companies that cut back on salaries when facing bankruptcy, perhaps because they feared even greater loss if the company were to close its doors for good, or perhaps because they felt it was the fair

thing to do, given the company's predicament. These results thus corroborate the results obtained in laboratory experiments.

5.2 Intrinsic Motivation and Reputation

The fact that financial incentives can have counterproductive effects was pointed out by psychologists some time ago (Kruglanski, 1978; Deci et al., 1999). They have also demonstrated that social norms and/or the way others see us and the way we see ourselves (our "reputation") are powerful vectors of our attitudes in society. Our actions may be guided in part by "intrinsic" motives that may conflict directly with those "extrinsic" motivating factors par excellence, financial incentives (see the survey of Rebitzer and Taylor, 2011, on this potential conflict). In what follows, we will examine these problems with the help of a model inspired by Bénabou and Tirole (2006, 2012), and give some empirical illustrations of the clash between intrinsic and extrinsic motives.

5.2.1 A Model Where People Care About Reputation

The Participation Condition

Let there be a continuum of agents, each of whom must choose an action a that may take the values of 1 or 0. Action a has a prosocial dimension that may vary with the context. (Prosocial behavior is defined as voluntary behavior meant to benefit others.) Examples might include donating blood or not, voting on election day or not, separating the recyclables from the rest of the garbage or not, applying oneself sedulously on the job or shirking, being helpful to colleagues at work or not, and so on. When the agent opts to engage in prosocial behavior ($a = 1$), he is said to "participate," and he receives a financial reward w, taken to be exogenous, that procures him utility w. The model can also be extended to cover contributions to the public welfare that are not remunerated directly (to refrain from polluting, for example, also corresponds to $a = 1$), but that would attract a fine $-w$ when $a = 0$ (when one was caught polluting).

Participation does however bear a cost in terms of the time (or effort) invested that reduces the utility by an amount denoted c. And for that matter we will assume that it also procures an intrinsic satisfaction that may be different for every agent. If some disagreeability is attached to participating, that can be integrated into parameter c, which in turn allows us to assume that intrinsic satisfaction, denoted v, is a positive magnitude that increases utility by an amount v. More precisely, we will assume that v is a random variable defined over $[0, v_M]$ of which the cdf (cumulative density function) and the pdf (probability density function) are denoted respectively $G(.)$ and $g = G'$. The value of v may for example represent the intensity of the happiness one feels at being altruistic or putting one's professional skills to use.

Each agent evidently knows her own type of v, but that is something other agents cannot observe. Participation however is assumed to be observable, making it possible for others to form a notion of what sort of agent she is by observing her participation. Formally, this notion is represented by the quantity $\mathbb{E}(v \mid a)$, which is the conditional expectation of the random variable v formed after having observed action a. The crucial hypothesis in the model of Bénabou and Tirole is that the satisfaction of an agent depends on what others think of her. Let us assume that everyone believes that agents who participate are of a type v superior to a threshold v^*. We then have

$\mathbb{E}(v \mid a = 1) = \mathbb{E}(v \mid v \geq v^*)$ and the positive quantity $[\mathbb{E}(v \mid v \geq v^*) - v^*]$ is a measure of the social gratification (honor, if you like) that participation procures: the more you are regarded by others as someone who exceeds threshold v^*, the happier you are. The utility of an agent of type v who participates is ultimately written:

$$U(1) = w + v - c + \mu\,[\mathbb{E}(v \mid v \geq v^*) - v^*]$$

In this expression, parameter $\mu \geq 0$ designates the weight an agent allots to her "reputation." The sign of μ reflects the idea that people would like to appear prosocial. To enjoy good repute may yield disinterested satisfaction, with no other payoff than one's own pleasing self-image. Yet self-interest may perfectly well play a part in the satisfaction that a good reputation procures: for example an employee in a firm may have an interest in showing that he is assiduous on the job or that he is willing to expend effort with promotion in mind. Note as well that v and μ are not necessarily linked. An agent may take no pleasure in his work ($v = 0$) but accord great importance to his reputation (large μ) or, conversely, take great pleasure in his work (large v) and care not a whit about his reputation ($\mu = 0$).

Besides, we have $\mathbb{E}(v \mid a = 0) = \mathbb{E}(v \mid v \leq v^*)$. The negative quantity $[\mathbb{E}(v \mid v \leq v^*) - v^*]$ is then a measure of the social stigma (dishonor, if you like) that nonparticipation attracts; the more you are regarded by others as someone who falls beneath threshold v^*, the unhappier you are. The utility of an agent of type v who does not participate is simply written:

$$U(0) = \mu\,[\mathbb{E}(v \mid v \leq v^*) - v^*]$$

Let us define the function:

$$\Delta(x) = \mathbb{E}(v \mid v \geq x) - \mathbb{E}(v \mid v \leq x) = \frac{1}{1 - G(x)} \int_x^{v_M} v\mathrm{d}G(v) - \frac{1}{G(x)} \int_0^x v\mathrm{d}G(v) \qquad \textbf{(6.60)}$$

For given threshold x, $\Delta(x)$ represents the difference between the gratification that participation procures and the (absolute value of) the stigma that abstention attracts. This is a positive quantity that may be likened to a measure of net reputational payoff (the payoff to participating compared to abstaining). Note that nothing can be stated a priori about the sign of the derivative of $\Delta(x)$. All outcomes are possible. $\Delta'(x) > 0$ if a marginal hike in the threshold does more to boost the reputational payoff to participation than it does to undercut the reputational payoff in case of abstention.

With the help of function $\Delta(x)$, the participation condition $U(1) \geq U(0)$ takes the form:

$$w + v - c + \mu\Delta(v^*) \geq 0 \qquad \textbf{(6.61)}$$

The equilibrium thresholds v^* are then defined by the equation:

$$w + v^* + \mu\Delta(v^*) = c \qquad \textbf{(6.62)}$$

The interpretation of this equation is quite simple. An agent of type v^* is indifferent between participating and abstaining, hence the cost c that participation incurs must be equal to the gain that it brings. Now this gain is composed of the extrinsic payoff (w), the intrinsic payoff (v^*), and the net reputational payoff ($\Delta(v^*)$).

The Crowding-out Effect and the Accelerator Effect of Reputation

In addition, for equilibrium to exist, agents have to be able to coordinate around the same setting of threshold v^*. In this regard, it will be helpful to describe a mechanism of convergence towards equilibrium that ensures its stability. Let us suppose that at date t an agent thinks that the threshold is equal to v_t^*. In accordance with condition (6.61), she knows that only persons for whom $v \geq c - w - \mu\Delta(v_t^*)$ are going to participate, hence she revises the setting of the threshold following the formula:

$$v_{t+1}^* = \phi(v_t^*) = c - w - \mu\Delta(v_t^*) \tag{6.63}$$

The equilibrium thresholds must verify equation $v^* = \phi(v^*)$ and (local) stability in these points dictates that the absolute value of the slope of function ϕ must be smaller than 1 ($\mathrm{d}v_{t+1}^*/\mathrm{d}v_t^*$ must be bounded by -1 and 1). The local stability of each equilibrium thus requires conditions:

$$1 - \mu\Delta'(v^*) > 0 \quad \text{if } \Delta'(v^*) > 0; \quad 1 + \mu\Delta'(v^*) > 0 \quad \text{if } \Delta'(v^*) < 0 \tag{6.64}$$

The proportion of agents who participate is equal to $1 - G(v^*)$. Since v^* depends on w, we may say by analogy that this proportion represents "labor supply" and will denote it $L(w)$. We thus have $L(w) = 1 - G[v^*(w)]$. The main question to which we now turn is to determine how labor supply varies as a function of the wage.

Deriving equation (6.62) with respect to w, we get:

$$\frac{\mathrm{d}v^*}{\mathrm{d}w} = -\frac{1}{1 + \mu\Delta'(v^*)} \quad \text{and so } \frac{\mathrm{d}L(w)}{\mathrm{d}w} = -g(v^*)\frac{\mathrm{d}v^*}{\mathrm{d}w} \tag{6.65}$$

Equation (6.64) entails that v^* always diminishes with the wage, which signifies that participation necessarily increases with the wage. Still, the degree of impact exerted by the wage depends on interactions that occur among agents.

If $\Delta' > 0$, the net reputational payoff diminishes when the rate of participation increases (participation, equal to $1 - G(v^*)$, decreases with v^*). In this configuration, decisions about participation exhibit *strategic substitutabilities*: an increase in participation by others reduces the reputational payoff to everyone, which is a disincentive for agents to participate. Relation (6.65) then shows that the derivative $\frac{\mathrm{d}v^*}{\mathrm{d}w}$—which Bénabou and Tirole call the "social multiplier"—is positive but smaller than 1. The motivational impact of reputation has the effect of curtailing participation with respect to what it would have been if reputation had no motivational impact. This crowding-out effect, then, occurs when $\Delta'(v^*) > 0$, in other words, when honor dominates stigma (in the eyes of the pivotal agent for whom $v = v^*$). For Bénabou and Tirole, such situations would correspond to cases in which participation takes on "heroic" dimensions. For example, jumping into freezing water to rescue a drowning person ($a = 1$) augments one's prestige considerably. But if you don't ($a = 0$), the stigma is weak, for most people

understand that you (and perhaps they in the same situation) are unwilling to risk your own life. When participation is seen as heroic, material incentives have limited effect. For example, we may reasonably suppose that a (reasonable) hike in the wage paid to paratroopers would have little influence on the number of young men aspiring to join the airborne regiments.

When $\Delta' < 0$, the net reputational gain increases when the rate of participation increases. In this configuration, participation decisions exhibit *strategic complementarities*: increased participation by others increases the reputational payoff to everyone, which incentivizes others to participate. Relations (6.64) and (6.65) reveal that the social multiplier is then greater than 1. Strategic complementarity comes about when stigma dominates honor ($\Delta' < 0$ for the pivotal agent). This situation signifies that participation is experienced as a *social norm*: in terms of reputation, the distress that results from abstaining is more powerful than the payoff to participating. What it comes down to is that participation is seen as normal and refusal to participate is seen as abnormal. In this situation, a small variation in material incentives may have a very big impact on participation (the social multiplier is greater than 1). The effects of a rise in extrinsic incentives (which push agents to take part) are amplified by the reputational mechanism that stigmatizes even more those agents who fail to respect the social norm and so pushes them to participate. For example, an increase in the tax on cars that pollute pushes (extrinsically) a certain number of persons to buy "clean" cars. Those who still drive polluting vehicles grow fewer and so more stigmatized than before (not to pollute is a social norm), which incentivizes them even more to buy clean cars themselves.

Endogenous Social Norms

One of the merits of the model of Bénabou and Tirole is to explain the determinants of strategic substitutability or complementarity, to the extent that they are *endogenous*. The fact that $\Delta'(v^*) > 0$ or that $\Delta'(v^*) < 0$ is a result of the characteristics of agents, particularly the properties of the distribution $G(.)$ of intrinsic motivators. How intrinsic motivators are allocated among agents determines the emergence of strategic complementarity or strategic substitutability. Hence, as the appendix to this chapter shows, a density function decreasing over its whole support ($g' < 0$, which is the case for example with an exponential distribution e^{-v}) entails strategic substitutability ($\Delta' > 0$). This result is quite intuitive: when g' is negative, the share of persons with high values of v is small. This is a world with few heroes, so a high value is placed on heroic attitudes and those who abstain from heroism feel little distress. We have seen above that these are the characteristics of a situation of strategic substitutability. We may also note that $\Delta' > 0$ entails a unique equilibrium—characterized by $v^* = \phi(v^*)$—since function ϕ defined by relation (6.63) is here decreasing.

Conversely, an always increasing density function $g(.)$ entails strategic complementarity for stable equilibria. But when $\Delta' > 0$, function ϕ is increasing, and multiple equilibria become possible, some stable and others not. Such a configuration is represented in figure 6.7 where there are three possible equilibria: equilibrium E_1 which is unstable, equilibrium E_2 which is stable, and a third equilibrium situated at the origin where all agents participate ($v^* = 0$), which is likewise stable. It is interesting to observe that the formation of beliefs can lead to two stable equilibria, one with very strong participation (point 0) and the other with weak participation (point E_2). Take the example

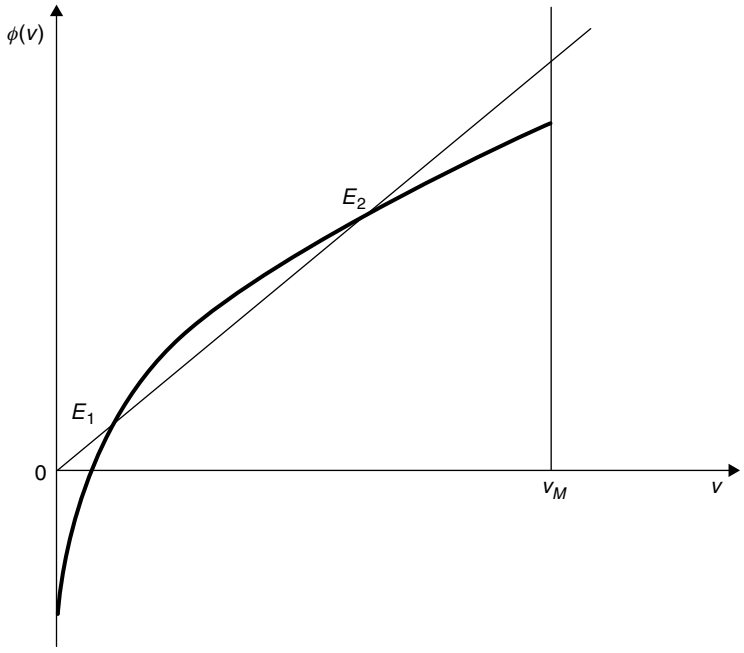


Figure 6.7
Multiple equilibria in the model of Bénabou and Tirole (2006).

where the punishment for polluting is a small fine w. If everyone thinks that most people are civic minded and that the amount of the fine therefore matters little, nobody pollutes. Conversely, if everyone thinks that most people are not civic minded and that the amount of the fine is not dissuasive, a great many people continue to pollute. The factual examples that follow illustrate well this twofold possibility.

A more general case is the one in which the distribution $g(.)$ is unimodal. Under this hypothesis, Bénabou and Tirole show that function Δ is quasi convex and that function ϕ is thus quasi concave. The situation then resembles that described in figure 6.7, except for the fact that the absolute value of the slope at point E_2 might be greater than 1, in which case E_2 would also be an unstable equilibrium.

5.2.2 Empirical Illustrations

The study most often cited to illustrate the point that extrinsic incentives might crowd out prosocial attitudes is that of Gneezy and Rustichini (2000a). This was a field experiment bearing on day-care centers in the city of Haifa in Israel. Prior to the experiment, parents were supposed to come to pick up their children at 4 p.m., but many failed to meet that obligation, forcing the manager and the staff of the day-care center to stay longer than they were supposed to. The natural way to counteract this tendency seemed to be to impose a fine on those who turned up late. To test this idea, Gneezy and Rustichini conducted an experiment on 10 day-care centers over a period of 20 weeks. During the first four weeks, all they did was count the number of parents who arrived late. Then, at the start of week 5, a very modest fine of 10 NIS (New Israeli Shekels) per

child was imposed in 6 day-care centers (the test group) on all parents who were more than 10 minutes late. The four other day-care centers, at which no fine was imposed for lateness, served as the control group. Within the test group Gneezy and Rustichini observed a regular *increase* in the average number of parents arriving late after the fine was put in place. Then, after about three weeks, this number stabilized at a level higher than the one that obtained before there was any fine. The fine was canceled at the start of week 7. During the four weeks that followed the cancellation, the number of parents arriving late did not budge, thus remaining fixed at a level higher than the level during the first four weeks when there had likewise been no fine.

One possible interpretation of this result (Rebitzer and Taylor, 2011) is that before the introduction of a fine, the attitude of the parents was guided in part by concern for their reputation, that is, by the image of responsibility they wished other parents and the employees of their day-care center to have of them. But once a small fine was introduced, arriving on time no longer sent a prosocial signal; now it might be seen as a signal of excessive concern to save a few shekels. The upshot of this experiment would probably have been quite different if the fine had been set substantially higher. A lot more parents would probably have started arriving on time and would not necessarily have been seen as signaling their miserly nature for doing so. The other parents and the staff of the day-care center would probably have judged it perfectly normal to come and get one's children on time if the cost of being late was high. Hence we may view this experiment as an illustration of the crowding out of prosocial attitudes by material incentives, in conformity with the model developed above.

Gneezy and Rustichini (2000b) carried out another laboratory experiment on a group of 160 students at the University of Haifa and obtained results in line with those obtained in the day-care center experiment. In their setup, the students were divided into four different groups, but each student, whatever group she belonged to, received a fixed emolument of 60 NIS for answering 50 questions taken from an IQ test. On top of this basic remuneration, bonuses were available to the students in certain groups. In group 1 (the control group) there was no bonus and students were simply requested to answer as many questions correctly as they could. In group 2, they were promised an extra 0.1 NIS—a derisory sum—for each correct answer. In groups 3 and 4, they received an extra 1 NIS and 3 NIS respectively for each correct answer. On average there were 28 correct answers in group 1 and only 23 in group 2. But there were around 34 correct answers on average in groups 3 and 4.

In this experiment, the sum received by each participant was kept confidential, which meant that the attitude of any student could not be attributed to a wish to display a prosocial image. It is more likely that "self-esteem" accounts for the demotivation of group 2. We may suppose that on average the participants in group 2 judged it degrading or humiliating to have to make an extra effort for such a derisory bonus—an interpretation compatible, once again, with the model developed above.

The conclusion to be drawn from these experiments and the other research cited in this chapter is not that financial incentives do not count. On the contrary, they certainly do play a part, as long as the labor economist takes into account their possible interactions with other sources of motivation, in particular motivations of a prosocial kind that obey a logic different from that of financial incentives. A repertory of possible reasons why financial incentives might fail to incentivize, and ways to try to ensure that they succeed, may be found in Gneezy et al. (2011).

6 SUMMARY AND CONCLUSION

- Labor contracts are used to deal with problems of risk-sharing and incentive. The properties of wage contracts depend to a large extent on whether or not it is possible to take the observation of the results of a wage earner's activity into account. All observations that could be objectively assessed by an impartial tribunal fall into the category of *verifiable* clauses. The labor relationship may be governed by implicit agreements that address the problem of nonverifiability in other clauses. Such agreements occur in the setting of long-term relationships, the existence of which depends only on the mutual interest the partners have in them. We then say that the agreement is *self-enforcing*.
- The demand for insurance allows us to explain certain empirical characteristics of the movement of wages: in particular, the fact that they are procyclical, fluctuating less than productivity, and the fact that they are not correlated with the current rate of unemployment.
- The traditional agency model with hidden action analyzes the remuneration rule that a risk-neutral principal offers to a risk-averse agent, when that agent's results are *verifiable*. The principal faces a problem of *moral hazard*, since he does not know with certainty what actions the agent took in order to achieve her observed results. The optimal remuneration rule exhibits a compromise between the demand for insurance and the need for incentive. It most often prescribes a remuneration that depends on performance. The optimal rule must take account of all the verifiable observations correlated with the effort of the agent.
- Multitasking, only part of which is verifiable, constitutes a source of inefficiency that impels firms to adopt implicit contracts and/or overall indicators of performance. Another source of inefficiency is *rent-seeking*. Its cause is the comparative advantage that agents may derive from concentrating part of their efforts on actions that will impress the supervisors who are charged with informing the principal about observed performances.
- The internal market in a firm, and more generally systems of hierarchical promotion, can be analyzed as *tournaments* in which the rules of promotion and the wages that go along with each promotion are specified in advance. A tournament offers the advantage of making the clauses of a contract explicit. The tournament model also suggests that hierarchical levels ought in large measure to explain wage variation. It suggests further that the remuneration that comes with a grade in the hierarchy rises with the number of individuals who aspire to be promoted to that grade. These predictions match empirical results well. The rule of promotion by seniority is partially explained by the fact that it makes it possible to avoid rent-seeking activity.
- The deferred payment mechanism entails a positive linkage between seniority and wages.
- The shirking model describes a long-term relationship between a principal and an agent, in which the agent's effort *and* results are *unverifiable*. In this context, the optimal remuneration rule is a series of wage settings increasing with seniority but offering the agent no rent over the whole duration of the contract.

Empirical studies confirm the existence of an increasing relationship between seniority and wages, as well as the influence of incentive mechanisms in this area. Moral hazard is a source of inefficiency, however, for employers cannot credibly enter into long-term engagements with agents who may produce insufficient but positive surpluses. The exact wage profile depends on the combined effects of the acquisition of general human capital by the agent and the principal's wish to obtain an adequate level of effort.

- Social preferences play a major role in exchange relations. Most individuals attach importance to fairness (or equity) and to reciprocity and are sensitive to social norms as well as to the image they have of themselves and wish to project to others. This does not mean that financial incentives have no point. Empirical research shows that they certainly play a major role, too, but that it is important to take into account their possible interactions with prosocial motives.

7 RELATED TOPICS IN THE BOOK

- Chapter 1, section 2.1: The choice between consumption and leisure
- Chapter 3, section 1: The competitive equilibrium
- Chapter 3, section 2: Compensating wage differential
- Chapter 4, section 3: Education as a signaling device
- Chapter 5, section 4.2: The equilibrium search model
- Chapter 7, section 3: Collective bargaining
- Chapter 8, section 2: Theories of discrimination
- Chapter 13, section 1: Unemployment insurance

8 FURTHER READING

Bloom, N., & Van Reenen, J. (2011). Human resource management and productivity. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 4B, pp. 1697–1767). Amsterdam: Elsevier.

Bolton, P., & Dewatripont, M. (2005). *Contract theory*. Cambridge, MA: MIT Press.

Gibbons, R., & Waldman, M. (1999b). Careers in organizations: Theory and evidence. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 3, chap. 36, pp. 2373–2437). Amsterdam: Elsevier Science.

Lazear, E. (2011). *Inside the firm: Contributions to personnel economics*. New York, NY: Oxford University Press.

Oyer, P., & Schaefer, S. (2011). Personnel economics: Hiring and incentives. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 4B, chap. 20, pp. 1769–1813). Amsterdam: Elsevier.

Prendergast, C. (1999). The provision of incentives in firms. *Journal of Economic Literature*, *37*, 7–63.

Rebitzer, J., & Taylor, L. (2011). Extrinsic rewards and intrinsic motives: Standard and behavioral approaches to agency and labor markets. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 4A, chap. 8, pp. 702–772). Amsterdam: Elsevier.

Rosen, S. (1985). Implicit contracts: A survey. *Journal of Economic Literature*, *23*, 1144–1175.

Salanié, B. (1997). *The economics of contracts: A primer*. Cambridge, MA: MIT Press.

9 APPENDIX: THE PROPERTIES OF THE NET REPUTATIONAL PAYOFF FUNCTION

The aim of this appendix is to analyze the properties of function $\Delta(x)$ defined by equation (6.60). Function $\Delta(x)$ can be written as:

$$\Delta(x) = \chi(x) - \psi(x)$$

with

$$\chi(x) = \frac{1}{1-G(x)}\int_x^{v_M} v\mathrm{d}G(v) \quad \text{and} \quad \psi(x) = \frac{1}{G(x)}\int_0^x v\mathrm{d}G(v)$$

We wish to show that $g'(x) < 0$ for all $x \in [0, v_M]$ implies $\Delta'(x) > 0$ and that $g'(x) > 0$ for all $x \in [0, v_M]$ implies $\Delta'(x) < 0$. Let us integrate by parts the integral that appears in the expression of $\psi(x)$. We get:

$$\int_0^x v\mathrm{d}G(v) = [vG(v)]_0^x - \int_0^x G(v)\mathrm{d}v = xG(x) - \int_0^x G(v)\mathrm{d}v$$

Therefore, we can write:

$$\psi(x) = x - \frac{1}{G(x)}\int_0^x G(v)\mathrm{d}v$$

which implies:

$$\psi'(x) = g(x)\frac{\int_0^x G(v)\mathrm{d}v}{G^2(x)} \tag{6.66}$$

and:

$$\psi''(x) = g'(x)\frac{\int_0^x G(v)\mathrm{d}v}{G^2(x)} + \frac{g(x)}{G(x)}\left(1 - 2g(x)\frac{\int_0^x G(v)\mathrm{d}v}{G^2(x)}\right)$$

which can also be written:

$$\psi''(x) = \frac{g'(x)}{g(x)}\psi'(x) + \frac{g(x)}{G(x)}\left[1 - 2\psi'(x)\right] \tag{6.67}$$

Let us perform the same manipulations on the expression of function χ. Integration by parts yields:

$$\int_x^{v_M} v\mathrm{d}G(v) = [vG(v)]_x^{v_M} - \int_x^{v_M} G(v)\mathrm{d}v = v_M - xG(x) - \int_x^{v_M} G(v)\mathrm{d}v$$

then:

$$\int_x^{v_M} v\mathrm{d}G(v) = x\left[1 - G(x)\right] + \int_x^{v_M} \left[1 - G(v)\right]\mathrm{d}v$$

which yields:

$$\chi(x) = x + \frac{1}{1 - G(x)}\int_x^{v_M} \left[1 - G(v)\right]\mathrm{d}v$$

Thus:

$$\chi'(x) = \frac{g(x)}{\left[1 - G(x)\right]^2}\int_x^{v_M} \left[1 - G(v)\right]\mathrm{d}v \tag{6.68}$$

and:

$$\chi''(x) = \frac{g'(x)}{g(x)}\chi'(x) + \frac{g(x)}{1 - G(x)}\left[2\chi'(x) - 1\right] \tag{6.69}$$

We can now use relations (6.67) and (6.69) to prove that $g'(x) < 0$ for all $x \in [0, v_M]$ implies $\Delta'(x) > 0$ and that $g'(x) > 0$ for all $x \in [0, v_M]$ implies $\Delta'(x) < 0$.

Let us begin by considering the case where $g' < 0$. From (6.66) we have:

$$\psi'(0) = g(0)\operatorname*{Lim}_{x\to 0}\frac{\int_0^x G(v)\mathrm{d}v}{G^2(x)}$$

From l'Hôpital's rule, we know that:

$$\operatorname*{Lim}_{x\to 0}\frac{\int_0^x G(v)\mathrm{d}v}{G^2(x)} = \operatorname*{Lim}_{x\to 0}\frac{\left(\int_0^x G(v)\mathrm{d}v\right)'}{\left(G^2(x)\right)'} = \operatorname*{Lim}_{x\to 0}\frac{G(x)}{2g(x)G(x)} = \frac{1}{2g(0)}$$

and therefore:

$$\psi'(0) = \frac{1}{2}$$

This equality implies, together with equation (6.67), that:

$$\psi''(0) = \frac{1}{2}\frac{g'(0)}{g(0)} < 0$$

Therefore, in the neighborood of 0, $\psi'(x)$ is decreasing. Moreover, it can be shown that $\psi'(x) < 1/2$ for all $x > 0$. Assume that this is not the case, that $\exists\ x > 0$ such that $\psi'(x) > 1/2$ for some values of $x > 0$. Since $\psi'(0)$ is equal to $1/2$ and $\psi'(x)$ is decreasing in the neighborhood of $x = 0$, if $\psi'(x)$ increases to cross the horizontal line of coordinate $1/2$, the slope of $\psi'(x)$ has to be positive at this point. But (6.67) implies that we necessarily have, $\psi''(x) = \frac{g'(x)}{g(x)}\psi'(x) < 0$ when $\psi'(x) = 1/2$. The slope of $\psi'(x)$ is therefore necessarily negative, which implies that $\psi'(x)$ cannot cross the horizontal line of coordinate $1/2$ with a positive slope. Therefore, $\psi'(x) < \frac{1}{2}$ if $g'(x) < 0$ for all $x > 0$.

Let us now study function χ. From (6.68) we have:

$$\chi'(v_M) = \frac{g(x)}{[1 - G(x)]^2}\int_x^{v_M} [1 - G(v)]\,\mathrm{d}v = g(v_M)\lim_{x \to v_M}\frac{\displaystyle\int_x^{v_M} [1 - G(v)]\,\mathrm{d}v}{[1 - G(x)]^2}$$

From L'Hôpital's rule, we have:

$$\lim_{x \to v_M}\frac{\displaystyle\int_x^{v_M} [1 - G(v)]\,\mathrm{d}v}{[1 - G(x)]^2} = \lim_{x \to v_M}\frac{\left(\displaystyle\int_x^{v_M} [1 - G(v)]\,\mathrm{d}v\right)'}{\left([1 - G(x)]^2\right)'} = \lim_{x \to v_M}\frac{-[1 - G(x)]}{-2g(x)[1 - G(x)]} = \frac{1}{2g(v_M)}$$

and then:

$$\chi'(v_M) = \frac{1}{2}$$

This equality implies, together with (6.69), that:

$$\chi''(v_M) = \frac{1}{2}\frac{g'(v_M)}{g(v_M)} < 0$$

which means that $\chi'(x)$ is decreasing in the neighborhood of v_M. We will now show that $\chi'(x) > 1/2$ for all $x < v_M$. Let us suppose that this is not the case, that $\exists\ x < v_M$ such that $\chi'(x) < 1/2$. If $\chi'(x)$ crosses the horizontal line of coordinate $1/2$ before v_M, its slope is necessarily positive at this point. But (6.69) implies that $\chi''(x) = \frac{g'(x)}{g(x)}\chi'(x)$ when $\chi'(x) = 1/2$. The slope of $\chi'(x)$ is therefore negative, which implies that it is impossible that $\exists\ x < v_M$ such that $\chi'(x) < 1/2$. Therefore, $\chi'(x) > \frac{1}{2}$ if $g'(x) < 0$ for all x.

Finally, we have shown that:

$$\Delta'(x) = \phi'(x) - \chi'(x) \quad \text{with } \chi'(x) > \frac{1}{2} \quad \text{and} \quad \psi'(x) < \frac{1}{2} \text{ if } g'(x) < 0 \quad \text{for all } x \in [0, v_M]$$

Therefore, we have proved that $\Delta'(x) > 0$ if $g'(x) < 0$ for all $x \in [0, v_M]$.

The same reasoning can be used to show that $\Delta'(x) < 0$ if $g'(x) > 0$ for all $x \in [0, v_M]$.

REFERENCES

Abraham, K., & Haltiwanger, J. (1995). Real wages and the business cycle. *Journal of Economic Literature*, *33*(3), 1275–1364.

Agell, J., & Lundborg, P. (1995). Theories of pay and unemployment: Survey evidence from Swedish manufacturing firms. *Scandinavian Journal of Economics*, *97*(2), 295–307.

Akerlof, G. (1982). Labor contracts as partial gift exchange. *Quarterly Journal of Economics*, *87*, 543–569.

Akerlof, G., & Katz, L. (1989). Workers trust funds and the logic of wage profiles. *Quarterly Journal of Economics*, *104*, 525–536.

Akerlof, G., & Yellen, J. (1990). The fair wage-effort hypotheses and unemployment. *Quarterly Journal of Economics*, *105*, 255–283.

Argyle, M. (1991). *Cooperation: The basis of sociability*. London: Routledge.

Arvan, L., & Esfahani, H. (1993). A model of efficiency wages as a signal of firm value. *International Economic Review*, *34*, 503–524.

Azariadis, C. (1975). Implicit contract and underemployment equilibria. *Journal of Political Economy*, *83*(6), 1183–1202.

Azariadis, C. (1983). Employment with asymmetric information. *Quarterly Journal of Economics*, *98*, *Supplement*, 157–172.

Azariadis, C., & Stiglitz, J. (1983). Implicit contracts and fixed price equilibria. *Quarterly Journal of Economics*, *98*, *Supplement*, 1–12.

Baily, M. (1974). Wages and employment under uncertain demand. *Review of Economic Studies*, *41*(1), 37–50.

Baker, G., Gibbs, M., & Holmström, B. (1994a). The internal economics of the firm: Evidence from personnel data. *Quarterly Journal of Economics*, *109*, 881–919.

Baker, G., Gibbs, M., & Holmström, B. (1994b). The wage policy of a firm. *Quarterly Journal of Economics*, *109*, 921–955.

Bandiera, O., Barankay, I., & Rasul, I. (2009). Social connections and incentives in the workplace: Evidence from personnel data. *Econometrica*, *77*, 1047–1094.

Beaudry, P. (1994). Why an informed principal may leave rent to an agent. *International Economic Review*, *35*, 821–833.

Beaudry, P., & DiNardo, J. (1991). The effect of implicit contracts on the movement of wages over the business cycle: Evidence from microdata. *Journal of Political Economy*, *99*, 665–688.

Beaudry, P., & DiNardo, J. (1995). Is the behavior of hours worked consistent with implicit contract theory? *Quarterly Journal of Economics*, *110*, 743–768.

Becker, G. (1975). *Human capital*. New York, NY: Columbia University Press.

Bénabou, R., & Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, *96*(5), 1652–1678.

Bénabou, R., & Tirole, J. (2012). Laws and norms (IZA Discussion Paper No. 6290).

Bewley, T. (1995). A depressed labor market as explained by participants. *American Economic Review, Papers and Proceedings*, *85*, 250–254.

Bewley, T. (1999). *Why wages don't fall during a recession*. Cambridge, MA: Harvard University Press.

Blinder, A., & Choi, D. (1990). A shred of evidence on theories of wage stickiness. *Quarterly Journal of Economics*, *105*, 1003–1015.

Bloom, N., & Van Reenen, J. (2011). Human resource management and productivity. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 4B, pp. 1697–1767). Amsterdam: Elsevier Science.

Bolton, P., & Dewatripont, M. (2005). *Contract theory*. Cambridge, MA: MIT Press.

Brown, C. (1990). Firms' choice of method of pay. *Industrial Labor Relations Review*, *43*(3), 165–182.

Bull, C. (1983). Implicit contracts in the absence of enforcement and risk aversion. *American Economic Review*, *73*, 458–471.

Bull, C. (1987). The existence of self-enforcing implicit contracts. *Quarterly Journal of Economics*, *102*, 147–159.

Cahuc, P., & Dormont, B. (1997). Does profit-sharing increase productivity and employment? A theoretical model and empirical evidence on French micro data. *Labour Economics*, *4*, 293–319.

Campbell, C., & Kamlani, K. (1997). The reasons for wage rigidity: Evidence from a survey of firms. *Quarterly Journal of Economics*, *112*(3), 759–789.

Carmichael, L. (1983). Firm-specific human capital and promotion ladders. *Bell Journal of Economics*, *14*, 251–258.

Carmichael, L. (1985). Can unemployment be involuntary?: Comment. *American Economic Review*, *75*(5), 1213–1214.

Carmichael, L. (1989). Self-enforcing contracts, shirking, and life cycle incentives. *Journal of Economic Perspectives*, *3*, 65–84.

Carmichael, L. (1990). Efficiency wage models of unemployment: One view. *Economic Inquiry*, *28*, 269–295.

Chang, C., & Wang, Y. (1996). Human capital investment under asymmetric information: The Pigouvian conjecture revisited. *Journal of Labor Economics*, *14*, 505–519.

Chari, V. (1983). Involuntary unemployment and implicit contracts. *Quarterly Journal of Economics*, *98*, *Supplement*, 107–122.

Charness, G. (2004). Attribution and reciprocity in an experimental labor market. *Journal of Labor Economics*, *22*(3), 665–688.

Charness, G., & Kuhn, P. (2011). Lab labor: What can labor economists learn from the lab? In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 4A, chap. 3, pp. 229–330). Amsterdam: Elsevier Science.

Chiappori, P.-A., Macho, I., Rey, P., & Salanié, B. (1994). Repeated moral hazard: The role of memory, commitment, and the access to credit market. *European Economic Review*, *38*, 1527–1555.

Cooper, R. (1983). A note on overemployment/underemployment in labor contracts under asymmetric information. *Economics Letters*, *12*, 81–87.

Deci, E., Koestner, R., & Ryan, R. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, *125*, 627–668.

DeVaro, J. (2006). Internal promotion competitions in firms. *RAND Journal of Economics*, *37*, 521–542.

DeVaro, J., & Waldman, M. (2012). The signaling role of promotions: Further theory and empirical evidence. *Journal of Labor Economics*, *30*(1), 91–147.

Doeringer, P., & Piore, M. (1971). *Internal labor market and manpower analysis*. Lexington, MA: D. C. Heath.

Eriksson, T. (1999). Executive compensation and tournament theory: Empirical tests on Danish data. *Journal of Labor Economics*, *17*(2), 262–280.

Fehr, E., & Falk, A. (1999). Wage rigidity in a competitive incomplete contract market. *Journal of Political Economy*, *107*(1), 106–134.

Fehr, E., Kirchler, E., Weichbold, A., & Gächter, S. (1998). When social norms overpower competition: Gift exchange in experimental labor markets. *Journal of Labor Economics*, *16*(2), 324–351.

Fehr, E., Kirchsteiger, G., & Riedl, A. (1993). Does fairness prevent market clearing? An experimental investigation. *Quarterly Journal of Economics*, *108*(2), 437–459.

Freeman, R., & Rogers, R. (1999). *What workers want*. New York, NY: Russell Sage and Cornell University Press.

Gibbons, R., & Murphy, K. (1990). Relative performance evaluation for chief executive officers. *Industrial Labor Relations Review*, *43*, 30–52.

Gibbons, R., & Waldman, M. (1999a). A theory of wage and promotion dynamics inside firms. *Quarterly Journal of Economics*, *114*, 1321–1358.

Gibbons, R., & Waldman, M. (1999b). Careers in organizations: Theory and evidence. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 3, chap. 36, pp. 2373–2437). Amsterdam: Elsevier Science.

Gneezy, U., & List, J. (2006). Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica*, *74*(5), 1365–1384.

Gneezy, U., Meier, S., & Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior. *Journal of Economic Perspectives*, *25*(4), 191–209.

Gneezy, U., & Rustichini, A. (2000a). A fine is a price. *Journal of Legal Studies*, *29*(1), 1–17.

Gneezy, U., & Rustichini, A. (2000b). Pay enough or don't pay at all. *Quarterly Journal of Economics*, *115*, 791–810.

Gordon, D. (1974). A neo-classical theory of Keynesian unemployment. *Economic Inquiry*, *12*(4), 431–459.

Green, J., & Khan, C. (1983). Wage employment contracts. *Quarterly Journal of Economics*, *98*, *Supplement*, 173–187.

Groshen, E., & Krueger, A. (1990). The structure of supervision and pay in hospitals. *Industrial and Labor Relations Review*, *43*, 134–147.

Harris, M., & Holmström, B. (1982). A theory of wage dynamics. *Review of Economic Studies*, *49*, 315–333.

Hart, O. (1983). Optimal labor contracts under asymmetric information: An introduction. *Review of Economic Studies*, *50*, 1–35.

Hart O., & Holmström, B. (1987). The theory of contracts. In T. Bewley (Ed.), *Advances in economic theory. Fifth World Congress of the Econometric Society* (pp. 71–155). Cambridge, U.K.: Cambridge University Press.

Hart, O., & Moore, J. (1999). Foundations of incomplete contracts. *Review of Economic Studies*, *66*, 115–138.

Holmström, B. (1982). Moral hazard in teams. *Bell Journal of Economics*, *13*(2), 324–340.

Holmström, B., & Milgrom, P. (1987). Aggregation and linearity in the provision of intertemporal incentives. *Econometrica*, *55*, 303–328.

Holmström, B., & Milgrom, P. (1991). Multitask principal agent analyses, incentive contracts, asset ownership and job design. *Journal of Law, Economics and Organization*, *7*, 24–52.

Jovanovic, B. (1979). Job matching and the theory of turnover. *Journal of Political Economy*, *69*, 972–990.

Kotlikoff, L., & Gokhale, J. (1992). Estimating a firm's age-productivity profile using the present value of a worker's earnings. *Quarterly Journal of Economics*, *107*, 1215–1242.

Kreps, D. (1990). *A course in microeconomic theory*. Princeton, NJ: Princeton University Press.

Kruglansky, A. (1978). Endogenous attribution and intrinsic motivation. In D. Greene & M. Lepper (Eds.), *The hidden cost of reward*. Hillsdale, NJ: Erlbaum.

Kube, S., Maréchal, M., & Puppe, C. (2011). Do wage cuts damage work morale? Evidence from a natural field experiment. *Journal of the European Economic Association*, *11*(4), 853–870.

Laffont, J.-J. (1989). *The economics of uncertainty and information*. Cambridge, MA: MIT Press.

Lawler, E. (1994). Total quality management and employee involvement: Are they compatible? *Academy of Management Executive*, *8*(1), 68–76.

Lazear, E. (1979). Why is there mandatory retirement? *Journal of Political Economy*, *87*, 1261–1284.

Lazear, E. (1981). Agency, earnings profiles, productivity and hours restrictions. *American Economic Review*, *71*, 606–620.

Lazear, E. (1986). Salaries and piece rates. *Journal of Business*, *59*, 405–431.

Lazear, E. (2000). Performance, pay and productivity. *American Economic Review*, *90*, 1346–1361.

Lazear, E. (2011). *Inside the firm: Contributions to personnel economics*. New York, NY: Oxford University Press.

Lazear, E., & Moore, J. (1984). Incentives, productivity and labor contracts. *Quarterly Journal of Economics*, *99*(2), 275–296.

Lazear, E., & Oyer, P. (2010). Personnel economics. In R. Gibbons & J. Roberts (Eds.), *Handbook of organizational economics*. Amsterdam: North-Holland.

Lazear, E., & Rosen, S. (1981). Rank-order tournaments as optimum labor contracts. *Journal of Political Economy*, *89*, 841–864.

Lee, D., & Rupp, N. (2007). Retracting a gift: How does employee effort respond to wage reductions? *Journal of Labor Economics*, *25*(4), 725–761.

Lemieux, T., MacLeod, B., & Parent, D. (2009). Performance pay and wage inequality. *Quarterly Journal of Economics*, *124*(1), 1–49.

List, J., & Rasul, I. (2011). Field experiments in labor economics. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 4, no. 4). Amsterdam: Elsevier Science.

MacDonald, G. (1982). A market equilibrium theory of job assignment and sequential accumulation of information. *American Economic Review*, *72*, 1038–1055.

Macho-Stadler, I., & Perez-Castrillo, D. (2001). *An introduction to the economics of information: Incentives and contracts*. Oxford, U.K.: Oxford University Press.

MacLeod, B., & Malcomson, J. (1989). Implicit contracts, incentive compatibility and involuntary unemployment. *Econometrica*, *57*, 447–480.

MacLeod, B., & Malcomson, J. (1998). Motivation and markets. *American Economic Review*, *88*(3), 388–411.

Malcomson, J. (1984). Work incentives, hierarchy and internal labor market. *Journal of Political Economy*, *92*, 486–507.

Malcomson, J. (1999). Individual employment contracts. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 3, chap. 35, pp. 2291–2372). Amsterdam: Elsevier Science.

Mas-Colell, A., Whinston, M., & Green, J. (1995). *Microeconomic theory*. New York, NY: Oxford University Press.

Maskin, E., & Tirole, J. (1999). Unforeseen contingencies and incomplete contracts. *Review of Economic Studies*, *66*, 83–114.

Mauss, M. (1923). Essai sur le don. Forme et raison de l'échange dans les sociétés archaïques. *L'Année sociologique*, new series, *1*, 30–186.

Maximiano, S., Sloof, R., & Sonnemans, J. (2007). Gift exchange in a multi-worker firm. *Economic Journal*, *117*(522), 1025–1050.

McCue, K. (1996). Promotions and wage growth. *Journal of Labor Economics*, *14*(2), 175–209.

Meyer, M. (1992). Biased contests and moral hazard: Implication for career profiles. *Annales d'économie et de statistique*, *25*/*26*, 165–187.

Milgrom, P. (1988). Employment contract, influence activity and efficient organization. *Journal of Political Economy*, *96*, 42–60.

Mincer, J. (1974). *Schooling, experience and earnings*. New York, NY: Columbia University Press.

Moen, E., & Rosen, A. (2006). Equilibrium incentive contracts and efficiency wages. *Journal of the European Economic Association*, *4*(6), 1165–1192.

Murphy, K., & Cleveland, J. (1991). *Performance appraisal: An organizational perspective*. Boston, MA: Allyn and Bacon.

Myerson, R. (1979). Incentive compatibility and the bargaining problem. *Econometrica*, *47*, 71–74.

Offerman, T. (2002). Hurting hurts more than helping helps. *European Economic Review*, *46*(8), 1423–1437.

Okun, A. (1981). *Prices and quantities*. Washington, DC: The Brookings Institution.

Oyer, P., & Schaefer, S. (2011). Personnel economics: Hiring and incentives. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 4B, chap. 20, pp. 1769–1813). Amsterdam: Elsevier Science.

Paarsch, H., & Shearer, B. (1999). The response of worker effort to piece rate: Evidence from the British Columbia tree planting industry. *Journal of Human Resources*, *33*(4), 643–667.

Prendergast, C. (1993a). The role of promotion in inducing specific human capital acquisition. *Quarterly Journal of Economics*, *108*, 523–534.

Prendergast, C. (1993b). A theory of yes men. *American Economic Review*, *83*(4), 757–770.

Prendergast, C. (1999). The provision of incentives in firms. *Journal of Economic Literature*, *37*, 7–63.

Prendergast, C. (2002). The tenuous trade-off between risk and incentives. *Journal of Political Economy*, *110*, 1071–1102.

Rebitzer, J., & Taylor, L. (2011). Extrinsic rewards and intrinsic motives: Standard and behavioral approaches to agency and labor markets. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 4A, chap. 8, pp. 702–772). Amsterdam: Elsevier Science.

Rogerson, R., & Shimer, R. (2011). Search in macroeconomic models of the labor market. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 4A, chap. 7, pp. 619–700). Amsterdam: Elsevier Science.

Rosen, S. (1985). Implicit contracts: A survey. *Journal of Economic Literature*, *23*, 1144–1175.

Salanié, B. (1997). *The economics of contracts: A primer*. Cambridge, MA: MIT Press.

Sattinger, M. (1993). Assignment models of the distribution of earnings. *Journal of Economic Literature*, *31*, 831–880.

Shapiro, C., & Stiglitz, J. (1984). Equilibrium unemployment as a worker discipline device. *American Economic Review*, *74*, 433–444.

Shapiro, C., & Stiglitz, J. (1985). Can unemployment be involuntary?: Reply. *American Economic Review*, *75*(5), 1215–1217.

Shearer, B. (2004). Piece rates, fixed wages and incentives: Evidence from a field experiment. *Review of Economic Studies*, *71*, 513–534.

Simon, H. (1951). A formal theory of the employment relationship. *Econometrica*, *19*, 293–303.

Stouffer, S., Suchman, E., DeVinney, L., Star, S., & Williams, R., Jr. (1949). *The American soldier: Adjustment during army life.* Princeton, NJ: Princeton University Press.

Thomas, J., & Worral, T. (1988). Self-enforcing wage contracts. *Review of Economic Studies*, *55*, 541–553.

Tirole, J. (1992). Collusion and the theory of organizations. In J.-J. Laffont (Ed.), *Advances in economic theory: Sixth World Congress* (vol. 2). New York, NY: Cambridge University Press.

Waldman, M. (1984). Worker allocation, hierarchies and the wage distribution. *Review of Economic Studies*, *51*, 95–109.

Yellen, J. (1994). Efficiency wage model of unemployment. *American Economic Review, Papers and Proceedings*, *74*(2), 200–205.

CHAPTER 7

Collective Bargaining and Labor Unions

In this chapter we will:

- Present the different rates of unionization and collective bargaining coverage across countries and over time
- Study the determinants of unionization
- Study the behavior of unions
- Learn how employees and employers arrive at an agreement on how to share the benefits from productive activities
- Study the different approaches to bargaining theory
- Review the standard models of collective bargaining over wages and employment
- Explore the consequences of the opposition between insiders and outsiders
- See to what extent regression discontinuity design allows us to identify the causal impact of labor unions
- Assess empirically the impact of unions on wages, productivity, profits, employment, and investment

INTRODUCTION

Bargaining over labor contracts can take place at the individual level, between a worker and an employer, between an organization representing wage earners collectively and an employer, or between organizations representing wage earners collectively and organizations representing employers collectively. In most major industrialized countries, a significant proportion of wages is regulated by *collective agreements* that codify the agreements reached through bargaining between unions representing employees on one hand and employers or employers' organizations on the other. The purpose of this chapter is to study the course of events in a round of collective bargaining and their consequences.

We will begin by seeing that wage bargaining is organized very differently across OECD countries. Unions, collective national agreements, sectoral agreements, and

agreements at the firm level vary widely. Moreover, their roles may change considerably over the course of time. Still, in all countries the rates of unionization are influenced by the advantages that unions win for their members, and these advantages depend in turn on the legal context, the characteristics of the workforce, and the apparatus of production.

In order to grasp the impact of unions, it is essential to know their objectives and to understand how these interface with the objectives of firms, which are essentially to maximize profit. From this perspective, collective bargaining presents two conceptual barriers to analysis. The first is how to represent the objectives of the partners to the bargaining. These actors are not economic "agents" in the ordinary sense of the term but *organizations* (most often unions). The objectives of these organizations arise, one way or another, out of those of their component members. As we will see, economic analysis of collective decisions can shed light on the connection between individual preferences and those of collective organizations. Once past this barrier, there remains a second difficulty: how to represent the bargaining process. Since the early 1980s, developments in noncooperative game theory, especially dynamic games, and the attendant concepts of equilibrium have made it possible to overcome this obstacle as well. Dynamic game theory allows us to understand fundamental aspects of the behavior of actors, of the strategies they pursue as bargaining unfolds, and the manner in which they agree to conclude it and share the future benefits.

Knowledge of the objectives of unions allows us to understand the influence of collective bargaining on wages, employment, profit, and investment. Economic analysis predicts that unions exert a positive impact on wages but that their effect on employment may be positive or negative as circumstances dictate. These predictions are not easy to test because empirical research faces some major difficulties when it comes to pinpointing the causal impact of unions. The fact is, we cannot deduce the existence of a direct causal impact of unions on variables like wages or employment on the basis of correlations between these variables and indicators of union presence (rates of unionization, for example). Hence empirical researchers have developed strategies to try to pinpoint such a causal impact. One of them is regression discontinuity design, which is particularly well suited to analyzing the impact of unions in the legal context of the United States. We will examine what it can tell us and what its limitations are. We will also show that empirical research confirms the predictions of theoretical models, revealing that unions generally have a positive impact on wages, a negative impact on profits and investment, and an ambiguous impact on employment.

The analysis of collective bargaining presented in this chapter furnishes a primarily local explanation of wage setting and employment, in the sense that it compares the strategies of two clearly identified actors. It needs to be integrated into a general equilibrium model if we are to achieve an understanding of the *global* level of employment within an entire economy. In this chapter, we remain at the stage of partial equilibrium, with two actors (an employees' union and an employer) controlling a labor pool. Only in chapter 9 do we integrate bargaining over wages into a general equilibrium model.

Section 1 gives a sketch of the importance of labor unions and collective bargaining in the major industrialized countries and presents the determinants of union density. Section 2 gives the essential concepts and results of game theory used to analyze the unfolding of the negotiations and labor conflicts. They are applied in section 3, which lays out the basic models describing the consequences for wages, employment,

and investment of bargaining between an employer and a labor union. Finally, section 4 presents the empirical evidence regarding the consequences of collective bargaining.

1 FACTS ABOUT UNIONS AND COLLECTIVE BARGAINING

Collective bargaining plays an important role in most industrialized countries. The collective bargaining coverage, the concrete manner in which it occurs, the degree to which it is coordinated, and the variables involved are all sources of diversity and can affect the performance of a nation.

1.1 The Characteristics and Importance of Collective Agreements

A collective agreement is made up of a set of provisions negotiated between one or more employers and the representatives of their employees. Union density (the percentage of employees who are union members) and collective bargaining coverage (the percentage of employees covered by collective agreements) are measures of their importance. The level at which they are negotiated and the intervention of states vary significantly among the OECD countries.

1.1.1 Collective Bargaining Coverage and Union Density

Union density is to be distinguished from collective bargaining coverage. We present values for these two factors in OECD countries before examining their development.

Union Density

Figure 7.1 presents levels of union density (union members as a percentage of all employees) in the OECD countries in the 2000s. There is wide heterogeneity in rates of unionization. The Scandinavian countries have very high rates, surpassing 70%. At the other end, France, Turkey, and Estonia have rates of less than 10%. The average of all OECD countries is 19%.

Union density is typically higher in the public sector than in the private sector. Actually, union membership can be substantial in the public sector even in countries where overall union membership is low across the nation, such as the United States. This is illustrated by figure 7.2, which plots union density in the U.S. states for both the public and the private sectors. Union density can reach 60% in the public sector even in states where it does not exceed 10% in the private sector. Also, union membership varies significantly across industries within the private sector. Workers tend more often to be union members in construction and manufacturing than in the services, as illustrated by figure 7.3 in the case of the U.S. states.

Collective Bargaining Coverage

Union density is generally lower than collective bargaining coverage, which corresponds to the percentage of employees covered by collective agreements. Figure 7.4 displays the relation between union density and collective bargaining coverage for

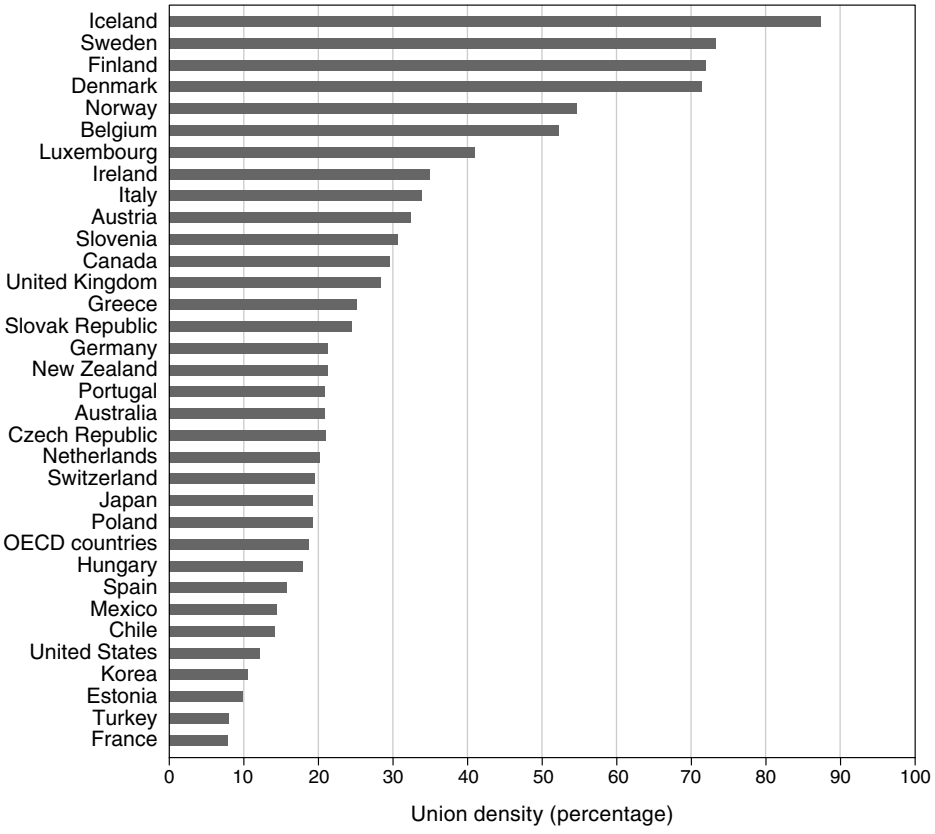


Figure 7.1
Average union density, 2000–2011.

Note: Union density is union members as a percentage of total employees.

Source: OECD labor market statistics.

34 OECD countries where these two variables are available. The average collective bargaining coverage is high, equal to 55.2% for these countries. On the other hand, the average union density proves to be significantly lower, amounting to 30.4%.

The gap between union density and collective bargaining coverage derives in large part from legal constraints and the institutional context. For example, in France and Spain collective agreements do not have the right to discriminate between union members and non-unionized workers. This prohibition may explain the large gap between the high collective bargaining coverage in these two countries and the remarkably low rate of union density. On the contrary, in Australia, New Zealand, the United States, and the United Kingdom, it is legal for collective agreements to discriminate between unionized and non-unionized workers, and this has certainly favored union membership. The upshot is that union density does not always provide a good measure of the power of unions. In France, though union density is low, unions play a preponderant role because they are *legally* empowered to represent workers in collective bargaining—and collective bargaining is compulsory in firms with more than 50 employees. In the

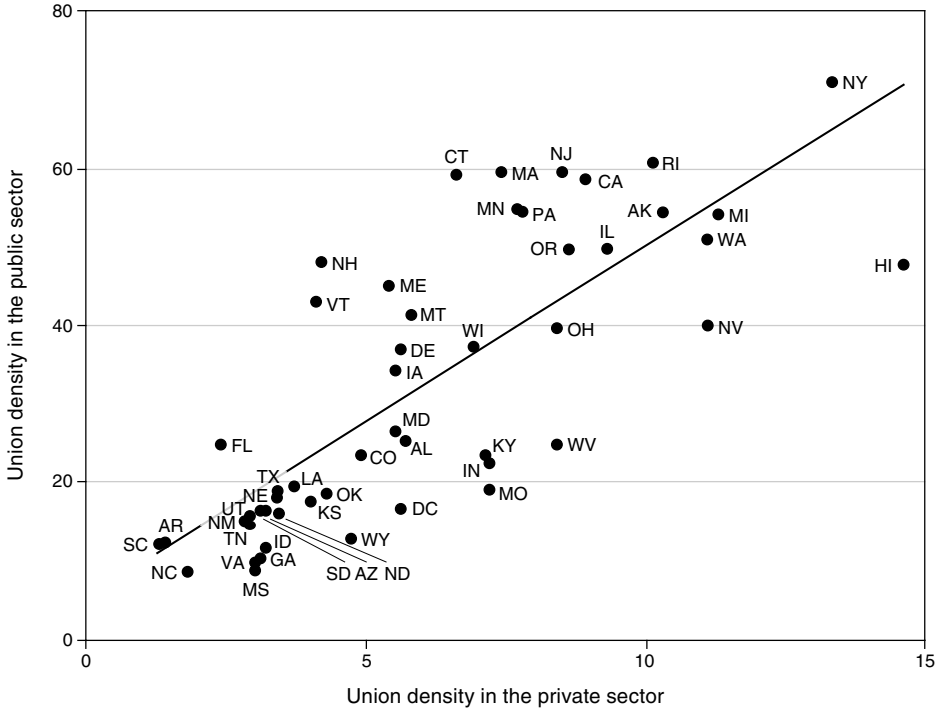


Figure 7.2
Union density in the public and private sectors in the United States in 2012.

Note: Current Population Survey (CPS) Outgoing Rotation Group (ORG) Earnings Files. Sample includes employed wage and salary workers, ages 16 and older. Density = percentage of employed workers who are union members.

Source: Union membership and coverage database constructed by Barry T. Hirsch and David A. Macpherson (Hirsch and Macpherson, 2003), www.unionstats.com/.

United States, on the other hand, where union density is higher, collective bargaining is only mandated by law if the majority of the employees in a plant vote in favor of union representation. This no doubt explains the low collective bargaining coverage in the United States. Overall, collective bargaining coverage is surely a more reliable indicator of the power of unions than union density.

Changes in Union Density and Collective Bargaining Coverage

Figure 7.5 depicts changes in union density for 13 countries between 1880 and 2000 (see Checchi and Lucifora, 2002, and Donado and Wälde, 2012, for studies of the evolution of union density). Comparison shows that over this period there were gains in all countries until the 1960s, and then gains continued only in Denmark and Sweden until the 2000s, and Norway and Italy until the 1980s. Major losses can be observed in, among others, Australia, Austria, the United States, France, Japan, Germany, and the United Kingdom. Overall, we see very different movements in union density and a significant drop in this indicator in many countries. The nonweighted average of union density has fallen off since the beginning of the 1980s.

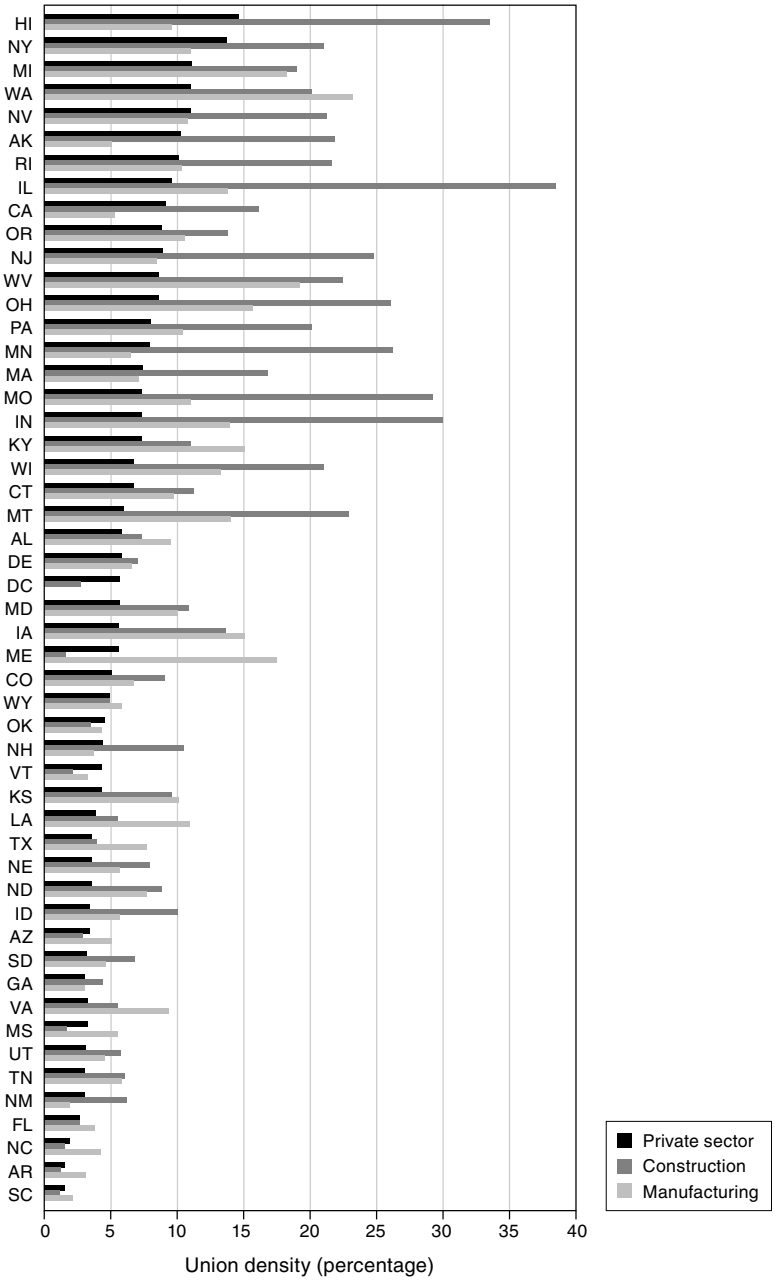


FIGURE 7.3
Union density in the private sector as a whole, in construction and in manufacturing in the United States in 2012.

Note: Current Population Survey (CPS) Outgoing Rotation Group (ORG) Earnings Files. Sample includes employed wage and salary workers, ages 16 and older. Density = percentage of employed workers who are union members.

Source: Union membership and coverage database constructed by Barry T. Hirsch and David A. Macpherson (Hirsch and Macpherson, 2003), www.unionstats.com/.

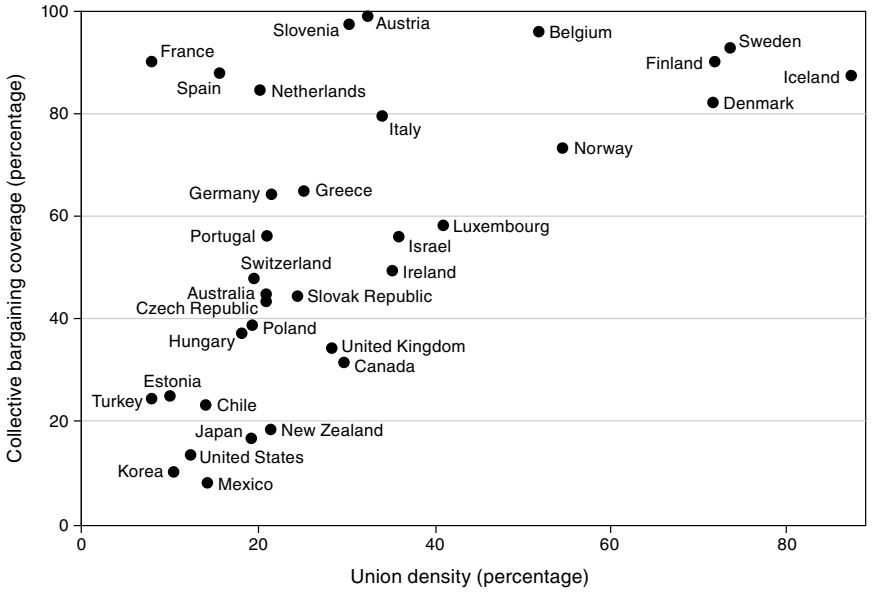


Figure 7.4
Collective bargaining coverage in the 2000s.

Source: Database on Institutional Characteristics of Trade Unions, Wage Setting, State (ICTWSS, for coverage). Years: average 2000–2010, or since 2000 to the latest available year.

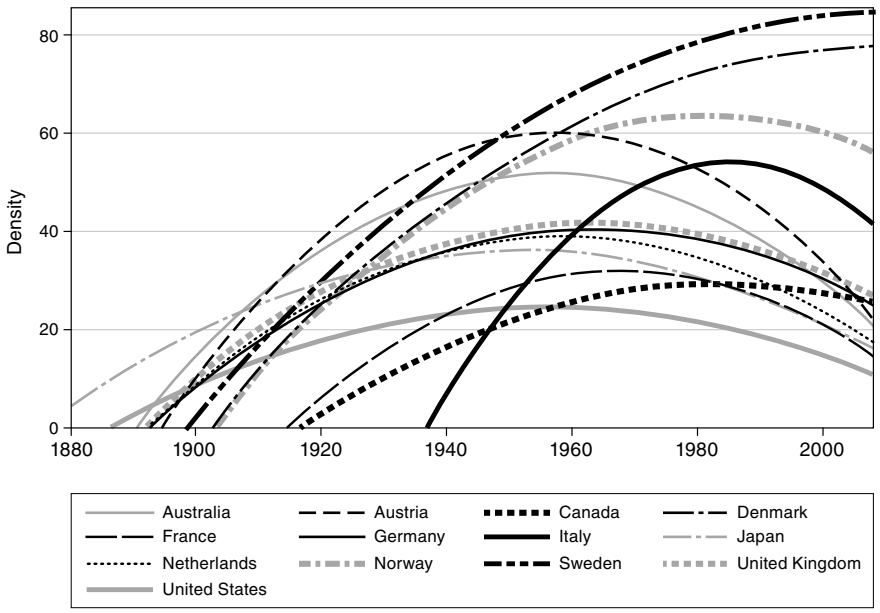


Figure 7.5
Changes in union density in 13 OECD countries, 1880–2008. These smoothed series are obtained by fitting the observed union density for each country to time and time squared variables plus a post 1959 dummy (to take into account the switch to OECD data after this date).

Source: Donado and Wälde (2012) data set.

Figure 7.6 depicts changes in the collective bargaining coverage in the same 13 OECD countries between 1960 and 2010. The extent of coverage decreased in 6 countries, including Germany, Japan, the United Kingdom, and the United States; it increased in France, Austria, Norway, and Sweden, and it peaked in the 1990s in Denmark and Canada. The United Kingdom experienced the sharpest drop in the 1980s. This particular change was the result of refusals to extend collective agreements made at the beginning of the 1980s. Overall, we see that the average of extents of coverage fell off very slightly between 1980 and 2010. But there is not a pervasive tendency for collective bargaining coverage to decline, contrary to what is observed for union density.

1.1.2 The Level at Which Bargaining Takes Place

To represent the unfolding of collective bargaining, we have to know whether it is taking place at the level of the firm, the industry, the region, or on a national scale. In reality, it is not always easy to classify countries by this criterion, for in most cases there is an overlap between negotiations taking place at several levels. Figure 7.7 presents a synthesis of coordination and government intervention in wage bargaining for 30 OECD

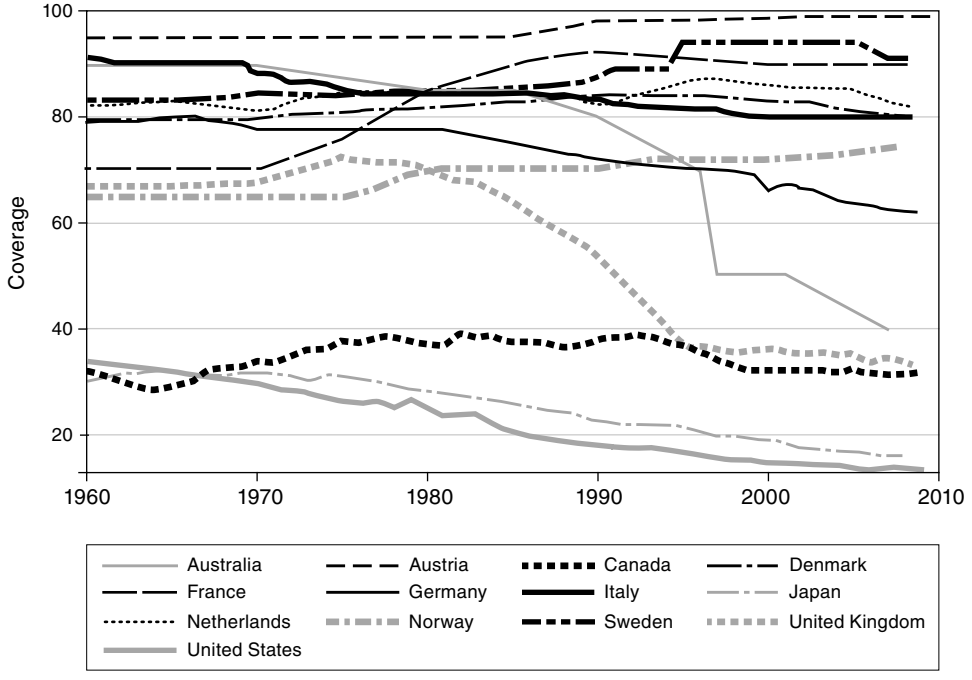


FIGURE 7.6
Coverage of collective agreements (as a proportion of all wage and salary earners in employment).

Source: Database on Institutional Characteristics of Trade Unions, Wage Setting, State Intervention and Social Pacts, 1960–2010 (ICTWSS).

countries. European countries tend to have more centralized processes of bargaining: there is often a mix of industrywide and economywide bargaining with guidelines set at a high level of centralization. English-speaking countries, except Ireland, typically have a more fragmented approach to wage negotiations, which most often takes place at the company level. In many countries, governments also influence the bargaining process, if not by participating directly in negotiations, at least indirectly through indexation measures, ceilings, minimum wages, legal extensions of collective agreements, or the setting of public-sector wages.

A distinction should be made between *explicit* and *implicit* coordination. Explicit coordination means actual bargaining between trade union confederations and confederations of employers at the national level. Implicit coordination derives either from the control exercised by union confederations over their members or from the fact that agreements reached in certain industries serve as models for the rest. Note that the absence of centralized bargaining does not necessarily imply the absence of national coordination; the latter may be implicit. Germany and Japan, for example, do not have collective bargaining at the national level, but there is a strong implicit coordination in both countries. In Japan, at the time of the "spring offensive" (*Shunto*), the unions announce the broad outlines of their wage demands vis-à-vis all the large firms in the country, and these guidelines are generally followed in individual cases. In Germany the logic of cohesion is different: it is agreements reached in the metalworking sector that traditionally serve as guidelines.

The plurality of forms of coordination makes it very difficult to classify systems of industrial relations according to their degree of centralization. Institutional structures are not carved in stone either. Certain countries like Sweden, the United Kingdom, and New Zealand are moving toward decentralization, while others, like Portugal, are moving toward a more centralized structure.

1.2 The Determinants of Union Density

We have just noted that rates of unionization vary considerably over time and across countries. Moreover, many countries have seen a marked diminution in rates of unionization over the last several decades. To understand these differences and trends, it is helpful to begin with the observation that joining a union is a choice that may be influenced by economic considerations. From this point of view, an individual opts for union membership if the advantages he derives from doing so are greater than the costs. Advantages and costs evidently depend on individual preferences, under the influence of social norms that can vary from country to country and epoch to epoch (Booth, 1995a). Advantages may comprise wage gains, improved working conditions, and jobs that are more stable. Costs essentially boil down to the payment of union dues, but in certain cases may also include opposition from management that might hamper the careers of union members. These advantages and costs depend on a range of factors. The first is the legal framework obtaining in the country in question. Then, competition in the product market might limit the ability of unions to extract large gains for their members. Last, structural aspects such as the sectoral makeup of production (private sector, public sector, manufacturing sector . . .) and the demographic composition of the workforce may also influence these costs and advantages.

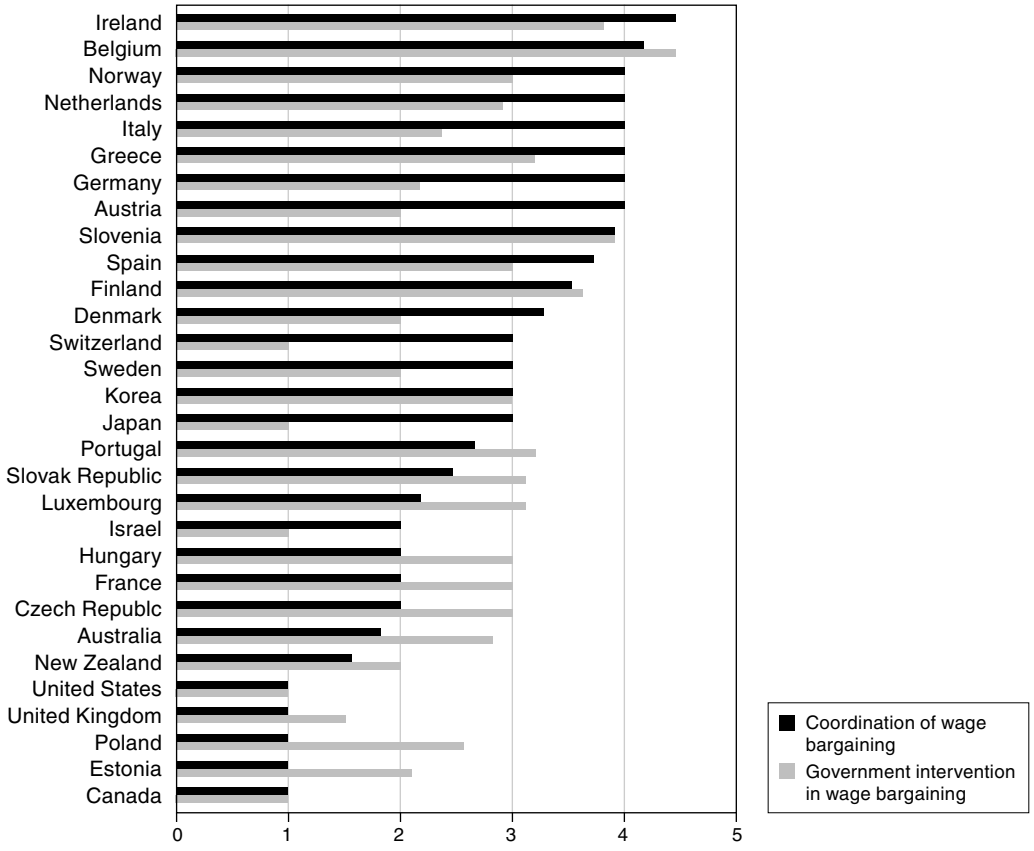


FIGURE 7.7

Wage bargaining coordination and government intervention in the OECD (average for the years 2000s). Coordination of wage bargaining: 5=strong coordination at national level, 1=strong fragmentation; Government intervention in wage bargaining: 5=strong intervention, 1=no intervention.[1]

Source: Database on Institutional Characteristics of Trade Unions, Wage Setting, State (ICTWSS).

[1] In this figure, the indexes mean the following:

Coordination of wage bargaining: 5 = economywide bargaining, based on (a) enforceable agreements between the central organizations of unions and employers affecting the entire economy or entire private sector, or on (b) government imposition of a wage schedule, freeze, or ceiling. 4 = mixed industry and economywide bargaining: (a) central organizations negotiate nonenforceable central agreements (guidelines) and/or (b) key unions and employers' associations set pattern for the entire economy. 3 = industry bargaining with no or irregular pattern setting, limited involvement of central organizations, and limited freedoms for company bargaining. 2 = mixed or alternating industry- and firm-level bargaining, with weak enforceability of industry agreements. 1 = none of the above, fragmented bargaining, mostly at company level.

Government intervention in wage bargaining: 5 = the government imposes private-sector wage settlements, places a ceiling on bargaining outcomes, or suspends bargaining; 4 = the government participates directly in wage bargaining (tripartite bargaining, as in social pacts); 3 = the government influences wage bargaining outcomes indirectly through price ceilings, indexation, tax measures, minimum wages, and/or pattern setting through public-sector wages; 2 = the government influences wage bargaining by providing an institutional framework of consultation and information exchange, by conditional agreement to extend private-sector agreements, and/or by providing a conflict resolution mechanism which links the settlement of disputes across the economy and/or allows the intervention of state arbitrators or Parliament; 1 = none of the above.

1.2.1 The Legal Framework

Unions cannot have real power if non-unionized workers can benefit from the same advantages as unionized ones. Olson (1965) pointed out that it is not enough to appeal for individual commitment and collective action to get people to sign up. It is necessary to offer real advantages in return for union dues. The same problem arises for collective goods. No one seriously proposes financing hospitals, roads, the police, or the courts on the basis of voluntary contributions. If that were the case, everyone would be happy to take advantage of these services but very few would be willing to contribute more than token sums toward paying for them, and in fact they could not function properly. That is why taxes are compulsory. Union membership depends on the same logic. Most wage earners think that unions are of use when it comes to defending their rights, their remuneration, and their working conditions. But if it is possible to profit from union action for free, the material incentive to sign up vanishes. Why pay union dues if collective agreements benefit all wage earners, unionized or not? It is for this reason that the legal framework, specific by definition to each country, exerts a fundamental influence on unionization.

The United States

The example of the United States well illustrates the influence of individual incentives on union membership. The rules governing unions and collective bargaining are set out in the National Labor Relations Act passed in 1935. For a union to be able to represent the wage earners in a firm, it is necessary in the first place that 30% of them sign a request that an election be held to decide whether there should be a union or not. If that threshold is met, a majority of the wage earners must then vote to effectively introduce a union. If they do, the union henceforth has a monopoly on collective bargaining and union membership is mandatory. The workplace becomes a "closed shop." But in 1947, the Taft-Hartley amendment to the National Labor Relations Act allowed each state to derogate from these rules by passing "right-to-work" laws that permit employees to delay joining the union or not to join it at all, although they may have to pay the full or partial equivalent of union dues. These non-joiners share in the advantages won by the union. The workplace then becomes a "union shop." Right-to-work legislation in some states may even provide for the "open shop," in which no employee can be compelled either to join the union or to pay it the equivalent of dues.

In the two years following the passage of the Taft-Hartley amendment, 12 states passed right-to-work legislation. By 2012 there were 24 such states. In them the rate of unionization in both the private and public sectors has fallen off more sharply than elsewhere (see Moore, 1998). Ellwood and Fine (1987) estimate that the adoption of right-to-work laws induces a drop in union membership of 5% to 10%. The unions are of course strongly opposed to right-to-work, which shrinks their resources and promotes "free riding." The AFL-CIO, one of the two main American union federations, denounces what it calls the "right to work for less." In the United States, as elsewhere, a decline in individual proclivity to join a union has led to a decline in unionization.

The OECD Countries

The Scandinavian countries have the highest rates of unionization in the OECD countries. Civic-mindedness has less to do with this enthusiasm for signing up than do the advantages that unions procure for their members. In Denmark, Finland, and Sweden,

unemployment insurance is managed by the unions but is not compulsory. If they lose their jobs, wage earners who opted out of unemployment insurance receive bare social assistance greatly inferior to the benefits paid out to the insured. But in order to pay in to unemployment insurance and be entitled to benefits, one must join a union. In other words, those wage earners who refuse to enroll in a union cannot enroll in the national unemployment insurance scheme and have no access to the associated services and benefits. This is known as the "Ghent system," from the name of the Belgian city where this system first saw the light of day in 1900.

In countries that adopted the Ghent system, the rise in unemployment in the 1970s and 1980s led to a rise in unionization, contrary to what occurred elsewhere. More wage earners signed up to the unions in order to acquire insurance in the face of rising unemployment. The observation that wage earners join unions when the latter offer services that answer their needs receives particularly strong corroboration from the case of Finland. The rate of unionization there lay in the vicinity of 33% at the start of the 1960s but rose from that time forward to peak at 85% in 1993. Conversely, over the two subsequent decades the rate of unionization in Finland dropped by more than 10 points. Bökerman and Uusitalo (2006) have shown that this fall is explainable, in a proportion of 75%, by the appearance in 1992 of an insurance fund independent of the unions that did not require the payment of any contribution upon entry. Quite logically, the majority of wage earners who joined the independent insurance scheme allowed their union membership to lapse. The pattern of unionization in countries that adopted the Ghent system well illustrates the importance of financial incentives in individual decisions about signing up to a union.

The Ghent system of administering unemployment insurance is not the only factor that may influence union membership. The extension clauses of collective agreements also play an important part. Checchi and Lucifora (2002) have shown that the automatic extension of the benefits obtained through collective bargaining to non-unionized wage earners is systematically associated with lower rates of unionization; their data bear on 14 European countries between 1950 and 1998.

The part played by self-interest in choices about union membership is also evident from a perspective perhaps less obvious. Research detects an inverse relation between, on one hand, the size of the legal minimum wage, the rigor of employment protection, and the degree to which wages are legally indexed to inflation, and on the other the rate of unionization (Checchi and Lucifora, 2002; Aghion et al., 2011). The reason could not be simpler: why bother to sign up to a union if the advantages are guaranteed by law? That is why the unions in the Scandinavian countries are fiercely opposed to a legal minimum wage, which would not fail to cause them to lose members. In Denmark, Finland, Norway, and Sweden there is no legal national minimum wage: it is the unions that negotiate wage floors (see chapter 12, section 2.1). In the converse situation where unions have little power, the minimum wage is often the only tool available to the government to push up low wages.

1.2.2 Competition

The profit of firms depends on their market power. The more a firm is earning thanks to the market power of its products, the greater the benefits a union can take action to extract from it. That being the case, unions ought to have a greater presence in firms

that do have significant market power. Abowd and Lemieux (1993) estimate that unions obtain collective agreements that are less advantageous for wage earners in Canadian firms most exposed to international competition. Slaughter (2007) finds a statistically significant correlation between falling union coverage and greater amounts of inward foreign direct investment in the United States. A possible interpretation is that the pressure of international capital mobility on U.S.-based companies raises labor-demand elasticities and alters the bargaining power of workers. Hence globalization, which increases the pressure of international competition on domestic firms, may help to explain the drop in unionization observed in many countries. But there is not yet a sufficient quantity of empirical research to allow us to specify with precision the impact of globalization on unionization.

1.2.3 Structural Determinants

Demographic change and changes in the sectoral makeup of production are other potential determinants of the pattern of unionization.

One reason that women are systematically less unionized than men is that they work part-time more often than men. Hence the increased entry of women into the labor market leads automatically to a falling off in unionization (Checchi and Lucifora, 2002).

The shift in the structure of employment from the manufacturing sector to the service sector leads automatically to a fall in unionization, to the extent that wage earners are less often unionized in the services than in industry. Yet these effects appear to explain a limited part of the fall in unionization in the OECD countries (Checchi and Lucifora, 2002; Hirsch, 2008). The size of the public sector, where rates of unionization are systematically higher, exerts an influence.

A bias in technological progress may also influence unionization. Technological progress that is biased in favor of skilled workers may increase the relative gains linked to the unionization of low-skilled workers with respect to the gains of more highly skilled workers (Aghion et al., 2011; Dinlersoz and Greenwood, 2012).

2 BARGAINING THEORY

Bargaining theory studies situations in which it is possible for rational agents to come to an agreement over how to share a quantity of (any) goods. Since Edgeworth (1881), a number of authors have sought to define the rational principles that preside over such a partition. Only recently has the work of Nash (1950, 1953), Stahl (1972), and Rubinstein (1982) systematically solved the bargaining problem. Nash launched the *axiomatic* approach, while Stahl and Rubinstein have developed the *strategic* approach. These two approaches make it possible to represent bargaining through simple models, which cast light on the notion of bargaining power and on the origins of conflict such as strikes.

2.1 The Precursors

The earliest analysts of bargaining were faced with the problem of the indeterminacy of the solution. Edgeworth (1881) had noted that this solution should be Pareto optimal,

since rational individuals would not accept a partition knowing that there existed other, more advantageous ones for at least one of the partners. But this criterion is not, in general, sufficient to define a unique solution. It simply indicates that agents exploit as far as possible the mutual benefits of cooperation. It is also necessary to explain *how* these benefits are shared. Zeuthen (1930) and Hicks (1932) were the first to propose solutions to the problem raised by Edgeworth.

The model of Hicks (1932) describes bargaining between a workers' union and the management of a firm on the hypothesis that each player possesses a *bargaining power* arising from his potential to hold out in case of conflict. This model can be presented graphically, with the duration of the strike on the horizontal axis and the wage on the vertical axis (see figure 7.8). The firm's "concession" schedule is denoted by the symbol (C). It is increasing, for the longer the strike lasts, the readier the employer is to accept high wages. Symmetrically, (R) designates the "resistance" schedule of the union. It is decreasing, for it seems natural to assume that the union will accept lower wages if the strike drags on. The wage settled on, denoted w^*, is determined by the intersection of curves (C) and (R). Assuming that the capacity of both sides to hold out is "common knowledge," Hicks deduces that strikes are only *potential*, since the firm and the union are perfectly capable of foreseeing the duration of the strike and the wage to which the bargaining will eventually lead.

The solution proposed by Hicks has the advantage of simplicity. Its drawback is that it remains very vague about the elements that determine the capacity of the players to hold out. Does it come from their risk aversion, their preference for the present, gains while the strike lasts, or alternative wages? It is indispensable for the theory of bargaining to state precisely the part played by these different factors. The model of Zeuthen (1930) adopts this perspective, since it represents the behavior of the players during

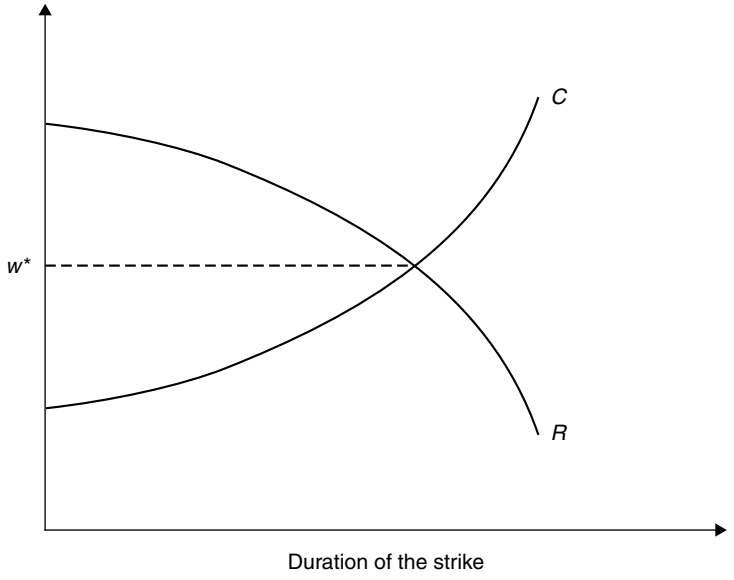


FIGURE 7.8
The model of Hicks.

the unfolding of negotiations. Today, however, the hypotheses adopted by this author to represent the strategic behavior of the players appear ad hoc, that is, incompatible with the postulate of rationality. It is game theory that has supervened in this field to clarify precisely how rational individuals behave as negotiations unfold.

The work of Nash (1950, 1953), Stahl (1972), and Rubinstein (1982) made it possible to solve the problem of bargaining in a systematic fashion. Nash (1953) came at the question from two different angles, which in practice turn out to be complementary. The first is the *axiomatic* approach, the aim of which is to define a priori the properties which it would seem natural for the solution to possess. The second is the *strategic* approach, in which the bargaining process is explicitly formalized but without prejudging the final properties of the solution. In this section, we examine the problem of a negotiation between two players (the extension to a larger number of participants raises no special difficulties; see for example Osborne and Rubinstein, 1990). We look first at the axiomatic approach, then at the strategic one, and finally analyze the linkage between these two ways of dealing with the bargaining problem.

2.2 The Axiomatic Approach

The works of Nash (1950, 1953) marked a decisive step in the analysis of bargaining between two agents. The aim of the axiomatic approach is to define the solution to the bargaining problem on the basis of a set of properties which it must "naturally" satisfy. Nash (1950, 1953) advances four such properties. To be precise, let G be the set of vectors of utility $u = (u_1, u_2)$ which players 1 and 2 can attain at the conclusion of the bargaining, and let $d = (d_1, d_2)$ be the vector of the utility obtained in a situation of status quo, in other words, the failure of bargaining. It is assumed that G is compact and convex, and that if $u \in G$, $u \geq d$. A set of solutions is then a function f linking every pair (G, d) to a vector $u^N = (u_1^N, u_2^N) \in G$ that satisfies the following four axioms:

(i) *Pareto optimality*

$$u \in G \quad \text{and} \quad u \geq u^N \Rightarrow u = u^N$$

(ii) *Invariance to positive affine utility transformations*

$\forall (a_1, a_2, b_1, b_2) \in \mathbb{R}^2_{*+} \times \mathbb{R}^2$. Let us define the affine function T which links every vector (u_1, u_2) to vector (u'_1, u'_2) such that $u'_i = a_i u_i + b_i$ for $i = 1, 2$; then $f[T(G), T(d)] = T[f(G, d)]$.

(iii) *Independence of irrelevant alternatives*

$$B \subset G \quad \text{and} \quad f(G, d) \in B \Longrightarrow f(B, d) = f(G, d).$$

(iv) *Symmetry*

If $d_1 = d_2$ and if $(u_1, u_2) \in G \Rightarrow (u_2, u_1) \in G$, then $u_1^N = u_2^N$.

The first two axioms signify respectively that the players exploit all mutual benefits and that the solution must not depend on a particular representation of their preferences. The fourth axiom postulates that the players are "interchangeable" in the following sense: when player 1 takes the place of player 2, he obtains the same gain as

the latter. This property supposes that the players have an identical "bargaining power." Axiom (iii) posits that if the players come to an agreement belonging to a subset B of the set G of all possible agreements, they will not change their attitudes if they confine themselves *straightaway* to taking into account only the possibilities offered by the subset B.

It is then possible to show that there exists a unique solution u^N satisfying properties (i) to (iv). It is defined by (see Nash, 1950, and Osborne and Rubinstein, 1990, p. 13):

$$u^N = \arg\max_{u \in G}(u_1 - d_1)(u_2 - d_2)$$

This so-called Nash solution thus corresponds simply to the maximization of the product of the net gains of the players. If we suppress the symmetry axiom (iv), we arrive at solution u^G called the "generalized Nash solution." It is defined by:

$$u^G = \arg\max_{u \in G}(u_1 - d_1)^{\gamma}(u_2 - d_2)^{1-\gamma} \qquad \gamma \in [0, 1]$$

In this expression, γ represents the bargaining power of player 1. Within the framework of the axiomatic approach, this concept lacks precision. We see below that the strategic approach allows us to establish a link between the preferences of players, the unfolding of the negotiation, and this notion of bargaining power.

It also needs to be emphasized that the properties stipulated by Nash have sometimes been criticized and that it is possible to imagine others (see for example Kalai and Smorodinsky, 1975, who discuss the independence axiom (iii)). As well, it can be difficult to define precisely the situation of status quo that corresponds to the gains d. During wage bargaining, are these gains the ones obtained if a strike occurs, or do they correspond to outside opportunities, that is, to gains obtained should the protagonists go their separate ways? To answer these questions, it is necessary to define the bargaining process completely. That is precisely the aim of the strategic approach.

2.3 The Strategic Approach

Stahl (1972) and Rubinstein (1982) worked out the first models of bargaining to use the theory of noncooperative games in a dynamic setting describing a process of offers and counteroffers. We describe the game that serves as a point of reference for all theories of collective bargaining first, then we look at the solutions.

2.3.1 A Noncooperative Bargaining Game

In dynamic noncooperative games, the relevant concept of equilibrium is that of "subgame perfect equilibrium." With this concept, it becomes possible to eliminate noncredible threats.

Rubinstein's Model

We present here a simplified version of the model of Rubinstein (1982). It is a game between two persons; their lifespan is infinite and unfolds in a sequence of periods. In each period it is possible for the two players to share a good, the size of which is

normalized to unity. In other words, in each period the two players have before them a "pie" of a given size, which they can share if they can reach an agreement about how much each will get. If they cannot agree, the pie is forfeited for that period and they must content themselves with their reservation utility. To be more precise: we assume that on even dates, player 1 proposes a partition $(x_t, 1 - x_t)$ which player 2 accepts or rejects. According to this partition, player 1 gets x_t and player 2 gets $(1 - x_t)$. On odd dates, player 2 proposes a partition $(y_t, 1 - y_t)$ which player 1 accepts or rejects. The agents have an infinite lifespan, and at every date their preferences are represented by strictly increasing and strictly concave utility functions denoted $u_1(x)$ and $u_2(1 - x)$. Parameters $\delta_1 \in (0, 1)$ and $\delta_2 \in (0, 1)$ designate the discount factors. The smaller δ_i the higher the preference for the present. We assume that each player is able to attain an instantaneous level of utility $\overline{u}_i = u_i(0),\ i = 1, 2$, at every date during the unfolding of the bargaining. These levels of utility are exogenous and correspond to what each agent can obtain as long as no agreement is reached. Bargaining ceases when an agreement is reached between the players. This agreement then applies to all the subsequent periods. In other words, at the date an agreement stipulating partition $(z, 1 - z)$ is accepted, the gains of players 1 and 2 are respectively defined by:

$$U_1 = \sum_{t=0}^{\infty} \delta_1^t u_1(z) = \frac{u_1(z)}{1 - \delta_1} \quad \text{and} \quad U_2 = \sum_{t=0}^{\infty} \delta_2^t u_2(1 - z) = \frac{u_2(1 - z)}{1 - \delta_2}$$

In this dynamic game, each agent adopts a strategy that specifies the offers that he makes and his reactions to the offers made by the other player. A strategy pair—one for each player—forms a Nash equilibrium if the strategy of one player is the best response to the strategy of the other. Hence, at Nash equilibrium, neither player has an interest in modifying his plan of action unilaterally *at the outset* of the bargaining.

Subgame Perfect Equilibrium

In dynamic games, however, the notion of Nash equilibrium thus defined is not completely satisfactory, since it offers no way to eliminate equilibria resting on noncredible threats. The example of the ultimatum game illustrates this point. Let us imagine a game unfolding over a single period (for example, date $t = 0$), and let us suppose that the strategy of player 1 consists simply of putting forward a partition offer $(x, 1 - x)$. Player 2's only options are to accept or reject it. If player 2 accepts the offer, he receives $(1 - x)$ and player 1 receives x, at which point the game ends. If player 2 refuses, each player must be satisfied with his reservation utility $\overline{u}_i,\ i = 1, 2$, and the game also ends. Let us now assume that player 2 adopts the following strategy: accept every offer $x \leq 1/2$ and refuse every offer $x > 1/2$. The outcome is a Nash equilibrium characterized by the partition $(1/2, 1/2)$, for at this point, each player obtains the highest possible gains given the strategy of the other player. For if player 2 undertakes to refuse every offer $x > 1/2$, player 1's best option is to offer $1/2$. In that situation, player 2 does indeed have an interest in undertaking to refuse any offer $x > 1/2$.

More generally, every partition $(x, 1 - x)$, $x \in [0, 1]$, corresponds to a Nash equilibrium. But this type of equilibrium implies that player 2's strategy rests on a *noncredible* threat, for when player 1 has put forward an offer x, player 2 has an interest in accepting every partition such that $u_2(1 - x) \geq \overline{u}_2 \equiv u_2(0)$, which is equivalent to $x \leq 1$. To undertake to refuse an offer $x \leq 1$ is thus not a credible threat.

For this reason, the general practice is to adopt a concept of equilibrium that eliminates noncredible threats. The idea is to search for each agent's optimal strategy *at every date* (and no longer just at the outset of the game), given all the other actions, past and present, chosen by the other agent. A pair of strategies respecting this condition is a Nash equilibrium for every *subgame*, that is, for every date t at which a player acts and not just for the initial game that begins at date $t = 0$. The consequence of this definition is that no agent individually has an interest in deviating from strategies that form a subgame perfect equilibrium, since each individual chooses his best strategy *at every instant*. In other words, the players do not prepare plans of action which, the moment they were put into operation, would be in their interest to renounce.

In the example just given of the ultimatum game, there is just one subgame perfect equilibrium. As we saw, player 2 accepts all $x \leq 1$ the moment he must respond to the offer of player 1. Player 1 knows this and so proposes the ultimatum $x = 1$. The only perfect subgame equilibrium of the bargaining game, at a period beginning with an offer from player 1, thus ends in a partition (1,0).

Let us now look at how the concept of subgame perfect equilibrium makes it possible to determine solutions to a bargaining process. Clearly the properties of solutions will be quite different, according to whether the horizon of the bargaining process is finite (Stahl, 1972) or infinite (Rubinstein, 1982).

2.3.2 Bargaining with a Finite Horizon

Subgame perfect equilibria are obtained by backward induction. We will show that it is in the interest of both players to agree *at the outset of the game* on a well-defined partition. For that purpose, let us suppose that the final date of the game, denoted n, is even. If no agreement has been reached by that time, player 1 would make the final offer, and player 2 would necessarily accept any value $x \leq 1$. So on the final date, player 1 would offer $x_n = 1$. Knowing that, player 2 could, at date $n-1$, make an offer acceptable to player 1 that took advantage of player 1's preference for the present: player 2 would offer partition $(y, 1-y)$ at date $(n-1)$, knowing that player 1 will obtain $\overline{u}_1 + \sum_{t=1}^{\infty} \delta_1^t u_1(1)$ by refusing and $\sum_{t=0}^{\infty} \delta_1^t u_1(y)$ by accepting. If we calculate the difference between these two quantities, we see that player 1 accepts all offers y such that:

$$u_1(y) - \overline{u}_1 \geq \delta_1 [u_1(1) - \overline{u}_1]$$

Since player 2 obtains no more than his reservation utility if player 1 refuses the offer, at date $(n-1)$ he makes an offer acceptable to player 1 which is the most advantageous for himself. This partition, $(y_{n-1}, 1-y_{n-1})$, is defined by:

$$u_1(y_{n-1}) - \overline{u}_1 = \delta_1 [u_1(1) - \overline{u}_1]$$

The reader can verify that $y_{n-1} < 1$ when $\delta_1 < 1$ and that $y_{n-1} = 1$ if $\delta_1 = 1$. This result means that wasting time is "costly" when an agent has a certain preference for the present ($\delta_1 < 1$). The line of reasoning now proceeds backward. Let us place ourselves at an even date $t \leq n-2$; player 1 makes an offer $(x_t, 1-x_t)$ knowing that at date $(t+1)$ player 2 will make an acceptable offer $(y_{t+1}, 1-y_{t+1})$. Should player 1's offer be refused, player 2 attains the level of utility $\overline{u}_2 + \sum_{\tau=1}^{\infty} \delta_2^{\tau} u_2(1-y_{t+1})$ and if it is accepted,

he obtains $\sum_{\tau=0}^{\infty} \delta_2^\tau u_2(1 - x_t)$. For player 1, it is optimal that these last two quantities be equal and x_t is thus defined by relation:

$$u_2(1 - x_t) - \overline{u}_2 = \delta_2 \left[u_2(1 - y_{t+1}) - \overline{u}_2\right] \tag{7.1}$$

Likewise, at odd date $(t - 1)$ player 2 makes an offer $(y_{t-1}, 1 - y_{t-1})$ knowing that player 1 will make an acceptable offer $(x_t, 1 - x_t)$ at date t. By refusing player 2's offer, player 1 obtains $\overline{u}_1 + \sum_{\tau=1}^{\infty} \delta_1^\tau u_1(x_t)$ while he attains a level of utility $\sum_{\tau=0}^{\infty} \delta_1^\tau u_1(y_{t-1})$ by accepting. For player 2, it is optimal that these last two quantities be equal, and y_{t-1} is thus defined by:

$$u_1(y_{t-1}) - \overline{u}_1 = \delta_1 \left[u_1(x_t) - \overline{u}_1\right] \tag{7.2}$$

Relations (7.1) and (7.2) form a system of difference equations describing the offers that one of the players makes at a given date in the knowledge that the other player will make an acceptable offer on the following date. Step by step, it becomes apparent that optimal strategies in subgame perfect equilibrium depend on both the *initial date* and the *final date* of the game. If the game begins at $t = 0$, it is player 1 who makes the first offer and the equilibrium corresponds to the partition $(x_0, 1 - x_0)$ where x_0 is the value of x_t deduced from the system of equations (7.1) and (7.2) at $t = 0$, with $x_n = 1$. Conversely, if the game begins at $t = 1$, it is player 2 who makes the first offer, and the equilibrium corresponds to partition $(y_1, 1 - y_1)$ where y_1 is the value of y_t deduced from the system of equations (7.1) and (7.2) at $t = 1$, with $x_n = 1$. As intuition suggests, preference for the present causes the players to have an interest in coming to terms right at the start of the game.

The hypothesis of a finite horizon lets us define a simple solution to the bargaining. It is seldom adopted, however, since it gives the terminal date of the game such essential importance. Bargaining over wages, for example, is not generally set in such a framework. For this reason, it is no doubt more relevant to take the view that the horizon is a priori infinite, since the date at which a bargaining process will come to an end is rarely spelled out.

2.3.3 Bargaining with an Infinite Horizon

With an infinite horizon, it becomes possible to analyze the stationary strategies of the agents directly. A precise description of the bargaining process will better enable us to grasp the notion of bargaining power.

The Outcome of the Bargaining

When the game horizon is infinite, all subgames beginning on even dates are identical, and the same holds true for all subgames beginning on odd dates. Since the players are rational, offers made at a date t will be the same as the ones that would have been made at date $(t + 2)$. Hence we can characterize a subgame perfect equilibrium based solely on stationary strategies. Let us assume that the strategy of agent 1 consists, on one hand, of accepting any offer $y \geq y^*$ and refusing any offer $y < y^*$ on odd dates, and on the other of offering x^* on even dates; and let us further assume that the strategy of agent

2 consists of accepting every offer $x \leq x^*$ and refusing every offer $x > x^*$ on even dates, and offering y^* on odd dates. For these two strategies to constitute a subgame perfect equilibrium, x^* must be the highest value that player 2 (who then receives $1 - x^*$) is prepared to accept at every date, given y^*, and y^* must be the smallest value that player 1 (who then receives y^*) is ready to accept, at every date, given x^*.

If, on any odd date, player 1 accepts offer y^*, he attains a level of utility $\sum_{t=0}^{\infty} \delta_1^t u_1(y^*)$; by refusing, he obtains $\overline{u}_1 + \sum_{t=1}^{\infty} \delta_1^t u_1(x^*)$. The smallest value y^* that player 1 is prepared to accept at every odd date, given x^*, is then defined by:

$$u_1(y^*) - \overline{u}_1 = \delta_1 [u_1(x^*) - \overline{u}_1] \tag{7.3}$$

Symmetrically, the highest value x^* that player 2 is prepared to accept at every even date, given y^*, is defined by:

$$u_2(1 - x^*) - \overline{u}_2 = \delta_2 [u_2(1 - y^*) - \overline{u}_2] \tag{7.4}$$

In appendix 8.1 to this chapter, we show that these two equations define a unique solution. The reader may note that relations (7.3) and (7.4) could have been obtained by making t go to infinity in equations (7.1) and (7.2) describing the solutions of the finite horizon game. As before, it is preference for the present that gives players an incentive to accept an offer. If the game begins at date $t = 0$, player 1 makes the first offer, and the solution to the bargaining is defined by partition $(x^*, 1 - x^*)$, for player 2 is indifferent between accepting this solution now or offering y^* at $t = 1$. Conversely, if the game begins at date $t = 1$, the solution to the bargaining is partition $(y^*, 1 - y^*)$.

Hence the bargaining process is only *virtual* in this model, for the players have no interest in wasting valuable time in bargaining when they know what the unique solution to the bargaining process is. So this model does not explain why bargaining should not be concluded immediately, nor (consequently) why it should be interrupted by strikes. We see below how conflicts may emerge in such a setting.

Bargaining Power

Although bargaining is taking place virtually, preference for the present plays a very large role. Each player's share decreases with preference for the present, which means that impatience reduces bargaining power. This general result can be illustrated with the help of utility functions $u_1(x) = x$ and $u_2(1 - x) = 1 - x$, from which we get $x^* = (1 - \delta_2)/(1 - \delta_1\delta_2)$ and $y^* = \delta_1(1 - \delta_2)/(1 - \delta_1\delta_2)$. Player 1's share increases with δ_1 and decreases with δ_2. Moreover, scrutiny of this solution shows that there is an advantage in making the first offer, since $x^* > y^*$.

The models of bargaining just laid out are of interest because they describe a process that ends with a unique, noncooperative solution. They show that it is necessary to know with precision the structure of the game, that is, the whole set of possible actions and the characteristics of the players, in order to define the solution. Note however that there exist other noncooperative games capable of representing a bargaining process. Binmore et al. (1986) built a model very close to the one set forth here, in which, for one thing, bargaining can be interrupted at every instant with a positive

probability, and for another, it is risk aversion that gives players an incentive to accept a sharing arrangement immediately (see Osborne and Rubinstein, 1990). Finally, it must be emphasized that, thanks to the precise description of the bargaining process, we now have a better grasp of the notion of bargaining power. In the models we have studied, this notion is linked to a preference for the present. An "impatient" player has less bargaining power than a more patient one. In the model of Binmore et al. (1986), it is risk aversion that determines the power of each player. An agent with more risk tolerance will have more bargaining power than an agent more hesitant to face the same risks. In the axiomatic approach, there was no suitable way to get at this idea of bargaining power. That notwithstanding, there are linkages between the strategic and axiomatic approaches, which we will now clarify.

2.3.4 The Relationship of the Axiomatic Approach to the Strategic One

Nash's axiomatic solutions can also be obtained as limit solutions to a noncooperative game in which the interval between two offers has been rendered arbitrarily small. Comparison of these two approaches clarifies the manner in which the status quo points are conceived.

Convergence on Nash's Axiomatic Solution

Binmore et al. (1986) showed that if the interval between successive offers in the Rubinstein game described above tends to zero, then the solution converges on the *axiomatic* solution of Nash (1953). When the elapsed time between two successive offers goes to zero, the two players are in the end going to make identical offers. More precisely: we show in appendix 8.2 to this chapter that if the two players have the same discount rates, the solution to the Rubinstein game goes toward x^N defined by:

$$x^N = \arg\max_x [u_1(x) - \overline{u}_1][u_2(1-x) - \overline{u}_2]$$

Thus we come back to the axiomatic solution of Nash from section 2.1 above, on the condition that we identify the gains made in a status quo situation with the payoffs obtained by the players during the unfolding of the negotiation.

Binmore et al. (1986) and Osborne and Rubinstein (1990) have shown that starting with the same Rubinstein bargaining game, we arrive at the generalized Nash solution if we assume that agents have different discount factors or different response times. For example, when the two players have different discount rates, $r_i > 0$, the discount factor of player i takes the expression $\delta_i = e^{-r_i\Delta}$, where Δ represents the interval between two successive offers. When this interval tends to zero, the solution of the Rubinstein bargaining game converges on the following generalized Nash solution (see appendix 8.2 of this chapter):

$$x^G = \arg\max_x [u_1(x) - \overline{u}_1]^\gamma [u_2(1-x) - \overline{u}_2]^{1-\gamma} \quad \gamma = \frac{r_2}{r_1 + r_2} \tag{7.5}$$

The most impatient player, the one for whom the discount rate r_i is the highest, has the weakest bargaining power.

The Status Quo Situation

The correspondence between the generalized Nash solution and that of the noncooperative Rubinstein game thus allows us to define both the status quo situation and the bargaining power of the players with precision. If the game that allows us to obtain the Nash solution is the one proposed by Rubinstein (1982), the payments in a status quo situation are different from those the players would obtain *outside* the relationship. In fact, they coincide with the gains they obtain *during the negotiation*. In the case of wage bargaining between a union and a firm, that means that the status quo payments should not be defined by outside wages for the workers or by the profits that could have been realized with other wage earners for the firm. These payments should correspond to what the agents obtain if there is a strike, that is, to what they can receive during the unfolding of the bargaining without resorting to outside opportunities. The latter should therefore appear in the form of *constraints* in the bargaining problem, since each player must, at the conclusion of the bargaining, attain a utility greater than that which outside opportunities offer him. More generally, interpretations of the Nash solution are contingent on the noncooperative game that underlies them. Hence, the axiomatic Nash solution can be obtained as the limit solution of a noncooperative game in which the bargaining could be interrupted at any moment with a positive probability. In this case, it is risk aversion and the probability of the negotiation breaking off that determine both the power of each player and the status quo point (see Binmore et al., 1986, and Osborne and Rubinstein, 1990).

The Limits of Rationality

Bargaining theory yields simple models that define the solution of bargaining between rational individuals. The attraction of formal consistency must not, however, hide a certain fragility. A number of experiments—see for example Ochs and Roth (1989), Camerer (2003), and Charness and Kuhn (2011)—have in fact sought to test the validity of the theory. They show that the choices of players are frequently different from what the rationality hypothesis and reasoning by backward induction, the foundations of the logic of subgame perfect equilibrium, would predict. For example, we have seen that if one of the players is able to announce an ultimatum in a credible manner, the subgame perfect equilibrium outcome leaves the other player nothing. Numerous experiments show that players rarely adopt such strategies. When placed in these conditions, it appears that the player who is able to announce an ultimatum will rather have a tendency to propose a "fair" partition, leaving a not insignificant part of the pie to the other player. Symmetrically, the other player will tend to refuse a partition that procures him a level of utility which he views as unfair. Finally, it is worth noting that backward induction in certain circumstances requires chains of reasoning too complex to be systematically followed through by the players. From this perspective, Berninghaus et al. (2012) have shown that individuals rely on backward induction in a finite game only when the horizon of the game is short enough.

Despite these limitations, the overwhelming majority of collective bargaining models follow the Rubinstein approach and adopt the generalized Nash solution—a model with simple and precisely defined microeconomic foundations, which can subsequently be enriched by abandoning or adding supplementary hypotheses.

2.4 Labor Conflicts: Strikes and Arbitration

The bargaining models presented above do not allow us to investigate labor conflict. Labor conflict is a phenomenon that most often takes the form of strikes (where they are permitted) or, as in the public sector in the United States, arbitration procedures.

2.4.1 Strikes

In the strategic models presented above, strikes are no more than threats which are never carried out, for the players are able to anticipate the consequences of offers and counteroffers perfectly, without having to experience them. There are, however, two ways of accounting for strikes in this context.

Multiple Equilibria

First, it is possible to alter the Rubinstein (1982) bargaining model marginally by supposing that the players have a choice between striking or "holding out," which means continuing to work under a lapsed contract during the unfolding of the negotiation. Solutions to the bargaining game then exhibit an array of subgame perfect equilibria, some with a strike and others without (Fernandez and Glazer, 1991). This approach is of interest because it shows that strikes can emerge from a bargaining process in which the actors can choose among a number of strategies in case of disagreement. Its limitation is the lack of a clear criterion for selecting a particular equilibrium from among the various possible equilibria. The predictive power of this strike model is thus very low.

Asymmetric Information

Another course is to assume that the players know each other's characteristics imperfectly. The delays in the bargaining then become a means to force the revelation of the information each agent disposes of. For example, to withstand a strike may be the only action that allows an employer to prove that he is incapable of paying a high wage (see the summary of Kennan and Wilson, 1993). The frequency of strikes ought then to rise with the degree of uncertainty about the profitability of firms.

Empirical Determinants of Strikes

Tracy (1986) tested the prediction that the frequency of strikes ought to rise with the degree of uncertainty about the profitability of firms using the volatility of the return on shares as a measure of uncertainty about profitability. He does indeed find a positive correlation between strikes and uncertainty which suggests that uncertainty about profitability may influence strike activity. Kuhn and Gu (1999) analyze the behavior of union-firm pairs that bargain sequentially. When unobserved components of firms' ability to pay are subject to correlated shocks, unions that bargain later in a sequence can acquire valuable information by observing previous bargaining outcomes in their industry. Kuhn and Gu derive the implications of this kind of learning in an asymmetric information model of wage negotiations, and they argue that the most robust implication is a lower incidence of strikes among "followers," who bargain later, than among "leaders" in wage negotiations. They find empirical support for this implication in a long panel of Canadian contract negotiations.

The models with asymmetric information have been enhanced by introducing a choice between striking or holding out. Cramton and Tracy (1992) emphasize that holding out is five times more common than striking in their data on labor conflict in the United States over the period 1970–1989. This means that it is important to take this characteristic of wage bargaining into account. The model of Cramton and Tracy (1992) predicts that strikes will be more frequent to the extent that past real wages covered by current contracts have shrunk due to inflation, for in this case holding out is more costly for workers. Their empirical results do indeed highlight a positive correlation between frequency of strikes and shrinkage of past wages. The gains to be made by striking or holding out evidently depend on the labor legislation in force. In particular, if the employer is permitted to hire replacement workers during strikes, that reduces the harm a strike can do and thus ought to exert downward pressure on the wage negotiated. Cramton et al. (1999) studied the effects of legislation allowing the hiring of temporary workers in strike situations, using Canadian data from 1967 to 1993. They estimate that wages are 4% lower and that the average duration of strikes is two weeks shorter when such legislation is on the books.

2.4.2 ARBITRATION

In the United States, arbitration is frequently used in the public sector when strikes are forbidden. The arbitrators are generally experts picked by the employers and unions, following a procedure set out by the government. For example, in selecting an arbitrator for police and firefighters in New Jersey, the New Jersey Public Employment Relations Commission must present a list of seven candidates to representatives of the employers and employees; each side has a right of veto over three of the names and must rank its preferences among those who remain. Usually arbitration procedures fall into two categories. In *conventional* arbitration, the arbitrator is free to impose a settlement as she sees fit. In *final-offer arbitration*, the two sides each make a final offer, and the arbitrator must select one of them.

It is possible to study the effects of arbitration procedures using strategic bargaining models, as long as the objectives of the arbitrator are specified. For example, Farber and Bazerman (1986) assume that the arbitrator attempts to minimize the sum of square deviations between her proposals and the allocations preferred by the parties to the dispute (which permits her to maximize her chances of being nominated again in the future, according to Farber and Bazerman). Let us consider a conventional arbitration procedure in the bargaining game from section 2.3 above, where two players are trying to split up a pie of size 1. When the arbitrator allots share x to player 1, player 2 obtains a complementary share of $1 - x$. If we assume that the players are risk neutral and that each player wants to obtain the whole pie, the arbitrator's problem can be written as follows:

$$\min_x \ \alpha(x-1)^2 + (1-\alpha)x^2, \quad 0 < \alpha < 1$$

In this expression, α is a parameter representing the relative weight of player 1 in the arbitrator's goal and the term $(x-1)^2$ (or x^2) designates the square deviation between

the share allotted to player 1 (or 2) and his preferred allocation. The solution of this problem corresponds to the partition x_a chosen by the arbitrator. It is defined by:

$$x_a = \alpha$$

Let us assume that the players know the arbitrator's preference α imperfectly and that they anticipate values α_1 and α_2 which may turn out to be different. If going to arbitration has a cost denoted by c_i, $i = 1, 2$ (the cost of waiting for the arbitrator's decision, for example), player 1 anticipates that with this procedure, his net gain will be equal to $\alpha_1 - c_1$, while player 2 anticipates a net gain amounting to $1 - \alpha_2 - c_2$. Consequently, the players have no interest in going to arbitration if they can agree on a partition x satisfying the two inequalities $x \geq \alpha_1 - c_1$ and $1 - x \geq 1 - \alpha_2 - c_2$. Partitions negotiated without mediation thus fall in the interval $[\alpha_1 - c_1, \alpha_2 + c_2]$. This interval is not empty when $\alpha_1 - c_1 \leq \alpha_2 + c_2$. In the opposite case, which corresponds to inequality $\alpha_1 - \alpha_2 > c_2 + c_1$, the players will resort to arbitration. This model shows that the probability of using the arbitrator diminishes with the sum $(c_2 + c_1)$ of the costs and increases with the relative optimism $(\alpha_1 - \alpha_2)$ of each side.

For the purpose of illustrating this example in a simple context, let us assume that the players know the true value of α and that the bargaining process is represented by the Rubinstein model from section 2.3.3 above, in which there exists an exogenous probability that the outcome of the negotiation will be settled by conventional arbitration between each offer and counteroffer. More precisely, let us suppose that the probability that the bargaining will break off during the interval of time Δ following a refusal by player i is equal to $e^{\Delta p_i}$. That being so, the bargaining is represented by the problem (7.5) from section 2.3.4 with $\bar{u}_1 = \alpha - c_1$, $\bar{u}_2 = 1 - \alpha - c_2$, and $\gamma = p_2/(p_1 + p_2)$. We then get the solution[2] $x = \gamma(\alpha - c_2) + (1 - \gamma)(\alpha - c_1)$. We see that the outcome of bargaining in the presence of a conventional arbitration procedure depends on the probability of the arbitrator intervening, on the arbitrator's preferences, and on the cost of arbitration. The bargaining model in the presence of a final-offer arbitration procedure comes to the same qualitative conclusions (Farber and Bazerman, 1986; Ashenfelter et al., 1992).

This model shows that decisions assigned to arbitrators play a determining role. They influence not just the occurrence of arbitration procedures but also the wages settled by bargaining, even without the effective intervention of an arbitrator. Empirical work suggests that arbitrators all have approximately the same criteria, depending on the observable characteristics of employers and employees. In other words, the "interchangeability hypothesis" concerning arbitrators, based on the assumption that arbitrators maximize their probability of being nominated again in the future, which we did assume in the model above, is not generally rejected (Farber and Bazerman, 1986; Ashenfelter, 1987). This model is also compatible with the fact that recourse to an arbitrator is less common when the costs of going to arbitration are higher (Ashenfelter et al., 1992). Moreover, comparing the wages of police officers in states with a system of arbitration and in ones without between 1969 and 1998 in the United States, Ashenfelter and Hyslop (2001) estimate that the presence of a system of arbitration has

[2]This result can, as an exercise, be demonstrated simply by using appendix 8.2, with the hypothesis that the players have no preference for the present and that gains are zero during the unfolding of the negotiation.

no significant effect on the average wage, and it appears to reduce the dispersion of wages only slightly. Finally, all empirical work suggests that the two arbitration procedures, conventional and final-offer, have similar effects on the frequency of conflicts and on wages.

We have been studying the bargaining process in a very general framework, which might as easily have comprised individual negotiations as collective ones. This has allowed us to present simple models which clarify the factors that influence the partition of surpluses between two protagonists. In what comes next, we direct our attention to collective bargaining between workers' representatives and employers.

3 MODELS OF COLLECTIVE BARGAINING FOR WAGES, EMPLOYMENT, AND INVESTMENT

What is the impact of unions on employment, wages, and investment? Economic theory predicts that, thanks to their bargaining power, unions have a positive impact on wages. Moreover they reduce the dispersion of their members' wages. As for their impact on employment, theory predicts that it may be positive, neutral, or negative according to circumstances: it all depends on the variables at stake in the bargaining. The impact of unions on investment is ambiguous, depending entirely on whether or not it is possible to renegotiate wages once the investment has been made.

3.1 The Objective of Labor Unions

The economic analysis of unions has long been highly controversial. The assumption that complex political institutions had rational objectives like those dealt with in the economic theory of individual choice appeared too reductive to be relevant. Hicks (1932) took the view that "to protect the customary standard of life (which may be conceived as a money wage or, in times of monetary disturbance, a real wage), to maintain fair wages, and to secure to the workers a share in exceptional profits, are the usual aims of the wage policy of trade unions" (p. 140). In other words, unions simply demanded a "fair wage" (or a "customary standard of life") determined by overall social conditions. Hicks saw no need to resort to choice theory in order to represent the behavior of a union.

Dunlop (1944) was the first to declare that "an economic theory of a trade union requires that the organization be assumed to maximize (or minimize) something" (p. 4). According to this author, the aim of a trade union is to maximize the total amount of wages received by its members. Ross (1948) reacted by insisting on the essentially political nature of unions; he criticized Dunlop's approach by emphasizing that unions are not made up of identical individuals and that the content of their decisions reflects the struggle for power both within the membership and between members and leaders of unions. This objection highlights an important limitation of Dunlop's analysis by showing that the heterogeneity of a union's members affects its aims. The distribution of various individual characteristics, the way in which the leadership is selected, the organization of

elections, the recruitment of members, and a number of given institutional factors are all capable of influencing a union's behavior. Ross's objection signifies, in other words, that it is insufficient to postulate a union objective independent of its members' preferences and its own institutional characteristics. So it becomes necessary to analyze the relationship between the union's preferences and those of its members. This has been the goal of the economic theory of unions for several decades; some empirical studies have helped to clarify the nature of union aims.

3.1.1 Union Preferences and Individual Preferences

The representation of the preferences of a union depends mainly on the homogeneity or heterogeneity of the individuals it comprises. To this dimension we sometimes have to add certain aims proper to the union leadership.

A Union with Identical Members

A union composed of identical members has, since the work of Drèze and Modigliani (1981), MacDonald and Solow (1981), and Oswald (1982), been the basis for representations of union preferences. The assumption is that the union defends the interests of N identical workers who form its "labor pool." Every union member supplies one unit of labor if the real wage w exceeds the reservation wage $\overline{w}$, equated to the income of an unemployed person. Individual preferences are represented by an indirect utility function of the Von Neumann and Morgenstern type, or $v(.)$, strictly increasing with respect to income. The labor demand addressed to the union is denoted L. The (identical) workers each have the same probability $(1 - L/N)$ of being unemployed when $L < N$. If this inequality is satisfied, the probability of being employed amounts to L/N. Conversely, if labor demand is greater than or equal to the size of the labor pool, the probability of being hired is equal to 1, and a worker's expected utility is simply $v(w)$. We then assume that the objective of the union consists of maximizing the expected utility $\mathcal{V}_s$ of its members. This last quantity is defined by:

$$\mathcal{V}_s = \ell v(w) + (1 - \ell)v(\overline{w}), \quad \ell = \min(1, L/N) \tag{7.6}$$

Given that the size N of the labor pool is exogenous, that comes to the same thing as assuming that the union maximizes the sum $N\mathcal{V}_s$ of the utility of its members. Such a union is then described as "utilitarian." If the workers have no aversion to risk—$v'(.)$ is then a constant—this specification is compatible with the hypothesis that the union maximizes the "union rent" (Rosen, 1970; De Menil, 1971), defined by the product $\ell(w - \overline{w})$. Dunlop's hypothesis that the union objective is to maximize the total wage bill further requires $\overline{w} = 0$.

The hypothesis that union members are identical allows us to lay precise microeconomic foundations for union preferences. This precision is however gained at the expense of realism. The heterogeneity of union members poses different problems according to how the union functions. If the organization is perfectly democratic, its preferences can be deduced from those of its members by analyzing the outcome of a vote. On the other hand, the objectives of union leaders play a determining role if they enjoy strong discretionary powers.

A Perfectly Democratic Union with Different Members

We learn from the analysis of collective decisions (Arrow, 1963) that the preferences expressed through the outcome of a vote by rational agents are themselves rational—i.e. that they define a complete, reflexive, and transitive ordering—when (i) majority rules; (ii) agents vote sincerely and do not try to shape the outcome by announcing their intentions beforehand; (iii) they are voting on a single variable; and (iv) the utility function of each individual admits only one maximum with respect to the variable on which they are voting. If these conditions are satisfied, the decision taken expresses the preferences of the *median* voter.

These results show that when the heterogeneity of union members is introduced, the definition of a union's objective rests on highly restrictive hypotheses. In this regard, as Blair and Crawford (1984) point out, hypothesis (iii) that the vote can only address one question proves extremely limiting. In practice, collective negotiations embrace a number of topics (see Hartog and Theeuwes, 1993). But to allow both heterogeneity of members and a number of variables to be negotiated becomes a fraught exercise, since the rationality of union decisions is no longer guaranteed (preferences are not necessarily transitive). For this reason, studies analyzing a union composed of diverse members make the assumption that the vote is exclusively about wages (Blair and Crawford, 1984; Booth, 1984; Carruth et al., 1986). Their main contribution is to show that the union has a slight preference for employment if the median voter has a slim probability of losing her job, as will be the case if layoffs are made by seniority, as in the United States in the unionized sector, or if the median voter possesses specific human capital that gives her an advantage with respect to other workers hired more recently.

Conflicts Between Union Leadership and Membership

In many institutions, the leadership has discretionary power, and their objectives do not necessarily coincide with those of their membership. For example, the social prestige, the advantages in kind, and the remuneration of members of the leadership generally depend on the importance of the institution they represent. That being the case, it is most often assumed that their objective is to maximize the size of their organization (Ross, 1948; Atherton, 1973; Martin, 1980; Farber, 1986). Hence it is possible simply to study the consequences of the discretionary power of the leadership. If we assume (Farber, 1986) that the size of the union increases with the number of workers employed—a hypothesis justified by the observation that union density among the unemployed is much weaker than it is among workers who do have a job—union leaders in a position to fix the wage level unilaterally, subject to the constraint of decreasing labor demand, would set a wage equal to the reservation wage, so as to maximize the level of employment compatible with the participation constraint of workers. This conclusion, as Lewis (1963) points out, shows that a "boss dominated union" keeps its members from profiting from its monopoly power, since this power is used exclusively for the benefit of the leadership; the latter attain their objectives in a situation of perfect competition.

This rapid review of work dealing with the problems posed by the heterogeneity of union members reveals that the economic analysis of union behavior remains very crude and that the topic has not attracted recent research. Nonetheless, a couple of things have been learned. For one thing, it is possible to represent the preferences of the union in terms of employment and wages on a precise microeconomic basis. For another, the

goals of the union depend not just on the preferences of its members but also on its institutional structure. From this perspective, the purely economic approach to trade unionism, which deduces the union objective from the objectives of its members, is to a certain extent relevant. But the highly restrictive nature of its hypotheses (identical members, validity conditions of the theorem of the median voter) leads to a neglect of institutional characteristics that may have important influence on employment and wages.

3.1.2 Union Goals According to Empirical Studies

Useful information about union goals comes from several sources. Freeman and Medoff (1984, chapter 14) have undertaken studies using statements made by union members and leaders. They conclude that the American union movement functions very democratically, particularly at the local level. Work by labor sociologists is also instructive, but is difficult to apply to the formal definition of the objective function of a union. For this reason, econometric research would seem to be the most suitable approach. The econometric approach is to estimate wage and employment functions on the assumption that remuneration is set by a union maximizing its objective function subject to the constraint of the labor demand. The estimation of the coefficients of the wage and employment equations thus obtained then allows us to characterize union preferences. For example, Dertouzos and Pencavel (1981) and Pencavel (1984) have tested a utility function of the Stone-Geary type, taking the following form:

$$\mathcal{V}_s = (w - w^0)^\theta (L - L^0)^{1-\theta}; \quad \theta \in [0,1] \tag{7.7}$$

In this expression, w^0 and L^0 represent respectively the minimal wage and employment levels accepted by the union. Parameter θ measures the relative importance of wages. This formulation allows us to recover (as particular cases) the objective of the total wage bill ($\theta = 1/2,\ L^0 = w^0 = 0$) postulated by Dunlop (1944), and the objective of union rent ($\theta = 1/2,\ w^0 = \overline{w},\ L^0 = 0$) put forward by Rosen (1970) and De Menil (1971).

Dertouzos and Pencavel (1981) used data regarding six local branches of the International Typographical Union (United States), for the period 1946–1965. The data describe the production processes of a particular newspaper, the *Cincinnati Post*. In these data, W stands for the hourly nominal wage for journeymen printers at the *Post*, and L is the number of full-time typographical workers in the *Post* composing room. The real wage, $w = W/P$, is equal to the nominal wage divided by the consumer price index P.

Each local union is assumed to maximize the objective just set out subject to the constraint of a linear labor demand function deduced from minimization of costs which is written:

$$L = \alpha_0 + \alpha_1(W/r_1) + \alpha_2(r_2/r_1) + \alpha_3 X + \alpha_4 D \tag{7.8}$$

where r_1 is the price at which the product is sold, r_2 is an index of the non-wage cost of production, X is the number of lines of advertising sold annually, and D is a dummy variable that takes the value of zero from 1946 to 1957 and of unity for the later years

because the *Cincinnati Post* merged with the *Cincinnati Times-Star* in 1958. The first-order condition is written:

$$\frac{\theta}{\theta - 1} = \frac{\alpha_1 (w - w^0)}{r_1 P (L - L^0)} \tag{7.9}$$

Eliminating L between the two previous equations we get:

$$w = \beta_0 + \beta_1 \left(\frac{r_1}{P}\right) + \beta_2 \left(\frac{r_2}{P}\right) + \beta_3 \left(\frac{r_1 X}{P}\right) + \beta_4 \left(\frac{r_1 D}{P}\right) \tag{7.10}$$

where $\beta_0 = (1 - \theta) w^0, \beta_1 = \theta (L^0 - \alpha_0)/\alpha_1, \beta_i = -\theta \alpha_i \alpha_1^{-1}, i = 2, 3, 4$.

Dertouzos and Pencavel then estimate the system of two equations (7.8) and (7.10), exploiting the fact that the consumer price index P enters the system only in equation (7.10). To deal with the fact that wages and employment might be determined by confounding variables, they use P as an instrument for W/r_1 in the employment equation (7.8) to estimate the α_i, denoted by $\hat{\alpha}_i$. P is a good instrument if it is correlated to W/r_1 but not with other nonobservable variables correlated to employment. Then, they estimate equation (7.10) where the $\hat{\alpha}_i$ are substituted for the α_i. These results rely on the assumptions that r_1, the price at which the product is sold, r_2, the non-wage cost of production, and X, the number of lines of advertising sold annually, are exogenous with respect to wage and employment, that is, they are neither influenced by wage or employment nor influenced by confounding variables. Obviously, such assumptions are questionable.

Keeping these limitations in mind, we see that this procedure allows Dertouzos and Pencavel to find the values of the parameters characterizing the utility function of unions. They find that unions weight the employment objective and that the estimated values of the parameters reject the hypothesis of the maximization of the wage rent or of the total wage bill.

Farber (1978) and Carruth and Oswald (1985) have adopted an identical procedure, but with the assumption of different objective functions. Farber studies the behavior of the United Mine Workers (United States) over the period 1948–1973, on the assumption that this union maximizes the expected utility of a member with median seniority. He estimates that this member's relative degree of risk aversion—equal by definition to $-wv''(w)/v'(w)$, if $v(w)$ is the indirect utility function and w the individual's wage—is on the order of 3. Carruth and Oswald analyze the behavior of unions in the coal and steel industries in the United Kingdom over the period 1950–1980. They too assume that the union maximizes the sum of the utilities of its members and find a relative degree of risk aversion on the order of 0.8. Such results lead us to reject the total wage bill or union rent as objectives, based on the risk neutrality of workers.

All of these results must, however, be interpreted with caution, for at least two reasons. First, the identification of the structural parameters of union preferences and labor demand relies on assumptions about the exogeneity of other variables of the model which are questionable. Second, union preferences are not being estimated directly. The estimates actually bear on both the functional form of the union objective and the mode of wage formation. The equations tested all assume that the union determines the wages

of its members unilaterally. In reality, wages are the object of bargaining, and this can perceptibly modify the form of the estimated equation.

3.2 Models of Collective Bargaining

With the aim of understanding the impact of unions on employment and wages, economists have elaborated models of collective bargaining which generally consider an environment with a firm that is able to make positive profits and a union that negotiates for all employees. The first model to represent collective bargaining between a firm and a union is the "monopoly union" model of Dunlop (1944), in which the union sets the wage unilaterally, knowing the labor demand of the firm. The "right-to-manage" model (Nickell and Andrews, 1983) generalizes this case by assuming that wages are bargained over, with the employer retaining the prerogative to hire and fire. This hypothesis is actually highly questionable, inasmuch as unions and employers may have an interest in negotiating over variables other than wages. On that basis, two models have been advanced: the model of weakly efficient bargaining over wages and employment and the model of strongly efficient bargaining, in which the protagonists can negotiate about as many variables as they judge necessary.

These models do make it possible to understand the impact of unions on employment and wage levels. They also show that unions have incentives to compress the wage spread among their members, the "insiders," but have an interest in reducing as much as possible the number of those who are not members, often called "outsiders."[3]

3.2.1 The Right-to-Manage Model: Negative Employment Effects

The right-to-manage model is a generalization of the union monopoly model, with the assumption that the firm always decides its own labor demand but that wages are bargained over.

The Negotiation

Here we consider a union composed of N identical workers; the union's objective is to maximize the expected utility of each of its members, knowing that if the firm's labor demand is less than the number of union members, the employer chooses whom to hire at random. When the wage paid by the firm is equal to w, an individual who is hired attains a level of utility equal to $v(w)$, and one who is not—an unemployed person—obtains $v(\overline{w})$. In this expression, $\overline{w}$ is an exogenous parameter designating the reservation wage, taken to be equivalent to the income of a person not employed by the firm. We can assume that a person not employed by the firm can always be hired in a perfectly competitive labor market offering wage $\overline{w}$ to every employee. Unless the opposite is explicitly stated, we will assume that the members of the union are risk averse ($v'' < 0$). Let L be employment; the union's objective is then written:

$$\mathcal{V}_s = \ell v(w) + (1 - \ell)v(\overline{w}) \quad \text{with} \quad \ell \equiv \min(1, L/N)$$

[3] The issue of the length of time spent at work is not presented here. Economic analysis shows that union power reduces the duration of work if leisure is a normal good, which constitutes the empirically pertinent hypothesis (see chapter 1). Interested readers may consult Booth and Ravaillon (1993), Contensou and Vranceanu (2000), and Cahuc and Zylberberg (2008).

The union is facing a firm which has a competitive advantage that allows it to make strictly positive profits: its market is protected by entry costs. The firm's profit takes the form $\Pi = R(L) - wL$, where $R(L)$ designates the revenue function (this function is such that $R' > 0$ and $R'' < 0$). Profit maximization gives the labor demand of the firm: it is defined by $L^d(w) \equiv R'^{-1}(w)$.

In the unfolding of the bargaining process, we follow the standard practice in the literature and assume that if no agreement is reached, the members of the union get the level of utility of a person not employed by the firm and the employer gets zero profit. If γ designates the power of the union, then the negotiated wage solves the following problem:

$$\max_w [\Pi(w)]^{1-\gamma} [v(w) - v(\overline{w})]^{\gamma} [L^d(w)]^{\gamma} \quad \text{with } \Pi(w) \equiv R[L^d(w)] - wL^d(w)$$

subject to:

$$L^d(w) \leq N \quad \text{and} \quad w \geq \overline{w}$$

For an interior solution, the maximization of the logarithm of the generalized Nash criterion gives the first-order condition:

$$\frac{\gamma}{L^d(w)} \frac{\mathrm{d}L^d(w)}{\mathrm{d}w} + \frac{\gamma v'(w)}{v(w) - v(\overline{w})} + \frac{(1-\gamma)}{\Pi(w)} \frac{\mathrm{d}\Pi(w)}{\mathrm{d}w} = 0$$

Let $\eta_w^L = -(w/L)(\mathrm{d}L/\mathrm{d}w)$ and $\eta_w^\pi = -(w/\Pi)(\mathrm{d}\Pi/\mathrm{d}w)$ be respectively the absolute values of the elasticity of employment and profit with respect to wages. In general, these quantities depend on the wage w. The first-order condition is then written:

$$\Phi(w, \overline{w}, \gamma) \equiv -\gamma \eta_w^L - (1-\gamma)\eta_w^\pi + \frac{\gamma w v'(w)}{v(w) - v(\overline{w})} = 0 \tag{7.11}$$

The second-order condition is satisfied when $\Phi_w < 0$. Furthermore, for every parameter x we have $\partial w/\partial x = -\Phi_x/\Phi_w$. In consequence, $\partial w/\partial x$ is of the sign of Φ_x. Hence, as:

$$\Phi_\gamma = -\eta_w^L + \eta_w^\pi + \frac{w v'(w)}{v(w) - v(\overline{w})} = \frac{\eta_w^\pi}{\gamma} > 0$$

we see that the wage is an *increasing* function of the bargaining power γ of the union. The marginal revenue of labor being equal to this wage, employment decreases with parameter γ. The same reasoning shows that the wage is an increasing function of the income $\overline{w}$ of a person not employed by the firm. The reader will also be able to verify that any increase in the absolute value of the wage elasticity of labor demand or profit entails a reduction in the wage.[4]

[4] In order to study the influence of the wage elasticity on labor demand and profits, it proves useful to assume that they are increasing functions of parameters z_L and z_π. In other words, we can posit $\eta_w^L = \eta_w^L(w, z_L)$ and $\eta_w^\pi = \eta_w^\pi(w, z_\pi)$ with $\partial \eta_w^L / \partial z_L > 0$ and $\partial \eta_w^\pi / \partial z_\pi > 0$ by definition. Then, the first-order condition (7.11) implies that function Φ depends negatively on parameters z_L and z_π.

The Negative Impact of Union Power on Employment

The first-order condition also allows us to express the difference between the gains made by a worker who is hired and the gains of a person not employed by the firm. Thus we have:

$$\frac{v(w)-v(\overline{w})}{wv'(w)}=\frac{\gamma}{\gamma\eta_w^L+(1-\gamma)\eta_w^\pi}\equiv\mu_s \tag{7.12}$$

This equation shows that those who are hired have a utility greater than that of the person not employed by the firm, given that $\gamma>0$. To be precise, variable μ_s is interpreted as a *markup* indicating the gap between the utility of a worker with a job in the firm and that of a person not employed by the firm. At the optimum of the bargaining problem, this markup increases with union power γ and decreases with the absolute values of the wage elasticity of labor demand and profit. In the limit case in which the union has all the bargaining power—the "monopoly union" model—the gap between the utility of an employee of the firm and that of a person not employed by the firm depends only on the wage elasticity of labor demand. When the union's bargaining power is null, workers hired and persons not employed by the firm have the same gains. Such a situation is generally described as *competitive*, inasmuch as unionized employees get no "rent" with respect to other workers. The negotiated wage then equals the outside wage $\overline{w}$.

If the revenue function of the firm is homogeneous of degree $\alpha\in(0,1)$, then we have $\eta_w^L=1/(1-\alpha)$, $\eta_w^\pi=\alpha/(1-\alpha)$, and $\mu_s=\gamma(1-\alpha)/(\gamma(1-\alpha)+\alpha)$. In this case, shocks to productivity or the firm's selling price do not affect the wage and lead only to employment adjustments.

In figure 7.9 we represent the solution of the right-to-manage model. Note that an indifference curve for the union, defined by the equation $[v(w)-v(\overline{w})]\,L=cst$, has a negative slope in the plane (L,w) when $L\leq N$. Differentiating this equation, one gets:[5]

$$\left.\frac{\mathrm{d}w}{\mathrm{d}L}\right|_{\mathcal{V}_s=cst}=\frac{-[v(w)-v(\overline{w})]}{Lv'(w)}\leq 0$$

$$\left.\frac{\mathrm{d}^2w}{\mathrm{d}L^2}\right|_{\mathcal{V}_s=cst}=\frac{[v(w)-v(\overline{w})]}{L^2[v'(w)]^2}\left\{2v'(w)-\frac{v''(w)\,[v(w)-v(\overline{w})]}{v'(w)}\right\}\geq 0$$

The indifference curves are thus decreasing and convex. Moreover, they have a horizontal asymptote at the point $w=\overline{w}$ in the plane (w,L). We can also show that an

[5] To obtain $\left.\frac{\mathrm{d}^2w}{\mathrm{d}L^2}\right|_{\mathcal{V}_s=cst}$, one must derive $\left.\frac{\mathrm{d}w}{\mathrm{d}L}\right|_{\mathcal{V}_s=cst}$ with respect to w. We get:

$$\left.\frac{\mathrm{d}^2w}{\mathrm{d}L^2}\right|_{\mathcal{V}_s=cst}=\frac{1}{L^2[v'(w)]^2}\left[-Lv''(w)v'(w)\frac{\mathrm{d}w}{\mathrm{d}L}+Lv''(w)\,[v(w)-v(\overline{w})]\frac{\mathrm{d}w}{\mathrm{d}L}-v'(w)\,[v(w)-v(\overline{w})]\right]$$

Then substituting $\frac{\mathrm{d}w}{\mathrm{d}L}$ by its value $\left.\frac{\mathrm{d}w}{\mathrm{d}L}\right|_{\mathcal{V}_s=cst}$ gives the result.

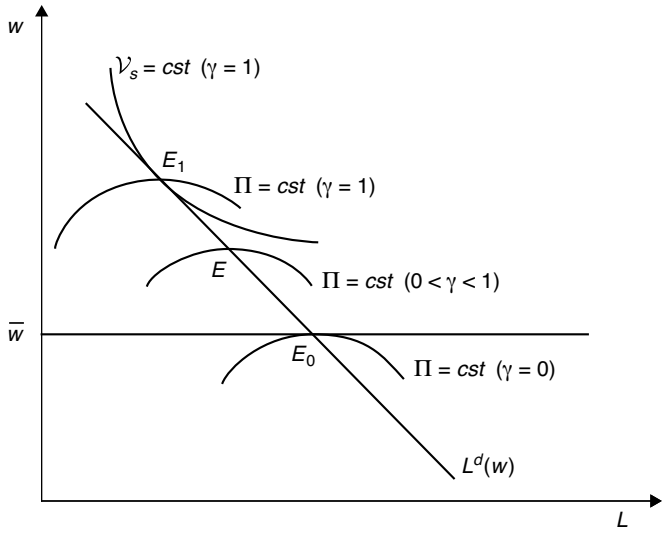


FIGURE 7.9
The right-to-manage model.

isoprofit curve, defined by the equation $R(L) - wL = cst$, reaches a maximum on the labor demand curve. Differentiating this equation, we get:

$$\left.\frac{\mathrm{d}w}{\mathrm{d}L}\right|_{\Pi=cst} = \frac{R'(L) - w}{L}$$
$$\left.\frac{\mathrm{d}^2w}{\mathrm{d}L^2}\right|_{\Pi=cst} = \frac{LR''(L) - 2\,[R'(L) - w]}{L^2}$$

The first equation implies that the isoprofit curves have a horizontal tangent on the labor demand curve, where $R'(L) = w$. Moreover, the second equation implies, together with the concavity of $R(L)$, that the isoprofit curves are concave at the point where they cross the labor demand curve, which means that the isoprofit curves reach a maximum on the labor demand curve.

In the right-to-manage model, the solutions lie on the labor demand. If the union's bargaining power is zero, the wage is equal to $\overline{w}$ and the isoprofit curve is tangential to the union's indifference curve at point E_0 (the first-order conditions of profit maximization entail that the isoprofit curves have a zero slope when they cross labor demand). In the other extreme situation, in which the union disposes of all the bargaining power, the solution lies at point E_1, where the indifference curve of the union is tangential to the labor demand. In all cases lying in between $(0 < \gamma < 1)$ the solution lies at point E on the portion of the labor demand delimited by points E_0 and E_1.

The monopoly union model, and more generally the right-to-manage model, come to the conclusion that the bargaining power of unions lowers employment. They are not, however, totally satisfactory because the union and the employer agree, when $\gamma > 0$, on a Pareto inefficient contract. At every point $E \neq E_0$, figure 7.9 shows that the indifference curves and the isoprofit curves are not tangent. Starting from point E, the employer

and the union could thus agree on an employment-wage pairing that would raise the level of satisfaction of at least one of them. In this regard, Leontief (1946) pointed out that it was possible to reach Pareto efficient allocations by bargaining over wages and employment.

3.2.2 Weakly Efficient Contracts: Overemployment

The outcome of collective bargaining is not usually a mere agreement about wage. The number of hours to be worked, working conditions, employment, and union representation are also privileged as objects of negotiation. The right-to-manage model is thus seen to lie relatively far from reality. It will be instructive to begin by examining the case in which the bargaining is about just two variables, wages and employment, then extend the analysis to larger numbers of variables. Assuming that bargaining does not concern wages alone leads to very different predictions from those arrived at with the right-to-manage model. In particular, increases in union power are not necessarily bad for employment if a sufficient number of topics are bargained over. The negotiation can increase the number of jobs with respect to the competitive situation when the firm and the union bargain over employment and wage only. However, when they bargain over a sufficiently large set of variables, the employment level is the same as in the competitive situation.

If bargaining is not solely about wages, the other variables to be agreed on must, directly or indirectly, have an influence on the level of employment. For this reason, MacDonald and Solow (1981) proposed to represent collective bargaining by a negotiation over employment and wages simultaneously. In this case, the bargaining problem is written:

$$\max_{\{w,L\}} [R(L) - wL]^{1-\gamma} [v(w) - v(\overline{w})]^{\gamma} L^{\gamma}$$

subject to:

$$0 \leq L \leq N \quad \text{and} \quad w \geq \overline{w}$$

For the interior solutions, differentiating the Nash criterion with respect to L and w, the first-order conditions imply:

$$(1-\gamma)\frac{R'(L) - w}{R(L) - wL} + \frac{\gamma}{L} = 0$$

$$-(1-\gamma)\frac{L}{R(L) - wL} + \gamma\frac{v'(w)}{v(w) - v(\overline{w})} = 0$$

Eliminating parameter γ between these two relations, we find the equation of the *contract curve*. It is written:

$$w - R'(L) = \frac{v(w) - v(\overline{w})}{v'(w)}$$

This curve represents the locus of the tangency points between the curves of isoprofit and isoutility. Hence, bargaining over wages and employment arrives at a Pareto optimal contract between the union and the firm. Differentiating the equation of the contract curve gives:

$$\frac{dw}{dL} = \frac{v'(w)R''(L)}{v''(w)[w - R'(L)]}$$

We see that the contract curve has a positive slope if workers are risk averse ($v'' < 0$). This situation is represented in figure 7.10, where we see that wages *and* employment increase with the union's bargaining power, since the union uses its room for maneuver both to protect workers against the risk of unemployment and to increase their remuneration. When $\gamma = 0$, the negotiated wage is equal to the reservation wage, and employment reaches its "competitive" value as defined by the equality between marginal revenue and $\overline{w}$ (this equality is often called the *productive efficiency* condition). This solution corresponds to point E_0 in figure 7.10. The presence of the union thus entails a level of employment *higher* than that which would prevail in a competitive situation. In other words, there is overemployment. If workers are risk neutral ($v'' = 0$), the contract curve is a vertical line in plane (L, w), having the competitive level of employment as its abscissa. This means that it is optimal, in this situation, to share the rent of the firm in the form of wage and employment increases instead of wage increases only, as is the case in the right-to-manage model. Employment decreases with the bargaining power of the union only if workers are highly risk tolerant.

The model with bargaining over wages and employment entails a level of employment that equalizes the marginal revenue and the reservation wage $\overline{w}$ only when workers are risk neutral. If workers are risk averse, this type of bargaining yields

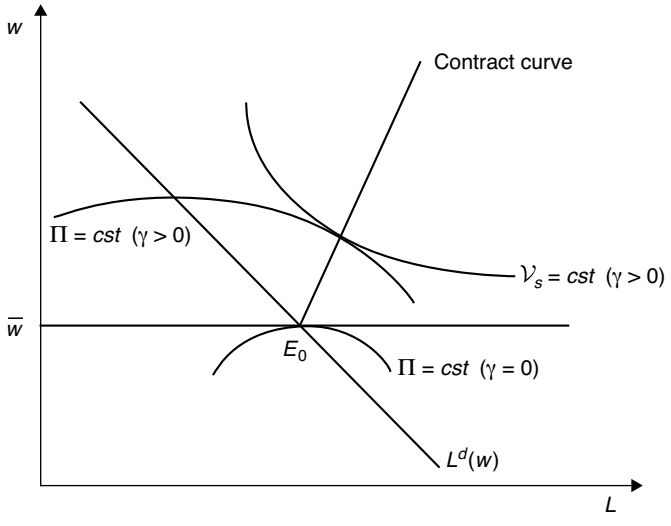


FIGURE 7.10
The model of bargaining over wages and employment.

overemployment, since the marginal revenue is less than the reservation wage. In other words, bargaining over wages and employment generally does not entail productive efficiency. For this reason, bargaining over employment and wages is frequently described as "weakly efficient."

3.2.3 Strongly Efficient Contracts: No Employment Effect

A priori, nothing prevents the union and the firm from coming to an agreement over certain variables other than employment and wages, if they have a mutual interest in doing so. Hence we assume that bargaining also extends to unemployment insurance benefits. We then see that the solution to the bargaining always arrives at an equalization of the marginal revenue from labor and the reservation wage.

The Indifference Principle

Let b be the unemployment benefit paid to each union member not employed by the firm. Assuming that a person not employed by the firm can receive income $\overline{w}$ under all circumstances, she then attains a level of utility equal to $v(\overline{w}+b)$. In this static framework, such an unemployment benefit can also be interpreted as a severance payment given to workers forced to leave the firm. In this case, the number N of workers bargaining with the firm is equal to the number of employees present in the firm at the beginning of the period taken into consideration (see Booth, 1995b, for example). Let us consider the situation in which bargaining extends directly to wages w and unemployment benefit b, with the firm preserving the right to manage. Let $\mathcal{C}=(w,b),\ w\geq\overline{w}+b$ be a contract of this type. We will show that if workers are risk averse ($v''<0$), this contract is Pareto dominated by a contract $\hat{\mathcal{C}}=(\widehat{w},\widehat{b})$, giving the same utility to the jobless and to employees, whatever the level L of employment. To that end, let us define the components of $\hat{\mathcal{C}}$ in the following manner:

$$\widehat{w}=\ell w+(1-\ell)(\overline{w}+b)\quad\text{with }\ell=\min(L/N,1)\text{ and }\widehat{b}=\widehat{w}-\overline{w}$$

By construction, contract $\hat{\mathcal{C}}$ satisfies $v(\widehat{w})=v(\widehat{b}+\overline{w})$. Moreover, risk aversion entails:

$$v(\widehat{w})=v[\ell w+(1-\ell)(\overline{w}+b)]\geq\ell v(w)+(1-\ell)v(\overline{w}+b)$$

Let $\mathcal{V}_s$ and $\hat{\mathcal{V}}_s$ be the expected utility of a union member with contract $\mathcal{C}$ and contract $\hat{\mathcal{C}}$ respectively. We then have:

$$\hat{\mathcal{V}}_s=\ell v(\widehat{w})+(1-\ell)v(\overline{w}+\widehat{b})=v(\widehat{w})\geq\mathcal{V}_s$$

Thus the union always prefers contract $\hat{\mathcal{C}}$ to contract $\mathcal{C}$. As well, it is easy to verify that the firm is indifferent. Employment L being the same in both types of contract, revenue $R(L)$ is thus identical. A simple calculation shows that the total wage bill does not change either. It is given by:

$$\begin{aligned}\widehat{w}L+\widehat{b}(N-L)&=\widehat{w}L+(\widehat{w}-\overline{w})(N-L)=\widehat{w}N-\overline{w}(N-L)\\&=[Lw+(N-L)(\overline{w}-b)]-\overline{w}(N-L)=wL+b(N-L)\end{aligned}$$

The passage from contract $\mathcal{C}$ to contract $\hat{\mathcal{C}}$ thus involves an improvement in the Pareto sense. In consequence, optimal contracts will respect the "indifference principle" $v(w) = v(\overline{w} + b)$. Workers are then perfectly insured against the risks of not being employed by the firm, without that affecting the value of the firm's profit. Note that this conclusion does not depend on the employer's attitude to risk, since revenues and total wage bills are strictly identical for the two types of contract $\mathcal{C}$ and $\hat{\mathcal{C}}$.

The Optimal Contract

On the basis of the foregoing, when unemployment benefits (or severance payments) are included in the negotiated variables, it is enough to study contracts of the form $\mathcal{C} = (b + \overline{w}, b)$. Profit is then written:

$$\Pi = R(L) - \overline{w}L - bN$$

We see that from the point of view of the firm, it is as if it were paying wage b to all members of the union and compensating those who were actually working by offering them a supplement $\overline{w}$. Profit maximization thus defines a labor demand L^* independent of the unemployment benefit b. The firm simply makes marginal revenue equal to the reservation wage: $\overline{w} = R'(L^*)$. In sum, the negotiation will only concern the unemployment benefits b. We assume that in case of failure to reach agreement the firm does not pay these benefits. Moreover, its profit is zero in this case because it is assumed that nobody is working. The utility of each worker then being equal to $v(\overline{w})$, the contribution $\mathcal{V}_s - v(\overline{w})$ of the union to the Nash problem is equal to $v(b + \overline{w}) - v(\overline{w})$. The reader will note that the union's objective is independent of the level of employment, since all workers are insured against the risk of unemployment. If we assume $L^* < N$, the bargaining problem takes the form:

$$\max_b [R(L^*) - \overline{w}L^* - bN]^{1-\gamma} [v(\overline{w} + b) - v(\overline{w})]^{\gamma}$$

The optimal level of unemployment benefits is then defined by:

$$\frac{v(\overline{w} + b) - v(\overline{w})}{v'(\overline{w} + b)} = \frac{\gamma}{1 - \gamma} \frac{[R(L^*) - \overline{w}L^* - bN]}{N} \quad \text{with } w = \overline{w} + b \quad \text{and } R'(L^*) = \overline{w}$$

The possibility of bargaining as well over the amount of the unemployment benefits thus has the effect of making the level of employment equal to its *competitive* value. The union members obtain a portion of the firm's profit, which increases with their bargaining power, without that causing reduced production or employment. In this context, the contract curve, which is the locus of the tangency points between the union's isoutility curves and the firm's isoprofit curves, is a vertical line defined by the relation $R'(L) = \overline{w}$. For an optimal contract, the utility function of the union is written $\mathcal{V}_s = v(w)$, and the expression of the firm's profit is $\Pi = R(L) - \overline{w}L - (w - \overline{w})N$. We thus get:

$$\left.\frac{\mathrm{d}w}{\mathrm{d}L}\right|_{\mathcal{V}_s = cst} = 0 \quad \text{and} \quad \left.\frac{\mathrm{d}w}{\mathrm{d}L}\right|_{\Pi = cst} = \frac{R'(L) - \overline{w}}{N}$$

The graphic representation of the strongly efficient bargaining model is given in figure 7.11. In this model the opportunity to bargain over unemployment benefits makes it possible to insure workers against the risk of unemployment. Bargaining of this kind, which reconciles productive efficiency and Pareto efficiency between the union and the firm, is called strongly efficient in order to distinguish it from bargaining limited to employment and wages, in which it is impossible to arrive at productive efficiency when workers are risk averse.

3.2.4 Is Bargaining Efficient?

We have just seen that collective bargaining leads to efficient contracts if unions and firms do actually bargain over wages, employment, and perhaps other variables like unemployment insurance benefits and severance payments. Manning (1987), Espinosa and Rhee (1989), and Strand (1989) have suggested that a contract covering employment and wages is more difficult to negotiate than a contract simply covering wages, inasmuch as an efficient choice of the level of employment of each type of manpower requires a thorough knowledge of the firm and must prescribe contingent contracts when the environment is uncertain, as the analysis developed in chapter 6 shows. On the other hand, bargaining over unemployment benefits or severance payments raises incentive problems that may prevent the achievement of efficient contracts.

Negotiations Over Employment

The model of Manning (1987) starts with the principle that the power of the union varies according to which variables are being bargained over. It nonetheless adopts the same sequence of decisions as that of the right-to-manage model, that is, the firm and the union agree at the outset on the amount w of wages. Knowing that, they launch a

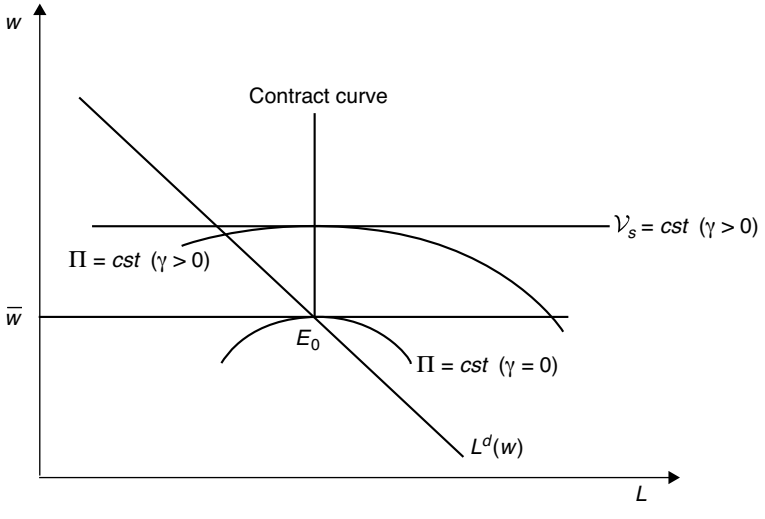


Figure 7.11
The strongly efficient bargaining model.

bargaining process over the level of employment, the outcome of which corresponds to the solution of the following Nash problem:

$$\max_{L} [R(L) - wL]^{1-\gamma_L} [v(w) - v(\overline{w})]^{\gamma_L} L^{\gamma_L}$$

subject to:

$$0 \leq L \leq N$$

Parameter $\gamma_L \in [0, 1]$ designates the power of the union during the bargaining over *employment*. The solution of this problem defines labor demand, or $L = \hat{L}(\gamma_L, \overline{w}, w)$, a function of wage w negotiated beforehand, bargaining power γ_L, and reservation wage $\overline{w}$. Bargaining over wages takes this labor demand $\hat{L}$ into account and is represented by another Nash problem:

$$\max_{w} [R(L) - wL]^{1-\gamma} [v(w) - v(\overline{w})]^{\gamma} (L)^{\gamma}$$

subject to:

$$L = \hat{L}(\gamma_L, \overline{w}, w) \quad \text{and} \quad w \geq \overline{w}$$

Parameter $\gamma \in [0, 1]$ designates the power of the union during the bargaining over *wages*. The two-stage solution of this bargaining process corresponds to that of the right-to-manage model if $\gamma_L = 0$, and to that of the weakly efficient contract model when $\gamma_L = \gamma$. In all other cases, the solution is not found on either labor demand or the contract curve.

Manning (1987) justifies this description of the unfolding of negotiations by arguing that wages, in general, are determined before employment is, but that does not mean that unions never play a part in determining the level of employment. He also points out that bargaining over wages takes place at a more centralized level than bargaining over employment. The latter is often informal in nature and takes place primarily at the level of the firm or the plant. These two reasons can indeed justify a representation of bargaining by a two-stage process, as well as different bargaining powers according to whether the bargaining is taking place over wages or employment. The model of Manning offers the advantage of showing that bargaining over employment and wages does not necessarily conclude with an efficient contract. It is not, however, completely satisfactory, for the two-stage representation, strictly separating bargaining over employment from bargaining over wages, has no precise theoretical foundation. It is, moreover, difficult to interpret the difference between bargaining power over employment and bargaining power over wages, on the basis of a noncooperative game.

Espinosa and Rhee (1989) and Strand (1989), starting from a different perspective, arrive at a conclusion close to that of Manning (1987). They consider a repeated game with an infinite horizon, in which a union and a firm bargain over wages at predetermined dates. In this framework, the decision to bargain over employment corresponds to a cooperative strategy within a strategic structure of the prisoner's dilemma type. The

firm has an interest in bargaining over employment and hiring workers whose marginal productivity is lower than their wage only if the union agrees to lower wages. But once wage concessions have been extracted, the firm has an interest in renouncing its implicit undertaking regarding employment by equalizing the marginal productivity of labor to wages. Espinosa and Rhee (1989) and Strand (1989) exploit the properties of repeated games in order to show that bargaining will only implicitly cover employment if the firm has a sufficiently weak preference for the present. They further prove the existence of values of the firm's discount rate for which the solution of the bargaining lies between the labor demand curve and the contract curve.

These contributions suggest, overall, that the right-to-manage model and the model of bargaining over wages and employment represent limit cases of the same model.

Negotiations Over Unemployment Benefits or Severance Payments

Negotiations over unemployment benefits or severance payments raise incentive problems that may constitute a barrier to obtaining efficient contracts. The strongly efficient bargaining model just presented does indeed come to the conclusion that the jobless, or workers who are fired, have a level of welfare *identical* to that of workers who are employed. The majority of empirical studies (see, for example, Atkinson and Micklewright, 1991, and Clark and Oswald, 1994) find that unemployment benefits are far from offering perfect insurance. The situation of those who do have a job is preferable to that of the jobless. Imperfect unemployment insurance may come from a moral hazard problem (see chapter 6). Kiander (1993) shows that it may be optimal to insure workers partially if excessively high unemployment benefits reduce the job search effort of the unemployed, and if checking on this effort proves too costly. Kiander's analysis applies to a representative union in a position to set unemployment benefits at a centralized level, as in Sweden, for example. The moral hazard problem is even clearer at the local level. It lets us understand why unions do not generally negotiate supplementary unemployment insurance at the level of firms in their labor pool. Benefits of this kind would risk attracting a large number of unemployed persons, which would cut back the profits of firms and the wages of workers in that labor pool. Layard et al. (1991, p. 95) have in fact observed that, with very few exceptions, collective agreements signed at the level of the firm do not make provision for unemployment benefits.

3.2.5 Wage Dispersion

Collective bargaining generally covers workers whose productive characteristics are heterogeneous. Workers with different skill levels are often represented by the same union. Collective bargaining models suggest a tendency to reduce the spread of wages, as compared to a situation in which workers are remunerated at their marginal productivity.

Let us consider a firm with two types of worker indexed by $i = 1, 2$, whose revenue is given by $R(L_1, L_2)$, where L_i designates the number of employed workers of type i, and R is a concave function increasing with respect to each of its arguments. Workers of type 1 have higher productivity than workers of type 2, which leads to a higher reservation wage for workers of type 1. By hypothesis, we thus have $\overline{w}_1 > \overline{w}_2$. Assuming

that the firm's employment pool comprises N_i workers of type i, a utilitarian union representing all the workers has as its objective:

$$\mathcal{V}_s = \sum_{i=1}^{2} L_i v(w_i) + (N_i - L_i)\, v(\overline{w}_i + b_i), \quad L_i \leq N_i$$

In this expression, b_i designates the amount of unemployment benefits paid to a worker of type i by the firm. We assume that the bargaining is strongly efficient. That means that it covers unemployment benefits, as well as employment and wages. As in the preceding model with a homogeneous workforce, employees' risk aversion always entails the indifference principle. Thus, the optimal contract necessarily satisfies $w_i = \overline{w}_i + b_i$, $i = 1, 2$. The Nash problem is then written:

$$\max_{\{b_1, b_2, L_1, L_2\}} \left[R(L_1, L_2) - \sum_{i=1}^{2} (\overline{w}_i L_i + b_i N_i) \right]^{1-\gamma} \left[\sum_{i=1}^{2} N_i \left[v(\overline{w}_i + b_i) - v(\overline{w}_i) \right] \right]^{\gamma}$$

subject to:

$$0 \leq L_i \leq N_i, \quad i = 1, 2$$

The first-order conditions are found by setting to zero the derivatives with respect to L_i and b_i of the logarithm of the Nash criterion. For the interior solutions, we thus get:

$$\frac{\partial R(L_1, L_2)}{\partial L_i} = \overline{w}_i, \quad i = 1, 2 \tag{7.13}$$

$$v'(\overline{w}_i + b_i) = \frac{(1-\gamma)\left[\sum_{i=1}^{2} N_i \left[v(\overline{w}_i + b_i) - v(\overline{w}_i) \right]\right]}{\gamma \left[R(L_1, L_2) - \sum_{i=1}^{2} (\overline{w}_i L_i + b_i N_i) \right]}, \quad i = 1, 2 \tag{7.14}$$

Equality (7.13) is a consequence of the hypothesis of strongly efficient bargaining. It indicates that the marginal productivity of each type of worker is equal to his reservation wage: the condition of productive efficiency is thus satisfied for each skill category. The right-hand side of equation (7.14) is a quantity independent of index i, so the wages $w_i = \overline{w}_i + b_i$ of the two types of worker are identical. Collective bargaining thus leads to the same wage level for the two types of worker, even though their productivities are different. This result is due to the properties of the utilitarian criterion of the union. All the workers being identical in terms of preference, and all having the same weight in the union's objective, the concavity of function $v(.)$ entails that the union always prefers a contract offering identical wages. Formally, this property can be proved with inequality:

$$\frac{N_1}{N_1 + N_2} v(w_1) + \frac{N_2}{N_1 + N_2} v(w_2) \leq v\left[\frac{N_1}{N_1 + N_2} w_1 + \frac{N_2}{N_1 + N_2} w_2 \right]$$

According to this inequality, given a contract offering wage w_i to N_i workers of type i, the union's contribution to the Nash criterion will always be greater with a contract offering

the same wage, equal to $\left[\frac{N_1}{N_1+N_2}w_1 + \frac{N_2}{N_1+N_2}w_2\right]$, to all the workers. Collective bargaining thus reduces wage dispersion with respect to a competitive situation in which each worker would receive his reservation wage $\overline{w}_i$. It should be noted that the equalization of the wages of different types of worker is obtained under very restrictive hypotheses. In particular, the union has to attribute the same importance to the different categories of worker, and the bargaining must be strongly or weakly efficient (it can be verified that bargaining over employment and wages also ends in equalized wages). Conversely, bargaining over wages alone generally arrives at a different result, since the wage of each manpower category depends on the labor demand elasticity of that particular category. This model nevertheless illustrates the fact that collective bargaining potentially has the effect of reducing the spread of wages.

3.2.6 Insiders and Outsiders

To this point we have assumed that the union represented all the workers in the labor pool of the firm in question and that they have the same preferences and the same weight in the objective of the union. At bargaining time, however, workers do not all have the same status. Some are unemployed, while others have a job. Actually, the unemployed are generally excluded from the bargaining process. They are "outsiders," with no power to influence the decisions of firms. Conversely, the "insiders," those in employment, can defend their interests and exploit positional advantages without having to worry about the effects on the outsiders. Moreover, all workers do not necessarily have the same preferences when it comes to wages and employment. When the last-in-first-out rule governs firing, the wage earners with the most seniority, who have little likelihood of being let go if the firm does reduce its workforce, have less incentive to take employment into account than recently hired wage earners do. If the union functions democratically, the median voter, who generally has a relatively high degree of seniority, may on that account have a weak preference for employment and little concern for recently hired wage earners, who are regarded as "outsiders" excluded de facto from the objectives of the union. Does this exclusion of outsiders from the bargaining explain their exclusion from employment? The insiders–outsiders model sheds light on this issue.

A Simple Model

In the insiders–outsiders model, it is important to specify how, and after how long, a person recently hired—an "entrant"—accedes to the status of insider. We sidestep the complications linked to this aspect of the problem by taking the view that the firm and the insiders negotiate in a timeframe limited to a single period. That being so, the future of entrants plays no part in the choice criteria of the insiders, since at the end of this period entrants do not become insiders. More precisely, we assume that the firm disposes of a stock L_0 of insiders and that it must decide on the number $L_I \leq L_0$ that it wants to retain as well as the number $L_E \geq 0$ of outsiders that it wants to hire. To simplify, we take it that insiders and entrants are perfectly substitutable in production. The firm's revenue is then written $R(L_I + L_E)$. Nonetheless, we assume that it is impossible to replace insiders with outsiders, an impossibility explained by, among other things, hiring and firing costs (which for simplicity do not appear in the model; see Lindbeck and Snower, 1988, for a more thorough analysis).

If bargaining over unemployment benefits is rare, bargaining over severance payments, on the contrary, proves to be frequent (see Layard et al., 1991, p. 95, and Hartog and Theeuwes, 1993). Hence it is important to take this bargaining item into account in models in which insiders are explicitly distinguished from outsiders. To take these characteristics of the wage relationship into account, we assume that the insiders negotiate their own wage w, severance payments b_L, and a wage w_E for entrants. In fact, in many countries, firms often use temporary workers, whose status is much less favored than that of insiders. An insider who is fired thus obtains a utility equal to $v(\overline{w} + b_L)$. Severance payments constitute a means of insuring workers against the risk of losing their jobs, and, just as in the case where bargaining covered unemployment insurance premiums—see section 3.2.3—the indifference principle applies. In other words, contracts that insure insiders perfectly against the risk of losing their jobs are dominant according to the Pareto criterion. For given (L_E, L_I, w_E) and for every contract $\mathcal{C}_L = (w, b_L)$, it is indeed possible to link a contract $\hat{\mathcal{C}}_L = (\widehat{w}, \widehat{b}_L)$ defined by:

$$\widehat{w} = \ell w + (1 - \ell)(\overline{w} + b_L) \quad \text{with} \quad \widehat{b}_L = \widehat{w} - \overline{w} \quad \text{and} \quad \ell = \min(L_I/L_0, 1)$$

When insiders are risk averse, the same proof as the one in section 3.2.3 would show that contract $\hat{\mathcal{C}}_L$ is strictly preferred to contract $\mathcal{C}_L$ by the insiders, while the firm is indifferent between these two contracts. This result leads us to consider only contracts for which $w = \overline{w} + b_L$; the insiders are then perfectly insured against the risk of job loss. As in the case of a strike, they attain a level of utility $v(\overline{w})$, and their contribution $\mathcal{V}_I - v(\overline{w})$ to the Nash problem is equal to $v(\overline{w} + b_L) - v(\overline{w})$. Symmetrically, the contribution of the employer to the Nash problem is equal to his profit:

$$\Pi = R(L_I + L_E) - w_E L_E - \overline{w} L_I - b_L L_0 \tag{7.15}$$

Since the wage w_E of the entrants has a negative effect on the firm's profit and has no weight in the objective of the insiders, maximization of the Nash criterion dictates that this wage be set at the lowest possible level; so we will always have $w_E = \overline{w}$. When the firm retains the right to manage, the expression (7.15) of profit shows that labor demand necessarily satisfies:

$$R'(L_I + L_E) = \overline{w} \tag{7.16}$$

Discrimination or Unemployment?

Equation (7.16) shows that total employment is equal to its *competitive* value. All the firm does is decide on its composition (hires of outsiders or fires of employees in place) according to the value of the initial stock L_0 of insiders. More precisely, if L_u designates the competitive level of employment, defined by $R'(L_u) \equiv \overline{w}$, the firm's labor demand takes the following form:

$$\begin{aligned} L_I &= L_u \quad \text{and} \quad L_E = 0 \quad \text{if } L_u \leq L_0 \\ L_I &= L_0 \quad \text{and} \quad L_E = L_u - L_0 \quad \text{if } L_u \geq L_0 \end{aligned}$$

Bargaining between the insiders and the firm now covers only the amount b_L of the severance payment. The optimal value of the latter corresponds to the solution of

the Nash problem:

$$\max_{b_L} [\Pi(b_L)]^{1-\gamma} [v(\overline{w} + b_L) - v(\overline{w})]^{\gamma} \quad \text{with} \quad \Pi(b_L) \equiv R(L_u) - \overline{w}L_u - b_L L_0$$

Here, employment is always equal to its competitive level, whatever the initial number of insiders taking part in the bargaining. Being perfectly insured against the risk of job loss, they use their bargaining power to obtain the highest possible wage ($\overline{w} + b_L$). It is easy to verify that the optimal level of b_L is increasing with γ. Moreover, insiders who are few in number have an interest in seeing the firm hire workers at the reservation wage in order to increase profit and thus indirectly their wages. In this sense, the insiders exploit the entrants, profiting from their bargaining power to extract a portion of the profits realized through the labor of the entrants. In other words, the insiders have no interest in opposing the hiring of outsiders as long as that is profitable for the firm, since what is profitable for the firm is profitable for them as well. These observations show that the opposition between insiders and outsiders, as Fehr (1990) pointed out, induces discrimination rather than unemployment. Certain workers capture a portion of rent thanks to the acquisition of specific human capital, for example, or the existence of firing costs, or the costs of looking for manpower. These workers have an interest in exploiting this situation by tilting the partition of the value added to their own advantage.

Evidently this description of the segmentation of the workforce is relevant only if the insiders are able to keep the entrants in a situation less favorable than their own over the long run. Legal constraints that impede recourse to temporary labor and subcontracting, and the power that entrants may gradually acquire, may set limits to the discrimination imposed by the insiders. Formally, we could incorporate a limit on the possibility of discrimination by supposing that the firm is constrained to pay the same wage to all its employees. We then return to a model in which the set of possible contracts has been voluntarily curtailed, and the power of insiders would have a negative effect on the level of employment.

3.3 Negotiations and Investment

In order to grasp with precision the impact of unions on investment, it will be helpful to introduce capital into our models of collective bargaining. This approach allows us to see that the interaction between wage bargaining and investment decisions can entail inefficiencies caused by the *irreversibility* of investments. We will see that if the union is able to *renegotiate* agreements already reached, the level of investment is generally suboptimal, and that empirical research does indeed highlight a negative impact of unions on investment.

The traditional models of labor demand presented in chapter 2 suggested that unions have an ambiguous effect on investment. By raising wages, the union tends to favor the substitution of capital for labor, which increases investment. But upward pressure on wages also exerts scale effects, which work in the opposite direction. Bargaining influences investment in yet other ways related to the incompleteness of contracts. The fact is that once installed, equipment generally cannot be modified without cost, and if it is not utilized, firms risk suffering substantial losses. This characteristic of equipment entails that firms have an incentive to invest less if bargaining over wages can be

started up at any time, for once the investment has been made, employees are tempted to demand a new round of bargaining in order to benefit from the improved productivity induced by the increase in capital stock. Conversely, if renegotiation is impossible, firms do invest more, and all agents benefit from the extra investment (Grout, 1984). This is the "holdup" problem. It crops up when collective agreements lead to incomplete contracts that can be renegotiated.

3.3.1 Contracts Without Renegotiation

To highlight the holdup problem, we consider a firm whose revenue function $R(K,L)$ is strictly concave and strictly increasing with capital K and employment L. To concentrate on the choice of level of investment, we assume that employment is given. This hypothesis is no doubt restrictive, but the results derived do not differ in substance from those obtained with a labor demand dependent on wages (see Anderson and Devereux, 1988). We assume as well that the firm chooses its capital stock unilaterally. In consequence, only wages w are negotiated. If r designates the user cost of capital, the firm's profit is written:

$$\Pi = R(K,L) - wL - rK \tag{7.17}$$

Employment L being fixed, we need not consider the possibility of bargaining over unemployment insurance or severance payments. Hence an employee obtains a level of utility $v(w)$ if she works and $v(\overline{w})$ in case of failure to agree. Since L is a constant, we can neglect this variable in the union's contribution to the Nash problem, which is thus simply equal to $v(w) - v(\overline{w})$. The firm's contribution to the Nash problem depends on the possibility of wages being renegotiated. If the union can undertake in a credible manner not to demand new wage negotiations once the investment has been made, the firm takes wages as *given* in making its decisions about equipment. Formally that amounts to supposing that investment decisions are made *after* wage bargaining. Conversely, if it is not possible for the union to commit itself in a credible manner to the wage, then we can regard investment decisions as being made *before* wage bargaining. Wages then become a function of the capital stock, and the firm takes this linkage into account when the time comes to choose its volume of equipment (see Grout, 1984; van der Ploeg, 1987; Anderson and Devereux, 1988; Devereux and Lockwood, 1991).

If the union can undertake credibly not to reopen wage negotiations, the firm does not run the risk of making an investment that could be immobilized by a strike, leading to losses. Under those conditions, its losses are zero, and its contribution to the Nash problem is identical to the profit given by relation (7.17). The optimal level of capital K^* is then obtained by maximizing profit at a given wage; it is defined by equation:

$$R_K(K^*,L) = r \tag{7.18}$$

The level of employment L being fixed, we observe that K^* does not depend on the value of the negotiated wage. Overall, the bargaining problem is written:

$$\max_w \left[v(w) - v(\overline{w})\right]^{\gamma} \left[R(K^*,L) - wL - rK^*\right]^{1-\gamma}$$

If w^* designates the solution of this problem, the pair (w^*, K^*) represents a Pareto optimum for the firm and the union. K^* being independent of wages, the pair (w^*, K^*) does in fact correspond to the solution of the following Nash problem too:

$$\max_{\{w,K\}} [v(w) - v(\overline{w})]^{\gamma} [R(K,L) - wL - rK]^{1-\gamma}$$

The pair (w^*, K^*) is thus indeed a Pareto optimum.

3.3.2 Contracts with Renegotiation

Let us now suppose that wages can be renegotiated after the employer has installed new equipment. If an investment K is made before wage bargaining and if the union cannot credibly undertake to stick to the negotiated wage, the firm will suffer a loss equal to $-rK$ if there is a strike. For given K the bargaining problem is then written as follows:

$$\max_{w} [v(w) - v(\overline{w})]^{\gamma} [R(K,L) - wL]^{1-\gamma} \tag{7.19}$$

Let $w(K)$ be the solution of this problem; the firm takes this relation into account in deciding its investment. The optimal level $\hat{K}$ of capital is then found by maximizing the firm's profit, which now takes the form:

$$\Pi = R(K,L) - w(K)L - rK$$

Setting the first derivative of this expression to zero with respect to K, we get:

$$R_K(\hat{K}, L) = w'(\hat{K})L + r \tag{7.20}$$

Scrutiny of relations (7.18) and (7.20) indicates that the comparison between levels of investment K^* and $\hat{K}$ depends on the sign of the derivative of function $w(K)$. This sign may be found easily with the help of the Nash criterion that comes into problem (7.19). It is written in logarithmic form:

$$\Phi(w,K) = \gamma \log [v(w) - v(\overline{w})] + (1-\gamma) \log[R(K,L) - wL] \tag{7.21}$$

Function $w(K)$ is defined by the first-order condition:

$$\Phi_w[w(K), K] = 0 \tag{7.22}$$

The second-order condition dictates $\Phi_{ww} < 0$. Now the derivation of equation (7.22) with respect to K gives $w'(K) = -\Phi_{wK}/\Phi_{ww}$, so $w'(K)$ is of the sign of Φ_{wK}. With (7.21), we find after several simple calculations:

$$\Phi_{wK} = \frac{(1-\gamma)LR_K(K,L)}{[R(K,L) - wL]^2} > 0$$

The negotiated wage $w(K)$ is thus an *increasing* function of the level of capital. Derivative $w'(K)$ being positive, relations (7.18) and (7.20) then entail $\hat{K} < K^*$.

In sum, the irreversible character of investment gives the firm an incentive to underinvest when the union cannot make a credible commitment not to renegotiate wages once the equipment has been installed. In this situation, the union knows that every strike costs the firm rK, whereas the strike has a cost of zero if it is impossible to renegotiate wages. The union can thus demand a larger share of the profits in the first case, which provokes a reduction in investment. Although we have taken labor demand as fixed, the consequences of underinvestment in terms of employment can be imagined on the basis of its impact on the marginal productivity of labor. If capital and labor are *gross substitutes*—which means that the demand for one factor increases when the cost of the other factor rises; see chapter 2 on labor demand for more detail—underinvestment ought to be favorable to employment, to the extent that any fall in the level of capital will be offset by an increase in employment. Conversely, when capital and labor are *gross complements*—which means that the demand for one factor declines when the cost of the other factor rises (again, see chapter 2)—underinvestment ought to be unfavorable for employment.

4 EMPIRICAL EVIDENCE REGARDING THE CONSEQUENCES OF COLLECTIVE BARGAINING

The great majority of studies on collective bargaining focus on how it affects wages and look at cases in the United States. These studies do bring out a wage differential between unionized workers and others. Still, they face significant methodological challenges, for the fact that one belongs to a union, or works in a firm covered by a collective agreement, is the fruit of an individual decision. Consequently, it is possible that eventual wage differences among unionized and non-unionized workers are the result, at least in part, of differences in the characteristics of individuals and not just the action of unions. The same remark applies to wage or employment differences among firms. Firms where a union is in place may have different characteristics from ones with no union, and these characteristics may underlie variations in performance among firms.

We start by presenting strategies that have been developed empirically to identify the impact of unions on wages, bearing in mind the challenges just mentioned. After a review of traditional methods grounded in ordinary least squares, we describe in detail the paper of DiNardo and Lee (2004), which utilizes the method of regression discontinuity. We then present results concerning the impact of unions on employment, productivity, profits, and investment.

4.1 The Estimation of the Union Wage Gap by Ordinary Least Squares

Empirical studies generally attempt to estimate the wage differential between unionized and non-unionized workers, known as the union wage gap. Let W_u and W_n

be respectively the wage of a unionized and a non-unionized worker; this gap is defined by:[6]

$$\Delta = \frac{W_u - W_n}{W_n} \approx \ln W_u - \ln W_n$$

In order to interpret this wage gap, it is useful to distinguish two types of effect. First, there is a direct effect, which corresponds to the influence of the union on the wages it negotiates. There is also an indirect effect deriving from the fact that the union exerts influence on wages not covered by collective bargaining. So an increase in union-negotiated wages may show up as a contraction of production in the unionized sector, and thus an increase in the demand for goods and labor, from which the non-unionized sector profits. In this case, wages in the non-unionized sector should rise with union power. Conversely, if wage rises due to union power entail a reduction in labor demand in the unionized sector, a worker who fails to find a job in that sector may move into the non-unionized sector. That ought to exert downward pressure on the wages of workers not covered by collective agreements. These observations suggest that wages as a whole are influenced by collective bargaining, and that the gap between wages in the unionized sector and those in the non-unionized sector reflects a combination of interactions, the result of which is ambiguous in sign.

4.1.1 The Equations to Be Estimated

To estimate the impact of unions on wages, earlier research used aggregate data at industry level. It concluded that the rate of unionization had a positive impact on wages (Lewis, 1963). The results of these studies are hard to interpret, however, for it is very difficult to assess differences in the characteristics of manpower between unionized sectors (i.e., sectors where collective agreements prevail) and non-unionized sectors. The most recent work estimates the impact of collective bargaining utilizing individual data.

Studies carried out, beginning in the mid-1970s, on individual data generally estimate two separate wage equations for the unionized and non-unionized sectors, in order to take account of possible differences of return to individual characteristics between these two sectors. These two equations are written respectively:

$$w_{ui} = \sum_j a_{uj} x_{ij} + \varepsilon_{ui} \quad \text{and} \quad w_{nk} = \sum_j a_{nj} x_{kj} + \varepsilon_{nk} \tag{7.23}$$

In these equations, index u locates an individual i belonging to the unionized sector, and index n an individual k from the non-unionized sector. The dependent variable w_{ui} is thus the wage (expressed as a logarithm) of individual i from the unionized sector. The exogenous variables x_{ij} represent the characteristics of individual i (age, sex, region, education, experience, etc.) and ε_{ui} is a random disturbance term. Likewise w_{nk}

[6] Since $\ln x \approx x - 1$ when x is close to 1, we can accept the approximation $\Delta \approx \ln W_u - \ln W_n$ for wages that differ little.

designates the logarithm of the wage of an individual k located in the non-unionized sector, and variables x_{kj} measure her characteristics. The term ε_{nk} again represents a random disturbance.

The estimation of equations (7.23) by ordinary least squares allows us to calculate, for each individual, the wage differential due to the existence of a union. Let $\hat{a}_{uj}$ and $\hat{a}_{nj}$ be the estimates of the coefficients appearing in equations (7.23); the gain of an individual i belonging to the unionized sector is measured by the difference:

$$\widehat{w}_{ui} - \widehat{w}_{ni} = \sum_j (\hat{a}_{uj} - \hat{a}_{nj})\, x_{ij}$$

In summing up 143 studies covering the period 1967–1979 in the United States, Lewis (1986) found that the average markup $(w_u - w_n)$, where w_u and w_n represent respectively the average of the estimates of the w_{ui} and the w_{ni}, was on the order of 15%. More recently, Hirsch (2004) suggested that the union wage premium could be higher in the United States. Using data from the Current Population Survey, he examines the difference in average wages between union and non-union workers, controlling for observable characteristics. He finds that union members earn 14% more. This result is in line with that obtained by other studies. However, Hirsch removed from the sample workers who did not answer the survey. Generally, these workers are assigned the earnings of another worker in other studies. Removing these workers raises the estimated union premium to 20% because many unionized workers who did not answer the survey are assumed to be obtaining the wages of non-unionized workers in those studies. Hirsch also accounts for misclassification errors in union status. He finds that assuming that 2% of reported union members actually do not belong to a union, as one study suggests, raises the union wage gap to 28%.

For the United Kingdom, the summary of Booth (1995a) arrives at a lower average, on the order of 8%. Dell'Aringa and Lucifora (1994) estimate that the wage differential is 4.4% for unskilled workers and 7.4% for skilled workers in mechanical industry in Italy. Studies, based on individual data from various countries, of the impact of collective bargaining on wages conclude that it is greatest in the United States, followed at a distance by the United Kingdom. Blanchflower and Freeman (1992) found that the union markup was 20% in the United States, 10% in the United Kingdom, and lay between 4% and 8% in Australia, Austria, Switzerland, and Germany in the period 1985–1987. These results are confirmed and complemented by Blanchflower and Bryson (2003), who estimate the impact of trade unions in 17 countries. The markup from the 17 countries averages out at 12%. Unions do not have the same impact on wages in all countries. Blanchflower and Bryson find that the union differential in the United States is higher on average than that found in the United Kingdom, 18% compared with 10%. Unions in other countries such as Australia, Austria, Brazil, Canada, Chile, Denmark, Japan, New Zealand, Norway, Portugal, and Spain also raise wages by significant amounts. In France, Germany, Italy, the Netherlands, and Sweden, where union wage settlements spill over into the non-union sector, Blanchflower and Bryson find no significant union wage differentials. Blanchflower and Bryson also analyze the changes over time in union-relative wage effect in the United States and the United Kingdom. It turns out that the union wage premium was untrended from the

beginning of the 1980s to the mid-1990s in both countries. However, the wage premium fell between years 1994 and 2001 in both countries from 14% to 4% in the United Kingdom and from 18% to 13.5% in the United States between these two dates.

4.1.2 The Limits of Ordinary Least Squares

Overall, the results above point to the conclusion that unions exert a positive impact on wages. These results must nevertheless be interpreted with care, for the method of estimating the wage differential runs up against several difficulties.

In the first place, unionized workers may have unobserved characteristics different from those of non-unionized ones, which induces selection bias. We emphasized that collective bargaining reduces the wage gaps between workers with different productivities. If that is the case, the most efficient workers prefer to be employed in the non-unionized sector, and the unionized sector is composed of less productive workers. This type of selection bias, resulting from workers' choices, is known as the "worker choice model" (Lee, 1978). The study of Farber and Saks (1980) finds that the probability of a worker wishing to have a union in his workplace decreases with his position in the distribution of wages in that workplace. It thus confirms the relevance of the hypothesis of the "worker choice model." It is also possible that the presence of a union gives firms an incentive to select a better-quality workforce—an adaptation to the high wages of less skilled workers. That assumes that all workers who wish to be employed in the unionized sector do not necessarily find such employment. This description of worker allocation resulting from the joint choice of workers and firms is known as the "queuing model" (Abowd and Farber, 1982; Farber, 1983). In this context, workers with lower performance are excluded from the unionized sector. The choices of workers and firms thus ought to lead to an allocation of the best-performing workers (who refuse to be unionized) and the worst-performing workers (who are turned down by firms in the unionized sector) to the non-unionized sector. The unionized sector is then composed of workers with an intermediate level of productivity (Abowd and Farber, 1982; Farber, 1983).

In the second place, ordinary least squares estimates are not biased if the rate of unionization is an exogenous variable. But the wage hikes which a union may obtain, on account of high productivity in a firm or sector for example, may in return increase the rate of unionization (Duncan and Stafford, 1980; Checchi and Lucifora, 2002), which must then be considered an endogenous variable.

These two observations lead to the conclusion that ordinary least squares produces a biased estimator of the wage differential. Numerous contributions have tried to overcome selection and endogeneity biases by estimating systems of simultaneous equations and using longitudinal data that make it possible to observe the wage variations of workers whose unionized status changes (see Hirsch and Addison, 1986; Robinson, 1989; and the surveys of Booth, 1995a, and Blanchflower, 1996). The estimation of simultaneous equations, by the method of instrumental variables or by the two-stage estimation procedures of Heckman (1976, 1979; see chapter 1, appendix 7.5.2), arrives at results that lack robustness and are divergent, being very sensitive to the method of estimation, hypotheses concerning the error terms, and the inclusion of supplementary variables.

4.1.3 Longitudinal Studies

Longitudinal data supply information about movement between unionized and non-unionized jobs and make it possible to suppress biases due to fixed time-invariant individual effects not observed by the econometrician. From this perspective, Card (1996) studied the impact of unions on wages with the help of longitudinal data, taking into account classification errors regarding the status of workers, as well as potential correlations between productivity and unionization. Card estimates a model with simultaneous equations for workers belonging to five skill levels, utilizing data from the Current Population Survey (United States) for 1987–1988. His results suggest that the positive effect of unions on wages is greater, the less skilled workers are. Moreover, he finds that selection biases differ from one group to another. Among the least-skilled workers, the most efficient ones are in the unionized sector on average (in conformity with the queuing model), while the opposite is true for the most highly skilled workers (in conformity with the worker choice model). Lemieux (1998) obtains qualitatively similar results on longitudinal data from Canada.

Nevertheless, longitudinal data are very sensitive to measurement errors, for small measurement errors concerning the status of workers lead to major biases if there is low mobility between unionized and non-unionized jobs. Moreover, longitudinal studies assume that the move of a worker from a unionized job to a non-unionized one is exogenous, as if individuals were assigned randomly to jobs. In reality, workers choose to move between jobs, and they observe many characteristics of jobs that are not observed by the econometrician. This unobserved heterogeneity can also induce important biases, especially if job amenities are different across union and non-union jobs.

4.2 Regression Discontinuity

The empirical research reviewed to this point yields debatable results, inasmuch as it struggles to identify an exogenous variation in unionization, independent of factors that could influence wages. More recent studies, following the contribution of DiNardo and Lee (2004), have relied on a more credible source of variation to identify union effects in the United States. This analysis is based on the fact that most recognitions of unions occur as a result of an election by secret ballot. By law, if a majority of workers votes in favor of bringing in a union, the National Labor Relations Board (NLRB) rules require company management to bargain "in good faith" with the recognized union. This process creates a natural set of comparisons between establishments where the union barely won the certification election (say, by one vote) and those where the union barely lost the certification election (by one vote). This fact allows DiNardo and Lee to use a regression-discontinuity design, where the comparison between near winners and near losers potentially eliminates any confounding selection and omitted-variable biases.

4.2.1 The Design

DiNardo and Lee (2004) use this strategy to analyze the effects of unions on a sample of establishments that faced a process of union recognition in the United States during 1984–1999. They combine several different data sets: (1) the set of National Labor Relations Board representation elections held between 1984 and 1999; (2) the set of contract expiration notices from the Federal Mediation and Conciliation Service during

1984–2001; (3) business survivorship, employment, and estimated sales volume from a commercial database with information on the population of businesses with a telephone number, as of the year 2001; and (4) detailed employment, output, investment, and wage information from the Census Bureau's Longitudinal Research Database on manufacturing establishments in the United States from 1974 to 1999.

Formally, regression-discontinuity designs rely on the following system of equations:

$$y = \mathbf{x}\boldsymbol{\gamma} + \beta D + \varepsilon \tag{7.24}$$

$$D = \begin{cases} 1 & \text{if } v \geq 1/2 \\ 0 & \text{otherwise} \end{cases} \tag{7.25}$$

$$v = \mathbf{x}\boldsymbol{\delta} + u \tag{7.26}$$

where y is the outcome of interest in each firm (average wage, employment, profit, ...), D is the indicator of union recognition status, v is the vote share for the union in the representation election, $\mathbf{x}$ contains observable characteristics that are assumed to determine the vote share, and ε and u are corresponding unobservable determinants. β is the parameter of interest and corresponds to the impact that flows from the union "treatment," and $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ are vectors of parameters to be estimated.

We know that OLS estimates of equation (7.24) are generally biased since, in general, the unobservable determinants are correlated with union status, or in more formal terms, $\mathbb{E}[\varepsilon|v \geq 1/2] \neq \mathbb{E}[\varepsilon|v < 1/2]$.

By contrast, the regression-discontinuity design allows us to obtain exogenous variations in union status, changes in union status that are independent of the couple $(\mathbf{x}, \varepsilon)$ of observable and unobservable characteristics that determine the vote share. To grasp this point more clearly, note that if there is some uncertainty in what determines the vote share v, we would expect the density of v (and hence u) conditional on $(\mathbf{x}, \varepsilon)$, to be continuous at the threshold $v = 1/2$. If we admit this assumption, then the variation in treatment status has the same statistical properties as a randomized experiment; in particular, the distribution of $(\mathbf{x}, \varepsilon)$ will be approximately the same across the treated (i.e., unionized firms) and control groups (i.e., non-unionized firms) within a small neighborhood of $v = 1/2$.

To show this, let us denote by $f(v|.)$ and $f(v)$ the conditional and unconditional distribution of v. Then, by Bayes rule, we have:

$$\Pr\left[\mathbf{x} = \mathbf{x}^0, \varepsilon = \varepsilon^0 | v\right] = f\left(v|\mathbf{x} = \mathbf{x}^0, \varepsilon = \varepsilon^0\right) \frac{\Pr\left[\mathbf{x} = \mathbf{x}^0, \varepsilon = \varepsilon^0\right]}{f(v)}$$

If we assume that $f\left(v|\mathbf{x} = \mathbf{x}^0, \varepsilon = \varepsilon^0\right)$ is continuous at $v = 1/2$ for any value of x^0 and ε^0 then $f(v)$ is also continuous at $v = 1/2$. The Bayes rule implies that the distribution of $(\mathbf{x}, \varepsilon)$ conditional on v—defined by $\Pr\left[\mathbf{x} = \mathbf{x}^0, \varepsilon = \varepsilon^0 | v\right]$—is continuous at $v = 1/2$. This means that all observed and unobserved predetermined characteristics have identical distributions in a small neighborhood around $v = 1/2$ (see Lee and Lemieux, 2010, for further details).

It is worth noting that this assumption of uncertain determination of the vote share around the threshold can be tested by assessing whether there are discontinuities in

the relation between the vote share and any predetermined characteristic in $\mathbf{x}$. That is, a sharp discontinuity in $\mathbb{E}(\mathbf{x}|v)$ at $v = 1/2$ would provide evidence against the assumption that there is genuine uncertainty in the vote share.

In addition, for all $\Delta > 0$ equations (7.24) and (7.25) entail:

$$\mathbb{E}\left(y|v = \frac{1}{2} + \Delta\right) = \mathbb{E}\left(\mathbf{x}|v = \frac{1}{2} + \Delta\right)\boldsymbol{\gamma} + \beta + \mathbb{E}\left(\varepsilon|v = \frac{1}{2} + \Delta\right) \tag{7.27}$$

$$\mathbb{E}\left(y|v = \frac{1}{2} - \Delta\right) = \mathbb{E}\left(\mathbf{x}|v = \frac{1}{2} - \Delta\right)\boldsymbol{\gamma} + \mathbb{E}\left(\varepsilon|v = \frac{1}{2} - \Delta\right) \tag{7.28}$$

We have seen that the hypothesis of genuine uncertainty in the vote share implied that the conditional distribution of the couple $(\mathbf{x}, \varepsilon)$ is continuous in a neighborhood of $v = 1/2$. From this property it results that:

$$\lim_{\Delta \to 0^+}\left[\mathbb{E}\left(\mathbf{x}|v = \frac{1}{2} + \Delta\right) - \mathbb{E}\left(\mathbf{x}|v = \frac{1}{2} - \Delta\right)\right] = \lim_{\Delta \to 0^+}\left[\mathbb{E}\left(\varepsilon|v = \frac{1}{2} + \Delta\right) - \mathbb{E}\left(\varepsilon|v = \frac{1}{2} - \Delta\right)\right] = 0$$

Subtracting relations (7.27) and (7.28) term for term we get:

$$\lim_{\Delta \to 0^+}\left[\mathbb{E}\left(y|v = \frac{1}{2} + \Delta\right) - \mathbb{E}\left(y|v = \frac{1}{2} - \Delta\right)\right] = \beta \tag{7.29}$$

This last relation signifies that if the assumption of genuine uncertainty is fulfilled, parameter β represents well the expected effect of unionization on the outcome y.

4.2.2 The Results of DiNardo and Lee

DiNardo and Lee (2004) first document the fact that the outcome of a National Labor Relations Board election has a substantial, binding impact on the collective bargaining process, however close the result. Where they narrowly win the election, unions are able to maintain their legal recognition over long time horizons; where they narrowly lose, there is little evidence of subsequent attempts to organize the workplace. Furthermore, unions that narrowly win have as good a chance of securing a collective bargaining agreement with the employer as those that win the elections by wide margins. And, as expected, unions that narrowly lose an election have little chance of ever signing such an agreement.

Further, it is useful to look at the graphic presentation in figure 7.12 of the results concerning the impact of union certification on wages. The Postelection curve represents the means of the (log of) hourly wages of production workers, by union vote share category, for establishment-year observations in the years that follow the election. A discontinuity at 50% represents the estimate of the causal impact of unionization. It appears that there is no discontinuity, suggesting that union certification has no significant impact on wages around this 50% threshold. The solid triangles plot averages of the wage after it has been deviated from its preelection mean. That is, in order to reduce the sampling variability in the discontinuity estimate, each postelection observation is deviated from

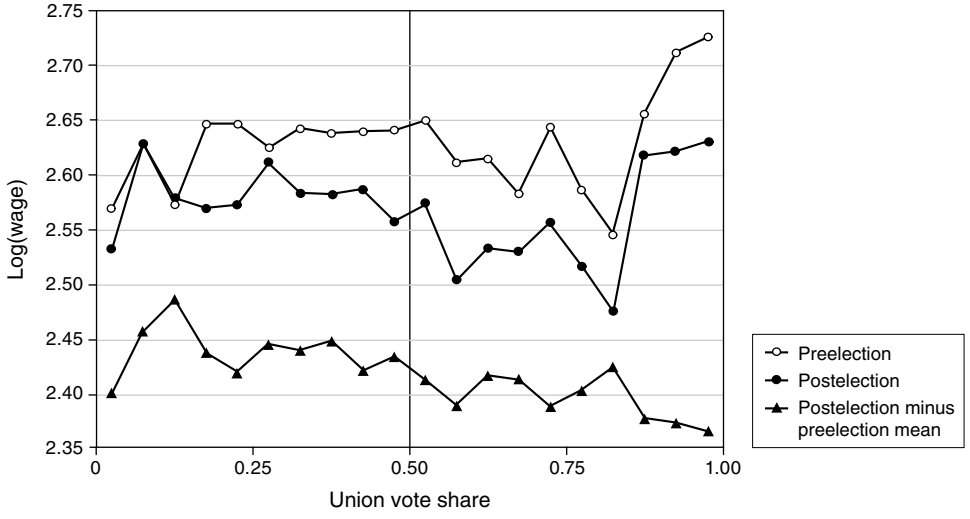


Figure 7.12
Log(production hourly wage), pre- and postelection, by union vote share.

Source: DiNardo and Lee (2004, figure IXb).

the overall mean that uses all observations before the election, for each plant. In a randomized experiment, this transformation should not affect the impact estimates, since presumably the preelection mean is independent of treatment status. Here too, we do not discern any discontinuity at the 50% vote share cutoff, confirming the absence of effect of unionization on average wages around this threshold. Estimations also confirm this result: point estimates are small (between 0 and 2%) and statistically insignificant, three years after the election. Finally, the open circles represent the observations strictly before the year of the election. For this plot, a significant discontinuity at the 50% cutoff would indicate that close winners and losers are systematically different before the election, which would imply a problem with the research design. Figure 7.12 suggests that this is not the case.

The no-wage-effect result of DiNardo and Lee could be an artifact of union threat effects, whereby employers raise wages to avoid the threat of future unionization. To deal with this issue, DiNardo and Lee complement their regression-discontinuity analysis with an "event-study" analysis that assesses whether wages rise in response to an election, even if the union eventually loses. They do not find any statistically significant wage raises before elections.

All in all, DiNardo and Lee obtain results that are very different from those obtained by previous studies relying on alternative identification strategies, which generally find significant union wage effects, corresponding to a union wage premium of about 15%. One reason for the difference may lie in the regression-discontinuity design, which allows us to analyze the effects of unions that barely won the elections. This design does not allow us to detect the effects of unions that got strong support from workers. Such unions may be more powerful than those that are barely elected. We will see later that this interpretation is plausible. The results of DiNardo and Lee may also be due to the fact

that their dependent variable is the mean hourly wage at the firm level. Measurement errors on this variable are likely important. Moreover, if unions compress wages, their impact on the mean wage can be small.

Another interpretation of the no-wage-effect result of DiNardo and Lee could be that studies using simple OLS estimates, which are probably biased, are totally misleading. However, other papers using regression-discontinuity design, which find significant effects of unionization on wages and other outcomes, suggest that this is not the case. Sojourner et al. (2012) study the effects of unions in private-sector nursing homes in the United States using a regression-discontinuity design to identify union effects by contrasting outcomes in nursing homes where unions closely won representation elections to outcomes in facilities where unions closely lost such elections. They find significant union wage premiums for some classes of nursing labor, equal to 14.8% for nurse aides and to 8.5% for registered nurses. Frandsen (2012) also uses regression-discontinuity to examine the impact of unions on the distribution of wages and finds that unions compress the wage distribution without much affecting average wage.

All in all, studies using regression-discontinuity design show that labor unions that are narrowly elected do not always have a significant impact on wages. We will see later that empirical evidence does suggest however that stronger unions, which get significantly more than 50% of the votes, have a significant impact on wages.

4.3 Wage Inequalities

We have just stated that workers covered by collective agreements may obtain higher wages. This effect ought to tend to increase the dispersion of wages throughout the economy as a whole. We have also stated, however, that workplaces covered by collective agreements have more compressed wage structures than others. These observations suggest that the impact of collective agreements on wage dispersion is a priori ambiguous.

This impact may be grasped by decomposing the variance v of the logarithm of wages in the economy as a whole, as a function of the proportion α of unionized workers (or ones covered by a collective agreement), the variances v_u and v_n of the logarithms of wages in the unionized and non-unionized sectors, and the averages of the logarithms of wages w_u and w_n in these two sectors. The result is[7] (Freeman, 1980; Fortin and Lemieux, 1997):

$$v = \alpha(1-\alpha)(w_u - w_n)^2 + \alpha v_u + (1-\alpha)v_n \qquad \textbf{(7.30)}$$

[7] Let N and w respectively be the size of the sample and the average of the logarithms of the wages. We thus have:

$$v = \left(\frac{1}{N}\sum_{i=1}^{i=N} w_i^2\right) - w^2 \quad \text{with} \quad w = \frac{1}{N}\sum_{i=1}^{i=N} w_i = \alpha w_u + (1-\alpha)w_n$$

Let $\mathcal{U}$ (respectively $\mathcal{N}$) be the set of unionized (respectively non-unionized) workers; the variances v_u and v_n satisfy the following equalities:

$$\frac{1}{N}\sum_{i=1}^{i=N} w_i^2 = \frac{1}{N}\sum_{i\in\mathcal{U}} w_i^2 + \frac{1}{N}\sum_{i\in\mathcal{N}} w_i^2 = \alpha\left(v_u + w_u^2\right) + (1-\alpha)\left(v_n + w_n^2\right)$$

Substituting this expression into the preceding relation, we have:

$$v = \alpha\left(v_u + w_u^2\right) + (1-\alpha)\left(v_n + w_n^2\right) - [\alpha w_u + (1-\alpha)w_n]^2$$

Developing and rearranging terms, we find formula (7.30).

This relation shows that the variance may be decomposed as the sum of a between-group variance and a within-group variance. The between-group variance, $\alpha(1-\alpha)(w_u - w_n)^2$, shows that the wage differential between the unionized and non-unionized sectors accentuates the inequalities. But unions also exert an influence that may work in the opposite direction, to the extent that they alter the dispersion of the wages they negotiate. This effect is represented by the last two terms, which take into account the variances of the two sectors weighted by their respective size. Empirical studies generally show that wage variance is weaker in the unionized sector in the United States, and they conclude that this second effect tends to play a dominant role, so that the total impact of unions on inequalities as a whole is negative (Freeman and Medoff, 1984; Blau and Kahn, 1999; Frandsen, 2012).

It should be remarked that there are potential important selection issues in the evaluation of the impact of unions on wage dispersion. For instance, Metcalf et al. (2001) find that wages are also more compressed in union firms than in non-union firms in the United Kingdom. However, they show that part of this wage compression is due to the fact that union members are more similar than workers in non-union firms and naturally earn more similar wages. They also find that unions negotiate contracts that reduce the returns to individual skills and ability, ones for instance in which seniority pay outranks merit pay. In the same perspective, Lemieux (1998) estimates the effects of unions on wages in Canada, explicitly correcting for measurement errors. He finds that the average union member earns 28% more than the average non-union member. However, unions cause less than two fifths of this wage premium. The rest comes from unmeasured individual characteristics. Workers who switch to union jobs see their wages rise by only 10%.

This being said, the experience of each country when it comes to inequality depends on its institutions, in particular the union density, the coverage of collective agreements, and the degree to which bargaining is coordinated. For instance, Card and de la Rica (2006) study the impact of collective bargaining in Spain where, as in several other European countries, sectoral bargaining agreements are automatically extended to cover all firms in an industry. Employers and employees can also negotiate firm-specific contracts. Card and de la Rica find that employees covered by firm-specific contracts earn about 10% more than those covered by sectoral contracts. The estimated premium is about the same for men in different skill groups but higher for more highly skilled women, suggesting that firm-level contracts raise wage inequality for women. This result is related to those of Card et al. (2004), who find that unions reduce inequality for men but not for women in the United States, Canada, and Great Britain.

The studies of Rowthorn (1992), Blau and Kahn (1996), and Kahn (1998, 2000) carried out on data from OECD countries find negative correlations between the union density (or the coverage of collective agreements) and wage inequalities. They also obtain significant negative correlations between the degree to which collective bargaining is centralized and wage inequalities. The study of Kahn (2000) in particular, which uses individual data for 15 OECD countries for the period 1985–1994, shows that an increase in the coverage of collective wage bargaining leads to relatively higher wages for low-skilled workers.

These results suggest that as institutions change, changes in wage inequalities may follow. Figures 7.5 and 7.6 show that the union density and the coverage of collective bargaining have fallen off sharply between the end of the 1970s and the beginning of the

1990s in certain OECD countries, in particular the United States and the United Kingdom. It is tempting to make a connection between these changes and the increased inequality of wages observed in these two countries (see chapters 10 and 11). In this respect, DiNardo et al. (1996) estimate that the fall in union density has contributed 10% to the increase in the differential of (the logarithms of) wages between the first and last deciles, and one third to the increase between the first and the fifth deciles in the United States, in the 1980s. Card (2001) finds that the decline in unionization explains between 15% and 20% of the growth in wage inequalities (measured by the variance of wage logarithms). For women, on the other hand, wage inequalities are not affected by the change in the global rate of unionization.

4.4 Employment

Models of wage bargaining have shown that the effect of collective agreements on employment depends strongly on the hypotheses about the sequence of decisions and about the set of variables submitted to bargaining. From this perspective, research carried out in the 1980s and 1990s has tested certain predictions of collective bargaining models to find out if they lead to efficient contracts or contracts of the right-to-manage type, in order to deduce their impact on employment. Other research has attempted to evaluate directly the impact of unions on employment.

4.4.1 Tests of Efficient Contract Models

Two different approaches have been used to test the efficiency of collective agreements. Ashenfelter and Brown (1986) estimated the properties of the relationship between employment and wages, while Abowd (1989) worked directly on the payoff functions of firms and unions.

Ashenfelter and Brown (1986) used data concerning a particular union (the International Typographical Union in the United States). The employment–wage relationship corresponds to the equation of the contract curve of the model of bargaining over wages and employment described in section 3.2.2 above. Applied to workplace i, this equation is written in log-linear form:

$$\ln L_{it} = a_0 + a_1 z_{it} + a_2 \ln w_{it} + a_3 \ln w_{ait} + u_{it} \qquad \textbf{(7.31)}$$

In this relation, L_{it} designates employment or the number of hours worked in workplace i at time t; z_{it} represents a vector of non-wage variables comprising lagged employment, fixed effects for localization, and productivity indicators; w_{it} is the (minimum) hourly negotiated wage; w_{ait} is the outside wage (including, according to specifications, the average wage in the manufacturing industry and unemployment insurance benefits); and finally u_{it} represents a random error term.

Whatever the value of the negotiated wage, the equation of the contract curve described in section 3.2.2 shows that the outside wage (denoted $\overline{w}$ in section 3.2.2) generally has an impact on employment. Conversely, if the solution of the bargaining is situated on labor demand—which is the case in the right-to-manage model—employment becomes independent of the outside wage, since labor demand, deduced from profit maximization behavior, does not depend on this parameter.

The negotiated wage being an endogenous variable, equation (7.31) is estimated by the method of instrumental variables. The instruments chosen are the lagged wage and the levels, actual and lagged, of the consumer price index. It is most often found that the values of coefficients a_2 and a_3 are very sensitive to the specification of the variables and have little significance. Thus Ashenfelter and Brown (1986) reject the hypothesis that the outside wage has a significant impact on employment but find some indication that the negotiated wage may have a negative influence on employment. These results would thus indicate that the contracts negotiated (by the International Typographical Union) are not optimal. It is best, though, to remain cautious, for as Pencavel (1991) points out, the absence of a relationship between the level of employment and that of the outside wage does not necessarily signify that the contract is not optimal. For example, if the utility function of the union takes the form $\mathcal{V}_s(w, L) = h(L)(w/\overline{w})^{\zeta};\ \zeta > 0,\ g'(L) > 0$, the contract curve is independent of $\overline{w}$. It is easy to verify that its expression is $R'(L) = w - (g'(L)/g(L))/(w/\zeta)$. In addition, the difficulties inherent in the definition of the outside wage and the sensitivity of results to the specification chosen, render any conclusion about the role of this variable fragile.

Card (1986, 1990) has also studied the impact of the outside wage in employment equations of the type (7.31). He shows, relying on data for the aeronautical industry in the United States (Card, 1986) and manufacturing industry in Canada (Card, 1990), that the correlations observed between employment and the outside wage are not consistent with the predictions of efficient contract models. Abowd and Kramarz (1993) come to a similar conclusion. They find, for French data on 1,097 firms in the period 1978–1987, that estimates of labor demand equations incorporating solely the outside wage (specified as the minimum wage multiplied by an index of the average wage of the decile below the category of manpower under consideration) are much less good than estimates that take only the negotiated wage into account.

Abowd (1989) adopts a strategy that makes it possible to overcome the difficulty linked to the specification of the outside wage. He assumes that the union maximizes the rent of its members, defined as employment multiplied by the difference between the negotiated wage and the outside wage: $\mathcal{V}_s = (w - \overline{w})L$. In this case, it is easy to verify that the contract curve, the equation of which is $R'(L) = \overline{w}$, is a vertical line in the (w, L) plane, as shown in figure 7.11 above. The total revenue $R(L)$ then becomes independent of bargaining power (which is not the case if the solution of bargaining is found on labor demand or on the contract curve). In consequence, the sum of profit $\Pi = R(L) - wL$ and union rent $\mathcal{V}_s$ amounts to $R(L) - \overline{w}L$, which depends *only* on wage $\overline{w}$. Any variation in union bargaining power will then entail $\Delta\Pi = -\Delta\mathcal{V}_s$. In other words, any increase in the wealth $\Delta\Pi$ of shareholders in the company should entail a reduction in union rent by the same amount when the power of the union diminishes. For 2,228 private-sector contracts, excluding construction, in the United States for the period 1976–1982, Abowd estimates relation:

$$\Delta\Pi = a_1 + a_2\Delta\mathcal{V}_s + u \qquad \textbf{(7.32)}$$

In this equality, u represents a random error term. Variations in profit are measured using the difference in the price of company shares three months before the date a collective agreement is signed and the price observed on that date. A similar approach is used to find variations in the union rent. Abowd calculates the value of the rent $[w - R'(L)]\,L$ at every date and from that deduces its variations. The estimation of this equation does not

allow us to reject the hypothesis $a_1 = 0$ and $a_2 = -1$. In consequence, we cannot exclude the possibility that collective bargaining may arrive at efficient contracts.

Attempts to assess the efficiency of contracts have not achieved clear conclusions. The estimation of an employment–wage relationship leads to rejection of the hypothesis that the alternative wage plays a determining role. This makes it possible to exclude only the model with a vertical contract curve. But the study of Abowd (1989) ends by accepting this very model, that is, the opposite result.

4.4.2 Tests of the Right-to-Manage Model

Tests of the right-to-manage model try to verify whether the solution of the bargaining lies on labor demand. MaCurdy and Pencavel (1986) use the result that, in the right-to-manage model, the marginal productivity of labor is equal to its cost. With the same data as Ashenfelter and Brown (1986), they first estimate production functions, in order to find the marginal productivity of labor. They then show that variations in the latter are explained by the current wage but also by other variables, like the outside wage or the level of employment which the union incorporates into its objective. They conclude that the solution of bargaining is not situated on labor demand. The validity of this approach rests on the quality of the estimate of marginal productivity. Now, variables modifying the utility of the union can have an impact on the behavior of individuals, which affects their productivity and shifts labor demand. In light of this, the results of MaCurdy and Pencavel (1986) are very fragile.

Nickell and Wadhwani (1990) test the sequential model of Manning (1987), in which bargaining power over employment differs from bargaining power over wages. Using data relative to 219 firms in the manufacturing industry in the United Kingdom between 1972 and 1982, they estimate the labor demand function resulting from bargaining over employment, with an exogenous wage. We saw in section 3.2.2 that this function depends on, among other things, the outside wage and bargaining power over employment (captured in this study by the rate of unionization). It appears that neither of these two variables has a significant impact on employment. This result may point to the conclusion that firms are on their labor demand. Nickell and Wadhwani emphasize, however, that their results are fragile, since the reservation wage is very poorly defined, and the rate of unionization is not a good measure of bargaining power over employment.

4.4.3 Direct Estimations

Other researchers have tried, using an approach different from that of the papers reviewed above, to estimate directly the effect of unions on employment by regressing the level of employment on variables measuring union power.

Boal and Pencavel (1994) tried to estimate the effects of bargaining on wages and employment directly. Their study uses data relative to labor in the coal mines of Virginia between 1897 and 1938. These data are available for 35 counties. At each date, there are counties in which the unions actually play a part in the bargaining process and other counties where there are no unions. The authors assume that employment and wages are determined competitively in these counties. This division of the counties into two groups makes it possible to estimate the wage gap between the "unionized" counties and the "competitive" counties. It emerges that starting in 1921, the wage gap differs

significantly from zero. It reaches 18% over the period 1921–1930 and 23% between 1931 and 1938. On the other hand, differences in terms of employment are never significantly different from zero, although the number of days worked is, on average, 17% lower in the counties where unions exist. Hence the study of Boal and Pencavel shows that a large wage differential does not necessarily have a negative effect on employment. It is possible that the presence of a union leads to a change in the internal organization of firms that, in return, alters the linkage between employment and wages.

The contribution of Kahn (2000), on 15 OECD countries for the period 1985–1994, brings out a negative correlation between the degree of union coverage and the relative employment rate of low-skilled workers. Kahn also shows that unions allow these workers to obtain higher relative wages, which suggests that unions contribute to the compression of the wage structure at the expense of the employment of less-skilled workers.

Changes in legislation influencing union power constitute interesting experiments for the assessment of the impact of unions on employment: they can be like exogenous shocks, the consequences of which the econometrician can identify. The reforms introduced by the Thatcher government in the United Kingdom in the 1980s fall into this category, since they limited union power, notably by abolishing the "closed shop" (the obligation for all workers in a firm with a collective agreement to belong to the union). The effect of these reforms was to diminish the rate of unionization and the collective bargaining coverage of collective agreements, and studies find that the response of wages and employment to variations in demand rose following these changes. The comparison of wages and employment in the unionized and the non-unionized sectors suggests, though, that the reforms do not appear to have had an impact on unemployment or on the chances of exiting from unemployment (Blanchflower and Freeman, 1994). The study of Maloney (1994), which looks at reforms introduced in New Zealand in 1991 that substantially reduced union power, comes to different conclusions. Maloney finds that the strong reduction in the rate of unionization had a positive impact on employment.

The results of this research are nevertheless fragile, given that unionized and non-unionized sectors have widely varying trends in employment. In this context, the common trend hypothesis, required for the validity of difference-in-differences estimations (see chapters 1 and 14), has a good chance of not being verified. The phenomenon is well illustrated by Linneman et al. (1990), who examine changes in the union wage premium and union employment in the United States from 1973 to 1986. They show that there is a strong correlation between de-industrialization and de-unionization, with unionized manufacturing jobs disappearing and non-union employment stable.

More recent studies, relying on regression-discontinuity strategies, also obtain diverging results. DiNardo and Lee (2004), in the article presented above, are unable to bring out a statistically significant impact of unions on employment in the manufacturing sector in the United States for the period 1984–2001. Using the same empirical strategy, Sojourner et al. (2012) found that unionization of nursing homes in the United States led to a significant decline in employment. They also found support for the idea that this decrease in staffing corresponded to higher wages. Frandsen (2012) estimates that unionization reduces employment of the lowest skilled workers in the United States, a finding consistent with those of Kahn (2000) for 15 OECD countries.

All in all, empirical studies arrive at very heterogeneous results, so it is impossible, on the basis of these works, satisfactorily to assess the impact of collective bargaining on employment. Empirical studies appear to converge on only two points. For one thing,

the hypothesis that the marginal productivity of labor is equal to the outside wage must be rejected, and for another, there are grounds for positing a negative correlation between employment and bargained wages. It should be noted that the conclusions of the right-to-manage model and those of the insiders–outsiders model with no discrimination against entrants do not contradict these two stylized facts.

4.5 Productivity and Profits

According to theory, unions and collective agreements have an ambiguous impact on productivity. Unions can reduce productivity by limiting the power of employers (Robinson, 1989). But unions can also improve productivity by improving the circulation of information among workers, and their motivation. Hence the impact of unions on profits is ambiguous in theory: if unions improve productivity sufficiently, they may push up not only wages but profits too. Only empirical research can shed light on this matter.

4.5.1 Exit, Voice, and Productivity

Following the work of Hirschman (1970), Freeman and Medoff (1984) maintain that unions improve productivity by improving the circulation of information among workers and their motivation. They claim that this characteristic of unions plays an essential role in the United States, in combination with the exercise of their monopoly power. They assert that by giving workers a voice, unions profoundly change social relations within the firm. Without them, workers adopt a strategy of defection or "exit"—they disengage from the relationship established with a person or an organization when that relationship proves unsatisfactory. The efficiency of the union lies in the fact that it favors the choice of a strategy of "voice," by transmitting complaints, grievances, and demands, with the aim of correcting and improving the relationship. In consequence, by improving the circulation of information between wage earners and employers, unions are capable of increasing the productivity of firms.

Freeman and Medoff (1984) estimate that the reduction in the turnover rate in the workforce due to unions allows American firms, on average, to reduce their labor costs by around 2%. They show as well that the productivity of labor is often higher in unionized firms. Examination of the results of collective agreements in France leads to results of the same type (Cahuc and Kramarz, 1997). Although empirical studies in the United States and in European countries have produced widely varying results, including coefficients with opposite signs, the synthesis of these studies suggests that unions have, at most, a small positive effect on productivity, at least in the United States (Hirsch, 2007, 2008). Doucouliagos and Laroche (2003) find, in a meta-analysis of 73 studies, that the simple mean of the estimated union productivity effect is about 4% and the weighted average is around 1%.

Most of these results are however subject to the same biases as the ones encountered in the estimation of wage differentials and employment effects of unions. Productivity gaps and turnover of manpower may result from unobserved characteristics of workers and firms and may influence behavior when it comes to unionization or the bargaining of collective agreements. So these results should be interpreted cautiously.

Regression-discontinuity studies also find varying results. DiNardo and Lee (2004) find no impact of unions on productivity in the manufacturing industry whereas Sojourner et al. (2012) find positive labor productivity effects in the nursing sector.

All in all, results on the effects of unions on productivity are fragile. The relationship needs to be analyzed further, with proper identification strategy to check whether the positive correlation between productivity and unionization that is often found by empirical studies is a genuine effect of unionization.

4.5.2 Profits and Unions: An Example with the Event-Study Method

The impact of unions on profits is less subject to debate. Studies of the process of setting up a procedure for collective bargaining (through a majority vote of the workers in the United States) and of the effect of the announcement of a renegotiation show that the share price of firms falls (Ruback and Zimmerman, 1984; Abowd, 1989). Freeman and Medoff (1984) examine the link between unionization and the rate of return on capital, coming to similar conclusions.

Van Reenen (1996) studied the movement of wages in firms that had introduced innovations in the manufacturing industry in the United Kingdom. These firms had also signed collective agreements with unions. Van Reenen shows that the innovations had a positive impact on wages over at least seven years. This result signifies that the workers covered by collective agreements obtain a share of the profit of their firms.

The study of Lee and Mas (2012) provides interesting evidence on the effect of new private-sector unionization on publicly traded firms' equity value in the United States over the period 1961–1999. This study analyzes the same type of event, union certification, as the study of DiNardo and Lee (2004) presented above. Lee and Mas analyze the evolution of the equity values of firms before and after the certification of a union, using an "event-study" method, which identifies the impact of unionization as the difference-in-differences between the equity values of unionized firms (belonging to the treatment group, where the election entails union certification) and non-unionized firms (belonging to the control group) before and after union certification. A critical issue in event studies is to define the control group. To do this, Lee and Mas match every firm to a portfolio of firms in the same size decile, based on market value, and compare the evolution over time of their cumulative returns. With this strategy, it appears, as shown by figure 7.13, that before the date of certification of a union, firms have similar cumulative returns, but the relative cumulative returns of firms drop when they become unionized. The drop is significant: the average change in the equity value of the firm is equivalent to $40,500 per unionized worker. This drop is not instantaneous: it takes about 15 to 18 months after unionization to fully materialize. However, since event studies do not rely on exogenous events that explain the certification of unions, the interpretation of this drop is debatable. It does not necessarily recover a causal impact of union certification. An alternative interpretation is that the event of a union victory is a "signal" of poor management.

Lee and Mas also analyze the impact of unionization with a regression-discontinuity design, as in the study of DiNardo and Lee (2004) presented above (section 4.2.1). Strikingly, they find considerably smaller and close to zero effects, which are consistent with those found in the study of DiNardo and Lee. The difference between the results of the event-study approach and the regression-discontinuity design seems

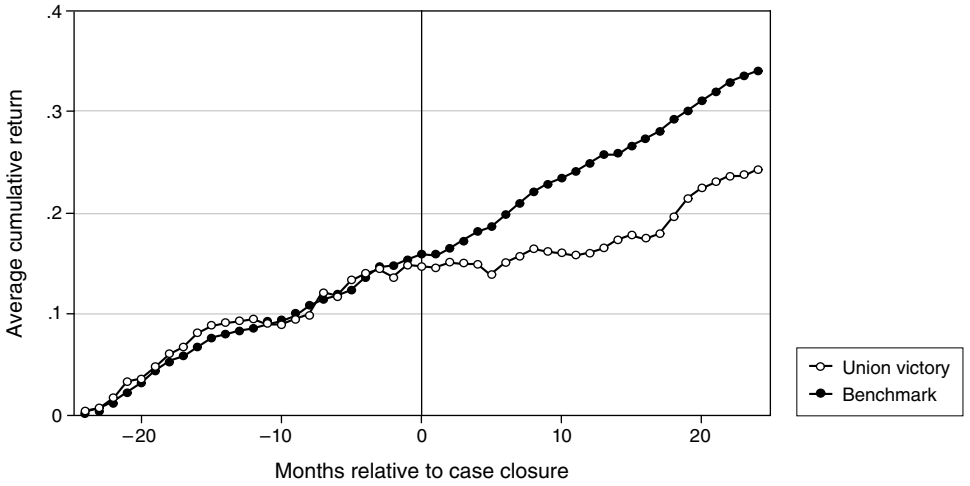


Figure 7.13
Average cumulative returns of union victory firms and of the size-matched reference portfolio, by month relative to National Labor Relations Board case closure (i.e., date of election certification).

Source: Lee and Mas (2012, figure 2).

to come, at least partially, from the fact that regression-discontinuity analysis uses a discontinuity in the relationship between firm performance and the vote share at the 50% threshold, whereas event studies compare firms where the share of votes in support of unionization goes from zero to one. Accordingly, the regression-discontinuity analysis is fundamentally unable to provide a counterfactual for the set of elections where a large majority of workers voted in favor of unionization. Lee and Mas show that the event-study strategy predicts that unions have a nonsignificant impact on equity values of firms when the share of votes in favor of the union is close to 50%. However, the event-study strategy predicts that unions that have been certified with a large majority of votes (above 60%) have a strong negative impact on equity values of firms. This result suggests that unions can exert their power only if they are supported by a large majority of workers. From a methodological point of view, it also emphasizes that the regression-discontinuity design can detect local causal effects, in the neighborhood of the discontinuity threshold, but is generally less suitable to detecting global effects.

4.6 Investment and Capital Structure

We have stated that the capacity to renegotiate wages may lead to a reduced level of investment and that this effect is greater, the more bargaining power the wage earners have. To verify this prediction, empirical studies estimate a relation of the type:

$$\ln I_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \delta u_{it} + \varepsilon_{it}$$

In this equation, I_{it} and $\mathbf{x}_{it}$ designate respectively the level of investment and a vector of the characteristics of the firm or sector i at date t which influence investment. The term u_{it} is an indicator of the union presence (such as the rate of unionization and the

number of days lost through strikes), ε_{it} is a random disturbance term, and $\boldsymbol{\beta}$ and δ are coefficients to be estimated.

The results obtained by this method indicate that unions exert a negative effect on investment in physical capital in the United States (Connolly et al., 1986; Hirsch, 1992, 2007; Bronars et al., 1994; Fallick and Hassett, 1999), in Canada (Odgers and Betts, 1997), and in the United Kingdom (Denny and Nickell, 1992). The loss of investment attributed to unions is generally of significant size. Hirsch finds figures on the order of 20% for the United States, while Denny and Nickell obtain, for the United Kingdom, a reduction lying between 3% and 16% according to the firm in question. Moreover, the estimates show that the effects are not linear. The marginal impact of an increase in union presence on investment in a sector is greater when the union density is slight (Hirsch, 1992, 2007; Odgers and Betts, 1997). This phenomenon can be explained by the effect of the spread of incipient unionization in a sector, to which non-unionized firms respond by increasing wages to make unionization harder in their plants.

These studies deal with the accumulation of physical capital in firms. Tan et al. (1992) suggest that the presence of a union is favorable to investment in the human capital of firms, since it is generally associated with outlays on training on the part of employers in the United States, the United Kingdom, and Australia. Dustmann and Shoenberg (2009) find support for the hypothesis that union recognition, via the imposition of wage floors and wage compression, increases training in apprenticeship programs in Germany.

The bargaining power of unions can modify not just the level of investment but also the way it is financed. Because maintaining high levels of corporate liquidity may encourage workers to raise their wage demands, a firm with external finance constraints has an incentive to use the cash flow demands of debt service to improve its bargaining position with workers. Cross-sectional analysis shows positive correlations between financial leverage and unionization rate (Bronars et al., 1991; Hirsch, 1992). For instance, Matsa (2010) estimates that a firm with a 50% unionized workforce is associated with 15% to 20% greater financial leverage (i.e., the debt-to-capital ratio) than a typical non-unionized firm in the manufacturing sector in the United States. Furthermore, Matsa uses states' adoption of right-to-work laws in the 1950s as sources of exogenous variation in union power. We have seen that once a union is certified by the National Labor Relations Board, introduced in 1935 by the National Labor Relations Act (the Wagner Act), the employer is required to bargain with the union in good faith. The National Labor Relations Act also allowed the parties to require employees to join and financially support the union. However, the Labor-Management Relations Act (the Taft-Hartley Act), which was passed in 1947, granted states the power to pass so-called right-to-work laws that outlaw employment contract provisions requiring employees to join or financially support a union. As such, the laws expose unions to a free rider problem whereby non-union employees benefit from collective bargaining without paying dues. Matsa finds that the ratio of debt to firm value decreases by up to one-half after a right-to-work law is passed.

5 SUMMARY AND CONCLUSION

- In Europe, the area of the economy covered on average by collective bargaining lay in the neighborhood of 65% in the 2000s. It was clearly less in the United States and Japan, where the values are respectively about 15% and 20%. Union

membership is typically lower than coverage, making up 30% of employees in Europe and 12% in the United States, but it is similar to coverage in Japan. Union membership is usually higher in the public sector, as well as in manufacturing and construction.

- The rate of unionization depends on legislation, demographic characteristics, the sectoral makeup of production, and the degree of competition on the product market. The fall in the rates of unionization observed in numerous OECD countries since the 1980s is explained by the evolution of all these factors in combination.
- Unions, or more generally institutions representing wage earners, have as their objective to obtain the highest wage and employment levels possible. Trading off between employment and wages depends on the internal organization of the union and the preferences of workers. Hence, a union made up solely of insiders is indifferent to the level of employment as long as it is sufficiently high for all the insiders to remain in employment. On the contrary, a boss-dominated union seeking to maximize the size of the organization will have as its objective an increase in employment at the expense of wages.
- Models of bargaining derived from the theory of noncooperative games allow us to pinpoint the elements that determine how the gains are shared out between protagonists taking part in a bargaining process. This shareout depends on the preference for the present and the risk aversion of the agents and on the gains they obtain during the unfolding of the negotiation, or when negotiations break off.
- All analyses of wage bargaining agree on the conclusion that the bargaining power of unions drives wages up. Their effect on employment is however ambiguous. Employment decreases with the bargaining power of workers if the bargaining is exclusively over wages, but employment may rise if it also concerns hires. If bargaining covers wages and unemployment benefits or severance payments, the bargaining is strongly efficient and employment always reaches its competitive level.
- The opposition between insiders and outsiders excluded from the bargaining leads to a discrimination between insiders possessing bargaining power, who can on that account obtain good jobs, and workers lacking this power, who are pushed into badly paid jobs.
- Workers' bargaining power has a negative effect on investment if it is impossible to negotiate long-term commitments concerning wages. Once an investment has been made, workers are tempted to push for new wage negotiations in order to benefit from the improved productivity flowing from the increase in capital stock. Without a long-term commitment, the chance that wages will be renegotiated diminishes the return on investment. But the effect on employment of lowered investment is ambiguous: it is positive if labor and capital are gross substitutes and negative if they are gross complements.
- Empirical studies suggest that collective bargaining has a positive impact on wages, while reducing their dispersion. Collective bargaining probably has a positive effect on productivity and a negative effect on profits and investment in physical capital. The effect of collective bargaining on employment proves to be ambiguous.

6 RELATED TOPICS IN THE BOOK

- Chapter 2, section 1.1: The substitution between capital and labor
- Chapter 5, section 4.1: Empirical facts about wage differentials
- Chapter 6, section 2: Risk-sharing
- Chapter 6, section 5: Social preferences
- Chapter 9, section 3.3: Wage bargaining
- Chapter 10, section 2.4: The role of institutions
- Chapter 12, section 2.2.1: What the monopsony model tells us

7 FURTHER READINGS

Booth, A. (1995a). *The economics of the trade union*. Cambridge, U.K.: Cambridge University Press.

DiNardo, J., & Lee, D. (2004). Economic impacts of new unionization on private sector employers: 1984–2001. *Quarterly Journal of Economics*, *119*, 1383–1441.

Freeman, R., & Medoff, J. (1984). *What do unions do?* New York, NY: Basic Books.

Hirsch, B. (2008). Sluggish institutions in a dynamic world: Can unions and industrial competition coexist? *Journal of Economic Perspectives*, *22*, 153–176.

Lee, D., & Mas, A. (2012). Long-run impacts of unions on firms: New evidence from financial markets, 1961–1999. *Quarterly Journal of Economics*, *127*, 333–378.

Osborne, M., & Rubinstein, A. (1990). *Bargaining and markets*. San Diego, CA: Academic Press.

Pencavel, J. (1991). *Labor market under trade unionism, employment, wages and hours*. Cambridge, MA: Blackwell Publishers.

8 APPENDICES

8.1 Unicity of Solution (x^*, y^*)

Consider the system of equations:

$$u_1(y) - \overline{u}_1 = \delta_1 [u_1(x) - \overline{u}_1] \tag{7.33}$$

$$u_2(1-x) - \overline{u}_2 = \delta_2 [u_2(1-y) - \overline{u}_2] \tag{7.34}$$

Relation (7.33) defines y as a function of x, that is, $y = y(x) \equiv u_1^{-1} [\delta_1 (u_1(x) - \overline{u}_1) + \overline{u}_1]$. Let us now define function $H(x)$ by:

$$H(x) = u_2(1-x) - \overline{u}_2 - \delta_2 [u_2(1-y(x)) - \overline{u}_2] \tag{7.35}$$

Since $u_2(0) = \overline{u}_2$, we have $H(1) < 0$ when $x > 0$. Likewise, since (7.33) shows that $y(0) = 0$, (7.35) entails $H(0) = (1 - \delta_2)[u_2(1) - \overline{u}_2] > 0$. Moreover, differentiating (7.35), we get:

$$H'(x) = u_2'(1 - y(x)) \left[\frac{\delta_1 \delta_2 u_1'(x)}{u_1'(y(x))} - \frac{u_2'(1-x)}{u_2'(1-y(x))} \right]$$

Since $y(x) < x$, for $x > 0$ [see (7.34)], the concavity of the utility function entails $u_1'(x)/u_1'(y(x)) < 1$ and $u_2'(1-x)/u_2'(1-y(x)) > 1$. The derivative $H'(x)$ is thus strictly negative for $x > 0$. Therefore, there exists a unique value x^* such that $H(x^*) = 0$. There is thus a unique solution (x^*, y^*), with $y^* = y(x^*)$, for the system of equations (7.33) and (7.34).

8.2 The Correspondence Between the Nash Axiomatic Solution and the Subgame Perfect Equilibrium of Rubinstein's Model

We take up the Rubinstein game presented in section 2.2.1 with the assumption that the players have different preferences for the present. Let Δ be the interval between successive offers; the discount factor of the agents is denoted by $\delta_i = e^{-r_i \Delta}$, $r_i > 0$, $i = 1, 2$, where r_i is the discount rate of player i. We will show that the solution of the bargaining in Rubinstein's game approaches the Nash axiomatic solution when Δ goes to zero.

The solution of the bargaining $(x(\Delta), y(\Delta))$ in Rubinstein's game is defined by the system of equations:

$$u_1[y(\Delta)] - \overline{u}_1 = e^{-r_1 \Delta} [u_1(x(\Delta)) - \overline{u}_1]$$
$$u_2[1 - x(\Delta)] - \overline{u}_2 = e^{-r_2 \Delta} [u_2(1 - y(\Delta)) - \overline{u}_2]$$

In the neighborhood of $\Delta = 0$, we have $e^{-r_i \Delta} \simeq 1 - r_i \Delta$, and these two equations then entail:

$$[u_1(y(\Delta)) - u_1(x(\Delta))] \simeq r_1 \Delta [u_1(x(\Delta)) - \overline{u}_1] \quad \textbf{(7.36)}$$
$$[u_2(1 - x(\Delta)) - u_2(1 - y(\Delta))] \simeq r_2 \Delta [u_2(1 - y(\Delta)) - \overline{u}_2] \quad \textbf{(7.37)}$$

These relations show that $y(\Delta)$ and $x(\Delta)$ converge towards the same value, $\tilde{x}$, when Δ goes to zero. They then entail:

$$u_1'(\tilde{x}) = \lim_{\Delta \to 0} \frac{u_1[y(\Delta)] - u_1[x(\Delta)]}{y(\Delta) - x(\Delta)}$$
$$u_2'(1 - \tilde{x}) = \lim_{\Delta \to 0} \frac{u_2[1 - x(\Delta)] - u_2[1 - y(\Delta)]}{y(\Delta) - x(\Delta)}$$

Using these last two relations and taking the ratio between equations (7.36) and (7.37) for $\Delta \to 0$, we get:

$$\frac{u_1'(\tilde{x})}{u_2'(1 - \tilde{x})} = \frac{r_1}{r_2} \frac{[u_1(\tilde{x}) - \overline{u}_1]}{[u_2(1 - \tilde{x}) - \overline{u}_2]} \quad \textbf{(7.38)}$$

The axiomatic solution of the generalized Nash negotiation, or x^G, is defined by:

$$x^G = \arg\max_x [u_1(x) - d_1]^\gamma [u_2(1-x) - d_2]^{1-\gamma}$$

The first-order condition then entails:

$$\frac{u_1'(x^G)}{u_2'(1-x^G)} = \frac{(1-\gamma)}{\gamma} \frac{u_1(x^G) - d_1}{u_2(1-x^G) - d_2} \tag{7.39}$$

Comparison of equations (7.38) and (7.39) then shows that $x^G = \tilde{x}$ if, and only if, $d_i = \overline{u}_i,\ i = 1, 2$, and $\gamma = r_2/(r_1 + r_2)$.

REFERENCES

Abowd, J. (1989). The effect of wage bargaining on the stock market value of the firm. *American Economic Review*, *79*, 774–800.

Abowd, J., & Farber, H. (1982). Job queues and the union status of workers. *Industrial and Labor Relations Review*, *36*, 354–367.

Abowd, J., & Kramarz, F. (1993). A test of negotiation and incentive compensation models using longitudinal French enterprise data. In J. van Ours, G. Pfann, & G. Ridder (Eds.), *Labour demand and equilibrium wage formation*. Amsterdam: Elsevier Science.

Abowd, J., & Lemieux, T. (1993). The effects of product market competition on collective bargaining agreements: The case of foreign competition in Canada. *Quarterly Journal of Economics*, *108*(4), 983–1014.

Aghion, P., Algan, Y., & Cahuc, P. (2011). Civil society and the state: The interplay between cooperation and minimum wage regulation. *Journal of the European Economic Association*, *9*, 3–42.

Anderson, S., & Devereux, M. (1988). Trade unions and the choice of capital stock. *Scandinavian Journal of Economics*, *90*(1), 27–44.

Arrow, K. (1963). *Social choice and individual values*. New Haven, CT: Yale University Press.

Ashenfelter, O. (1987). Arbitration and the negotiation process. *American Economic Review, Papers and Proceedings*, *77*, 342–346.

Ashenfelter, O., & Brown, J. (1986). Testing the efficiency of employment contracts. *Journal of Political Economy*, *94*, 40–87.

Ashenfelter, O., Currie, J., Farber, H., & Spiegel, M. (1992). An experimental comparison of dispute rates in alternative arbitration systems. *Econometrica*, *60*, 1407–1433.

Ashenfelter, O., & Hyslop, D. (2001). Measuring the effect of arbitration on wage levels: The case of police officers. *Industrial and Labor Relations Review*, *54*, 316–328.

Atherton, W. (1973). *Theory of union bargaining goals*. Princeton, NJ: Princeton University Press.

Atkinson, A., & Micklewright, J. (1991). Unemployment compensation and labor market transitions: A critical review. *Journal of Economic Literature, 39*, 1679–1727.

Berninghaus, S., Güth, W., & Schosser, S. (2012). Backward induction or forward reasoning? An experiment of stochastic alternating offer bargaining (Jena Economic Research Papers in Economics No. 2012, 401).

Binmore, K., Rubinstein, A., & Wolinsky, A. (1986). The Nash solution in economic modelling. *Rand Journal of Economics*, *17*(2), 176–188.

Blair, D., & Crawford, D. (1984). Labor union objectives and collective bargaining. *Quarterly Journal of Economics*, *99*, 547–566.

Blanchflower, D. (1996). The role and influence of trade unions in the OECD (CEP Discussion Paper No. 0310).

Blanchflower, D., & Bryson, A. (2003). Changes over time in union relative wage effects in the UK and US revisited. In J. Addison & C. Schnabel (Eds.), *International handbook of trade unions*. Cheltenham: Edward Elgar.

Blanchflower, D., & Freeman, R. (1992). Unionism in the U.S. and in other advanced O.E.C.D. countries. *Industrial Relations*, *31*, 56–79.

Blanchflower, D., & Freeman, R. (1994). Did the Thatcher reforms change British labour performance? In R. Barell (Ed.), *The UK labour market: Comparative aspects and institutional developments* (pp. 51–72). Cambridge, U.K.: Cambridge University Press.

Blau, F., & Kahn, L. (1996). International differences in male wage inequality: Institution versus market forces. *Journal of Political Economy*, *104*, 791–837.

Blau, F., & Kahn, L. (1999). Institutions and laws in the labor market. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 3A, chap. 25, pp. 1399–1461). Amsterdam: Elsevier Science.

Boal, W., & Pencavel, J. (1994). The effects of labor unions on employment, wages and day of operation: Coal mining in West Virginia. *Quarterly Journal of Economics*, *109*, 267–298.

Bökerman, P., & Uusitalo, R. (2006). Erosion of the Ghent system and union membership decline: Lessons from Finland. *British Journal of Industrial Relations, London School of Economics*, *44*, 283–303.

Booth, A. (1984). A public choice model of trade union behaviour and membership. *Economic Journal*, *94*, 883–898.

Booth, A. (1995a). *The economics of the trade union*. Cambridge, U.K.: Cambridge University Press.

Booth, A. (1995b). Layoffs with payoffs: A bargaining model of union wage and severance pay determination. *Economica*, *62*, 551–564.

Booth, A., & Ravaillon, M. (1993). Employment and the length of the working week in a unionized economy in which hours of work influence productivity. *Economic Record*, *69*, 428–436.

Bronars, S., Deere, D., & Tracy, J. (1991). The threat of unionization, the use of debt, and the preservation of shareholder wealth. *Quarterly Journal of Economics*, *106*, 231–254.

Bronars, S., Deere, D., & Tracy, J. (1994). The effects of unions on firm behavior: An empirical analysis using firm-level data. *Industrial Relations*, *33*, 426–451.

Cahuc, P., & Kramarz, F. (1997). Voice and loyalty as a delegation of authority: A model and a test on a panel of French firms. *Journal of Labor Economics*, *15*(4), 658–688.

Cahuc, P., & Zylberberg, A. (2008). Working time and employment. In T. Boeri, M. Burda, & F. Kramarz (Eds.), *Working hours and job sharing in the EU and USA: Are Europeans lazy? Or Americans crazy?* (pp. 113–140). Oxford, U.K.: Oxford University Press.

Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton, NJ: Princeton University Press.

Card, D. (1986). Efficient contracts with costly adjustment: Short-run employment determination for airline mechanics. *American Economic Review*, *76*, 1045–1071.

Card, D. (1990). Unexpected inflation, real wages, and employment determination in union contracts. *American Economic Review*, *80*, 669–688.

Card, D. (1996). The effect of unions on the structure of wages: A longitudinal analysis. *Econometrica*, *64*, 957–979.

Card, D. (2001). The effect of unions on wage inequality in the U.S. labor market. *Industrial and Labor Relations Review*, *54*, 296–315.

Card, D., Lemieux, T., & Riddell, W. (2004). Unions and wage inequality. *Journal of Labor Research*, *25*, 519–559.

Card, D., & de la Rica, S. (2006). The effect of firm-level contracts on the structure of wages: Evidence from matched employer-employee data. *Industrial and Labor Relations Review*, *59*(4), 573–592.

Carruth, A., & Oswald, A. (1985). Miners' wages in post-war Britain: An application of a model of trade union behaviour. *Economic Journal*, *95*, 1003–1020.

Carruth, A., Oswald, A., & Findlay, L. (1986). A test of a model of trade union behaviour: The coal and steel industry in Britain. *Oxford Bulletin of Economics and Statistics*, *48*, 1–18.

Charness, G., & Kuhn, P. (2011). Lab labor: What can labor economists learn from the lab? In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 4A, chap. 3, pp. 229–330). Amsterdam: Elsevier Science.

Checchi, D., & Lucifora, C. (2002). Unions and labour market institutions in Europe. *Economic Policy*, *35*, 363–408.

Clark, A., & Oswald, A. (1994). Unhappiness and unemployment. *Economic Journal*, *104*, 648–669.

Connolly, R., Hirsch, B., & Hirschey, M. (1986). Union rent seeking, tangible capital and market value of the firm. *Review of Economics and Statistics*, *68*, 567–577.

Contensou, F., & Vranceanu, R. (2000). *Working time: Theory and policy implications*. Cheltenham, U.K.: Edward Elgar.

Cramton, P., Gunderson, M., & Tracy, J. (1999). The effect of collective bargaining legislation on strikes and wages. *Review of Economics and Statistics*, *81*, 475–487.

Cramton, P., & Tracy, J. (1992). Strikes and holdouts in wage bargaining: Theory and data. *American Economic Review*, *82*, 1200–1210.

Dell'Aringa, C., & Lucifora, C. (1994). Wage dispersion and unionism: Do unions protect low pay? *International Journal of Manpower*, *15*, 150–169.

De Menil, G. (1971). *Bargaining: Monopoly power versus union power*. Cambridge, MA: MIT Press.

Denny, K., & Nickell, S. (1992). Unions and investment in British industry. *Economic Journal*, *102*, 874–887.

Dertouzos, J., & Pencavel, J. (1981). Wage and employment determination under trade unionism: The case of the International Typographical Union. *Journal of Political Economy*, *89*, 1162–1181.

Devereux, M., & Lockwood, B. (1991). Trade unions, nonbinding wage agreements, and capital accumulation. *European Economic Review*, *35*, 1411–1426.

DiNardo, J., Fortin, N., & Lemieux, T. (1996). Labor market institutions and the distribution of wages, 1973–1992: A semi-parametric approach. *Econometrica*, *64*, 1001–1044.

DiNardo, J., & Lee, D. (2004). Economic impacts of new unionization on private sector employers: 1984–2001. *Quarterly Journal of Economics*, *119*, 1383–1441.

Dinlersoz, E., & Greenwood, J. (2012). The rise and fall of unions in the U.S. (NBER Working Paper No. 18079).

Donado, A., & Wälde, K. (2012). How trade unions increase welfare. *Economic Journal*, *122*(563), 990–1009.

Doucouliagos, C., & Laroche, P. (2003). What do unions do to productivity? A metaanalysis. *Industrial Relations*, *42*, 650–691.

Drèze, J., & Modigliani, F. (1981). The trade-off between real wage and employment in an open economy. *European Economic Review*, *15*, 1–40.

Duncan, G., & Stafford, F. (1980). Do union members receive compensating wage differential? *American Economic Review*, *70*, 355–371.

Dunlop, J. (1944). *Wage determination under trade unions*. New York, NY: Macmillan.

Dustmann, C., & Schonberg, U. (2009). Training and union wages. *Review of Economics and Statistics*, *91*, 363–376.

Edgeworth, F. (1881). *Mathematical psychics*. London: Kegan Paul.

Ellwood, D., & Fine, G. (1987). The impact of right-to-work laws on union organizing. *Journal of Political Economy*, *95*, 250–273.

Espinosa, P., & Rhee, C. (1989). Efficient wage bargaining as a repeated game. *Quarterly Journal of Economics*, *104*, 565–588.

Fallick, B., & Hassett, K. (1999). Investment and union certification. *Journal of Labor Economics*, *17*, 570–582.

Farber, H. (1978). Individual preferences and union wage determination: The case of the United Mine Workers. *Journal of Political Economy*, *86*, 923–942.

Farber, H. (1983). The determination of the union status of workers. *Econometrica*, *51*, 1417–1438.

Farber, H. (1986). The analysis of union behavior. In O. Ashenfelter & R. Layard (Eds.), *Handbook of labor economics* (vol. 2, pp. 1139–1189). Amsterdam: Elsevier Science.

Farber, H., & Bazerman, M. (1986). The general basis for arbitrator behavior: An empirical analysis of conventional and final-offer arbitration. *Econometrica*, *54*, 1503–1528.

Farber, H., & Saks, D. (1980). Why workers want unions: The role of relative wages and job characteristics. *Journal of Political Economy*, *88*, 349–369.

Fehr, E. (1990). Cooperation, harassment and involuntary unemployment: Comment. *American Economic Review*, *80*, 624–630.

Fernandez, R., & Glazer, J. (1991). Striking for a bargain between two completely informed agents. *American Economic Review*, *81*, 240–252.

Fortin, N., & Lemieux, T. (1997). Institutional changes and rising wage inequality: Is there a linkage? *Journal of Economic Perspectives*, *11*, 75–96.

Frandsen, B. (2012). Why unions still matter: The effects of unionization on the distribution of employee earnings. Massachusetts Institute of Technology manuscript.

Freeman, R. (1980). Unionism and the dispersion of wages. *Industrial and Labor Relations Review*, *34*, 3–23.

Freeman, R., & Medoff, J. (1984). *What do unions do?* New York, NY: Basic Books. Traduction française, *Pourquoi les syndicats? Une réponse américaine*. Paris: Economica, 1987.

Grout, P. (1984). Investment and wages in the absence of binding contracts. *Econometrica*, *52*, 449–460.

Hartog, J., & Theeuwes, J. (Eds.). (1993). *Labour market contracts and institutions: A cross national comparison*. Amsterdam: North Holland.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, *5*(4), 475–492.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*, 153–161.

Hicks, J. (1932). *The theory of wages* (2nd ed. 1963). New York, NY: Macmillan.

Hirsch, B. (1992). Firm investment behavior and collective strategy. *Industrial Relations*, *31*, 95–121.

Hirsch, B. (2004). Reconsidering union wage effects: Surveying new evidence on an old topic. *Journal of Labor Research*, *25*(2), 233–266.

Hirsch, B. (2007). What do unions do for economic performance? In J. T. Bennett & B. E. Kaufman (Eds.), *What do unions do? A twenty-year perspective* (pp. 193–237). New Brunswick, NJ: Transaction Publishers.

Hirsch, B. (2008). Sluggish institutions in a dynamic world: Can unions and industrial competition coexist? *Journal of Economic Perspectives*, *22*, 153–176.

Hirsch, B., & Addison, J. (1986). *The economic analysis of unions, new approaches and evidence*. Boston, MA: Allen and Unwin.

Hirsch, B., & Macpherson, D. (2003). Union membership and coverage database from the current population survey: Note. *Industrial and Labor Relations Review*, *56*(2), 349–354.

Hirschman, A. (1970). *Exit, voice and loyalty*. Cambridge, MA: Harvard University Press.

Kahn, L. (1998). Collective bargaining and the interindustry wage structure. *Economica*, *65*, 507–534.

Kahn, L. (2000). Wage inequality, collective bargaining and relative employment from 1985 to 1994: Evidence from fifteen OECD countries. *Review of Economics and Statistics*, *82*, 564–579.

Kalai, E., & Smorodinsky, M. (1975). Other solutions to Nash's bargaining problem. *Econometrica*, *43*, 513–518.

Kennan, J., & Wilson, R. (1993). Bargaining with private information. *Journal of Economic Literature*, *31*, 45–104.

Kiander, J. (1993). Endogenous unemployment insurance in a monopoly union model when job search matters. *Journal of Public Economics*, *52*, 101–115.

Kuhn, P., & Gu, W. (1999). Learning in sequential wage negotiations: Theory and evidence. *Journal of Labor Economics*, *17*(1), 109–140.

Layard, R., Nickell, S., & Jackman, R. (1991). *Unemployment*. Oxford, U.K.: Oxford University Press.

Lee, D., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, *48*, 281–355.

Lee, D., & Mas, A. (2012). Long-run impacts of unions on firms: New evidence from financial markets, 1961–1999. *Quarterly Journal of Economics*, *127*, 333–378.

Lee, L. (1978). Unionism and wage rates: A simultaneous equations model with qualitative and limited dependent variables. *International Economic Review*, *19*, 415–434.

Lemieux, T. (1998). Estimating the effect of unions on wage inequality in a panel data model with comparative advantage and nonrandom selection. *Journal of Labor Economics*, *16*, 261–291.

Leontief, W. (1946). The pure theory of the guaranteed annual wage contract. *Journal of Political Economy*, *54*, 76–79.

Lewis, H. (1963). *Unionism and relative wages in the United States: An empirical inquiry*. Chicago, IL: University of Chicago Press.

Lewis, H. (1986). Union relative wage effects. In O. Ashenfelter & R. Layard (Eds.), *Handbook of labor economics* (vol. 2). Amsterdam: North Holland.

Lindbeck, A., & Snower, D. (1988). Cooperation, harassment and involuntary unemployment: An insider-outsider approach. *American Economic Review*, *78*, 167–188.

Linneman, P., Wachter, M., & Carter, W. (1990). Evaluating the evidence on union employment and wages. *Industrial and Labor Relations Review*, *44*, 34–53.

MacDonald, I., & Solow, R. (1981). Wage bargaining and employment. *American Economic Review*, *71*(5), 896–908.

MaCurdy, T., & Pencavel, J. (1986). Testing between competing models of wage and employment determination in unionised markets. *Journal of Political Economy*, *94*, supplement, 513–539.

Maloney, T. (1994). Estimating the effects of the employment contracts act on employment and wages in New Zealand. *Australian Bulletin of Labour*, *20*, 320–343.

Manning, A. (1987). An integration of trade union models in a sequential bargaining framework. *Economic Journal*, *97*, 121–139.

Martin, D. (1980). *An ownership theory of the trade union*. Berkeley and Los Angeles: University of California Press.

Matsa, D. (2010). Capital structure as a strategic variable: Evidence from collective bargaining. *Journal of Finance*, *65*, 1197–1232.

Metcalf, D., Hansen, K., & Charlwood, A. (2001). Unions and the sword of justice: Unions and pay systems, pay inequality, pay discrimination and low pay. *National Institute Economic Review*, *176*, 61–75.

Moore, W. (1998). The determinants and effects of right-to-work laws: A review of the recent literature. *Journal of Labor Research*, *19*, 445–469.

Nash, J. (1950). The bargaining problem. *Econometrica*, *18*, 155–162.

Nash, J. (1953). Two-person cooperative game. *Econometrica*, *21*(1), 128–140.

Nickell, S., & Andrews, M. (1983). Unions, real wage and employment in Britain 1951–79. *Oxford Economic Papers*, *35*, supplement, 183–206.

Nickell, S., & Wadhwani, S. (1990). Insider forces and wage determination. *Economic Journal*, *100*, 496–509.

Ochs, J., & Roth, A. (1989). An experimental study of sequential bargaining. *American Economic Review*, *79*, 355–384.

Odgers, C., & Betts, J. (1997). Do unions reduce investment? *Industrial and Labor Relations Review*, *51*, 18–36.

Olson, M. (1965). *The logic of collective action: Public goods and the theory of groups*. Cambridge, MA: Harvard University Press.

Osborne, M., & Rubinstein, A. (1990). *Bargaining and markets*. Waltham, MA: Academic Press.

Oswald, A. (1982). The microeconomic theory of the trade union. *Economic Journal*, *92*, 576–595.

Pencavel, J. (1984). The empirical performance of a model of trade union behavior. In J.-J. Rosa (Ed.), *The economics of trade unions: New directions*. Boston, MA: Kluwer-Nijhoff Publishing.

Pencavel, J. (1991). *Labor market under trade unionism, employment, wages and hours*. Cambridge, MA: Blackwell Publishers.

Robinson, C. (1989). The joint determination of union status and union wage effects: Some tests of alternative models. *Journal of Political Economy*, *97*, 639–667.

Rosen, S. (1970). Unionism and the occupational wage structure in the United States. *International Economic Review*, *11*, 269–286.

Ross, A. (1948). *Trade union wage policy*. Berkeley and Los Angeles: University of California Press.

Rowthorn, R. (1992). Centralisation, employment and wage dispersion. *Economic Journal*, *102*, 506–523.

Ruback, S., & Zimmerman, M. (1984). Unionization and profitability: Evidence from the capital market. *Journal of Political Economy*, *92*, 1134–1157.

Rubinstein, A. (1982). Perfect equilibrium in a bargaining model. *Econometrica*, *50*, 97–109.

Slaughter, M. (2007). Globalization and declining unionization in the United States. *Industrial Relations*, *46*, 329–346.

Sojourner, A., Town, R., Grabowski, D., & Chen, M. (2012). Impacts of unionization on employment, product quality and productivity: Regression discontinuity evidence from nursing homes (NBER Working Paper No. 17733).

Stahl, I. (1972). *Bargaining theory*. Economic Research Institute, Stockholm School of Economics.

Strand, J. (1989). Monopoly unions versus efficient bargaining, a repeated game approach. *European Journal of Political Economy*, *5*, 473–486.

Tan, H., Chapman, B., Peterson, C., & Booth, A. (1992). Youth training in the United States, Britain and Australia. *Research in Labor Economics*, *13*, 63–99.

Tracy, J. (1986). An investigation into the determinants of U.S. strike activity. *American Economic Review*, *76*, 423–436.

van der Ploeg, F. (1987). Trade unions, investment and employment, a non-cooperative approach. *European Economic Review*, *31*, 1465–1492.

Van Reenen, J. (1996). The creation and capture of economic rents: Wages and innovations in UK manufacturing plants. *Quarterly Journal of Economics*, *111*, 195–226.

Zeuthen, F. (1930). *Problems of monopoly*. London: Routledge.

CHAPTER 8

Discrimination

In this chapter we will:

- Find out why a "taste for discrimination" and monopsony power or labor market frictions may give rise to situations of discrimination, in which some persons obtain lower wages than others because of their membership in particular demographic groups
- Learn what statistical discrimination is and why it can lead to persistent inequalities among demographic groups
- Study the reach and the limitations of the different methods of estimating discrimination
- Apply these methods to the black–white wage gap in the United States, relying on the contributions of Neal and Johnson (1996) and Lang and Manove (2011); and apply them to the gender gap based on the contribution of O'Neill and O'Neill (2006) (The main results of the contributions presented in this chapter can be replicated with data and programs available at www.labor-economics.org.)
- Assess the extent of discrimination against nonwhites, women, and homosexuals
- Show that "better-looking" individuals have higher wages
- Find that empirical work reveals that discrimination does explain a part of the wage differences among demographic groups but does not account for the totality of these differences

INTRODUCTION

Discrimination is a situation in which individuals identical as regards their productive ability are treated differently because of certain of their nonproductive characteristics such as race, gender, or sexual orientation. When it obtains, exchanges on the labor market depend on these nonproductive characteristics. The career paths of two persons

whose productive characteristics are identical may thus differ because of the race or gender of one of them, and wage gaps no longer reflect compensating differences. Such situations come about when, for example, a diner in a restaurant prefers a waiter of the same skin color as herself (or a different one), or a worker prefers to collaborate with colleagues having the same sexual orientation, or the head of a firm would rather hire men than women (or the converse). A person may also be discriminated against if an employer offers him (or her) a low wage in the belief (mistaken or valid) that he (or she) is less efficient on account of belonging to a demographic group (women, men, whites, nonwhites) whose members collectively are thought to be, on average, less efficient.

The first section of this chapter supplies quantified indications about the main differences in the area of wages and employment among different demographic groups. According to these data, it is women whose situation is the most unfavorable on average in the OECD countries. But it also emerges that the gender wage gap has narrowed over the last three decades. This trend is similar to that observed in employment. We will see that large gaps in labor market outcomes in the OECD countries are also observed among nonwhites (or other ethnic groups) compared to whites and among children of immigrants compared with children of natives.

The second section presents theories that allow us to account for discrimination on the labor market. The earliest is the theory of "taste discrimination" advanced by Becker (1957). It shows that if employers or employees experience an aversion for a certain group of workers, the equilibrium wage of that group is lower than that of other workers collectively who present identical productive characteristics. But such discrimination cannot persist under perfect competition, as employers with no preference drive employers with discriminatory preferences out of the market, offering all workers equal wages. Conversely, the presence of a monopsonist firm or of search frictions in the labor market might explain the persistence of such discrimination. Another theory labeled "statistical discrimination" starts from the idea that employers think that membership in a given group yields an a priori estimate of individual productivity. If so, a "bad" a priori estimate can become self-fulfilling and discrimination may become persistent.

Methods of estimating wage discrimination are set out in the third section. They all aim to answer a simple question: do the wage gaps observed between certain demographic groups reflect a pure phenomenon of discrimination, or do there exist other characteristics (skills, education, nonobserved productivity, etc.) that can explain all or part of these wage gaps? We will see that estimations of wage equations, the most widely adopted method, struggle to identify discrimination. This is why recent research has turned to laboratory experiments and field experiments.

The fourth section brings together the main results acquired to date on phenomena of discrimination that affect the labor market. Apart from wage discrimination, forms of nonwage discriminations that influence the hiring process also play an important part. We will see that discrimination may be based on gender, race, ethnic background, and religion, as well as sexual orientation and the physical aspect ("beauty") of persons.

The fifth and last section gives an overview of policies to combat discrimination, while stressing that premarket factors such as cognitive or noncognitive skills, psychological attributes, and social norms explain the largest part of the inequality that exists in the labor market.

1 SOME FACTS ABOUT WAGE AND EMPLOYMENT DIFFERENCES

The labor market can generate marked differences in wages and the level of employment but also in the level and duration of unemployment across various demographic groups. In this section we present some key differences in wages and access to employment across gender and racial/ethnic groups, although there are also differences that depend on other demographic characteristics such as sexual orientation or the way people look; these are covered later in this chapter.

1.1 Women Versus Men

The largest demographic group that experiences both lower wages and lower employment levels is of course women. Since the number of hours worked may vary across groups and influence earnings, comparisons in the area of gender should focus on either hourly earnings or annual earnings for full-time workers.[1] On average in the OECD countries, full-time female employees have wages 16% lower than full-time male employees in 2010. The difference was 20% ten years earlier in 2000. Figure 8.1 shows that the wage gap is lowest in Mexico, Hungary, New Zealand, and Norway (below 7%) and it is highest in Israel, Germany, Japan, and Korea (above 20%). This gap usually increases with earnings. At the top 10% of income distribution, it is greater than 20% in most countries and even exceeds 30% in Korea and Japan, while at the bottom 10% of income distribution it is at or below 10% in many countries. Larger wage gaps at the top deciles of income distribution may suggest the existence of a "glass ceiling" that blocks women from accessing the best-paid jobs in their sector. Indeed, figure 8.2 shows that approximately a third or less of positions with managerial responsibilities were staffed by women in 2007. Actually, the existence of relatively small wage gaps in some countries does not in itself constitute proof of an absence of discrimination, that is, proof that women with skills comparable to those of men would get the same wage. As we will see in this chapter, labor market outcomes are the result of a selection process that can be harsh and cause only the most productive women to remain in the market, especially at the top end of the distribution (see also OECD, 2012, for more details on the gender gap). Similarly, large wage gaps at the low end of the distribution might stem from a restricted access to the market for some women because of a lack of affordable child care.

Another stylized fact is that the wage gap is growing with age, across cohorts, and within cohorts. As shown in figure 8.3, in all countries older women earn much lower wages than men of the same age, compared with younger women and younger men. As we will see later, this might be due to several factors, including the closing of the educational gap between men and women over the last decades, which results

[1] Of course the two measures might yield different results due to selection effects: those working full-time all year long might have characteristics influencing their hourly wage that set them apart from individuals working fewer hours over the year.

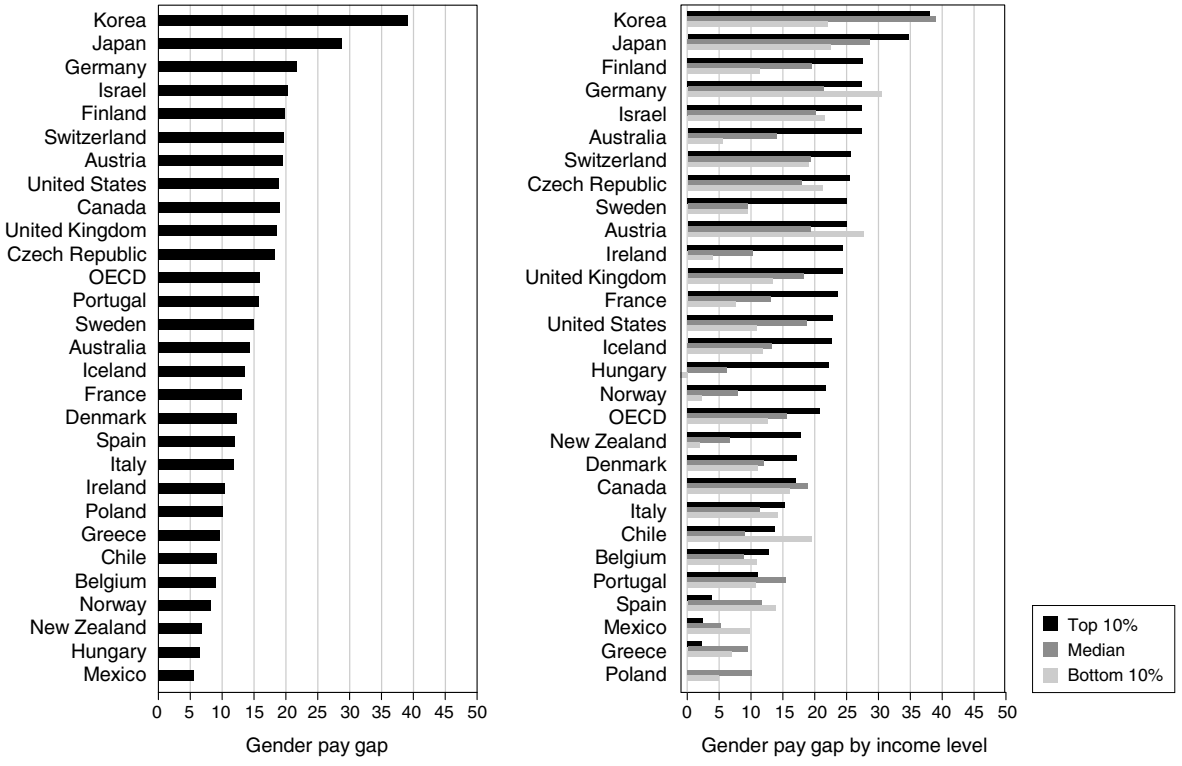


FIGURE 8.1

The gender pay gap in 2010 (percentage). The wage gap concerns full-time employees. It is defined as the difference between male and female median wages divided by male median wages.

Source: OECD (2012, figure 13.1).

in greater wage disparity between older women and men of their own generation than that experienced by younger women because they are less educated than them. But this deepening of the wage gap is also true over the life cycle: for the same individuals, the wage gap increases as women get older. This might be explained by differences in experience between the genders in the labor market, which increases with age, notably due to the family responsibilities of women.

We can see in figure 8.4 that the gender wage gap has narrowed over the last three decades. The gap shrank more markedly from the 1980s to the mid-1990s than it did in the late 1990s and 2000s. In the 1980s, countries where the wage gap was the highest tended to catch up and experienced a higher annual average reduction in wage differences. This trend is similar to that observed in employment. The employment rate of working-age (15–64) women was 57% in 2010 in the OECD countries, still 16 percentage points lower than that of men (see figure 8.5). But this employment rate increased after 2000, although at a slower pace than in the 1980s (see figure 8.6). The gap is now lowest in the northern European countries, higher in the southern European countries, and much higher in South America, Asia, and the Middle East. This trend went along with the strong feminization of the service sector: in the OECD in 2010 about 80% of employed women worked in the service sector (60% for men).

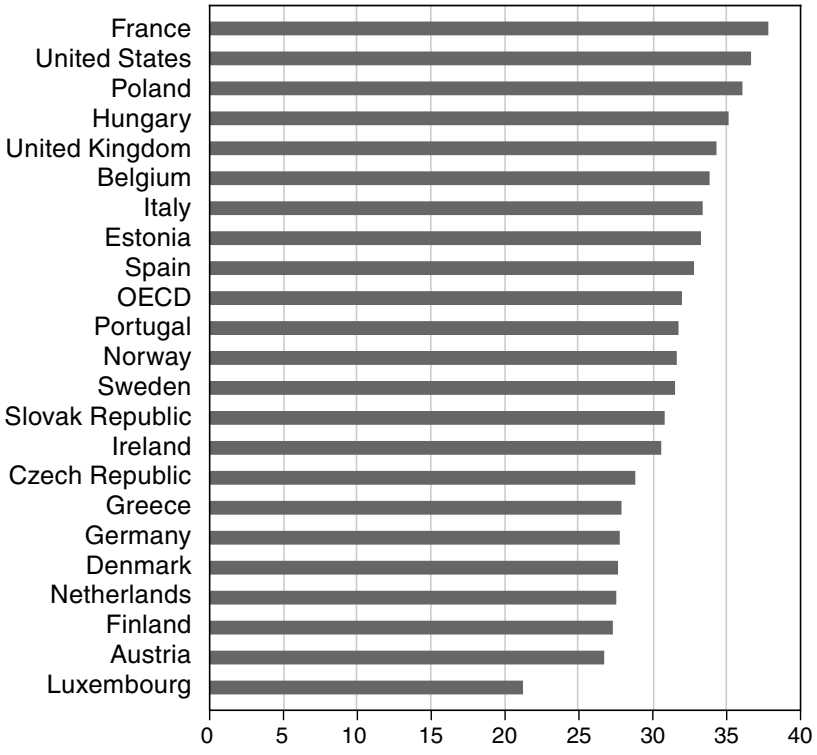


Figure 8.2
Proportion of women among staff with managerial responsibilities (percentage), 2007.

Source: OECD (2012, figure 11.5).

1.2 Gaps Between Ethnic Groups

Large gaps in labor market outcomes in the OECD countries are also observed among nonwhites compared to whites, or among children of immigrants compared to children of natives. These categories are often less easy to identify in administrative or survey data because in many countries the "race" or the color of the skin is not recorded at all, or when it is recorded it is with a degree of imprecision based on self declarations (see Charles and Guryan, 2011, on the potential consequences of this taxonomical challenge). In countries where statistics are available, nonwhites earn about 15% less than whites. In the United States, for instance, nonwhites overall earn about 17% less than whites, and blacks earn approximately 25% less. But the racial wage gap is also significant in Canada and in the United Kingdom (see figure 8.7). Gaps of a similar size can also be observed across ethnic groups. For instance, in a number of countries where immigration has been strong, children of immigrants have less access to employment than children of native-born parents (see figure 8.8). In the case of Canada, selection effects (immigrants are selected on the basis of skills) could explain why the employment gap is negative.

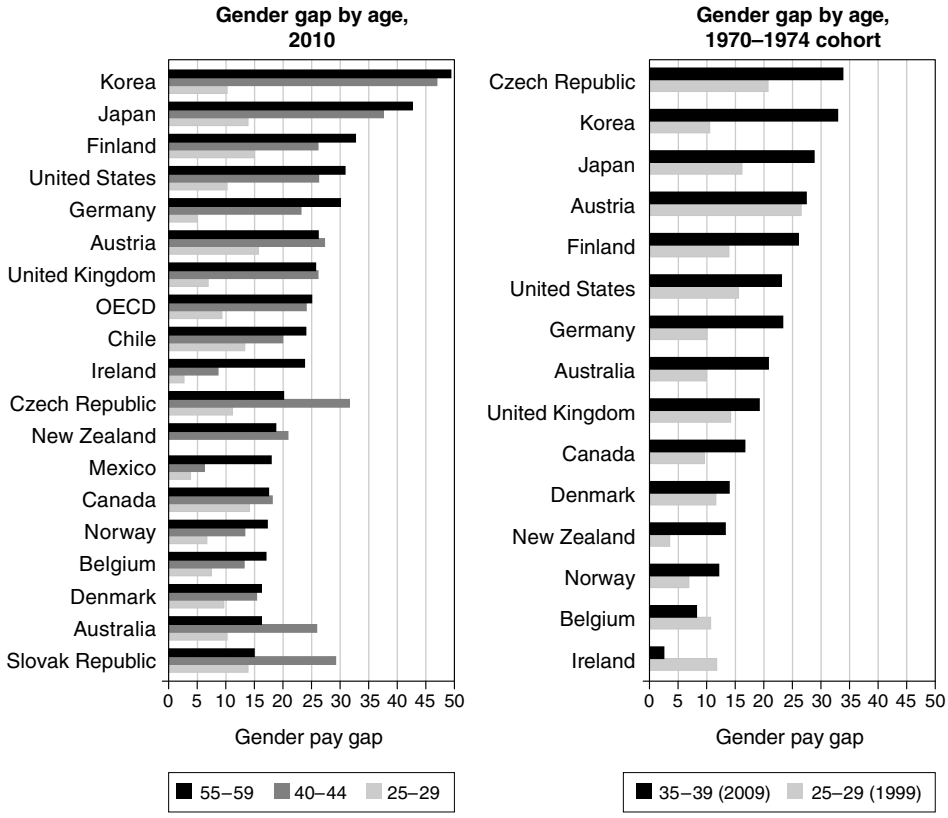


FIGURE 8.3

The gender pay gap by age (percentage). The gender gap concerns full-time employees. It is defined as the difference between male and female mean wages divided by male mean wages.

Source: OECD (2012, figure 13.2).

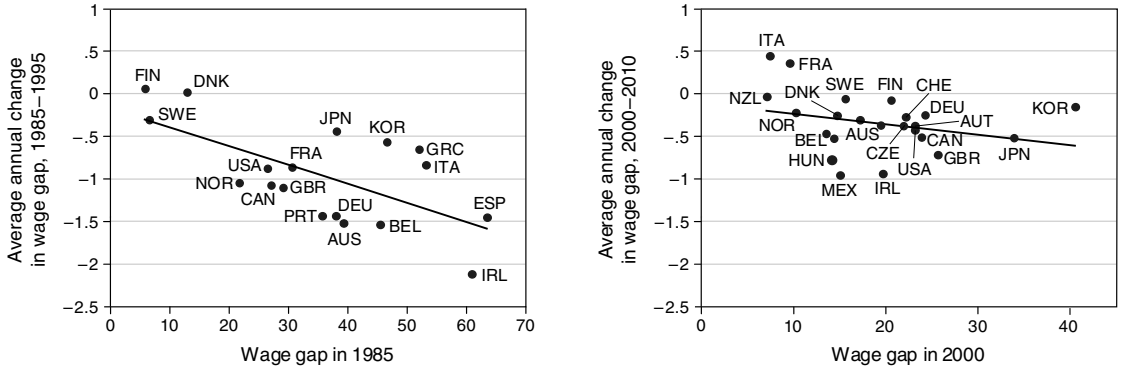


FIGURE 8.4

The evolution of the gender wage gap in OECD countries in the 1980s and the 2000s (percentage). The wage gap concerns full-time employees. It is defined as the difference between male and female median wages divided by male median wages. Years around 1985, 1995, 2000, and 2010.

Source: OECD (2008, 2012).

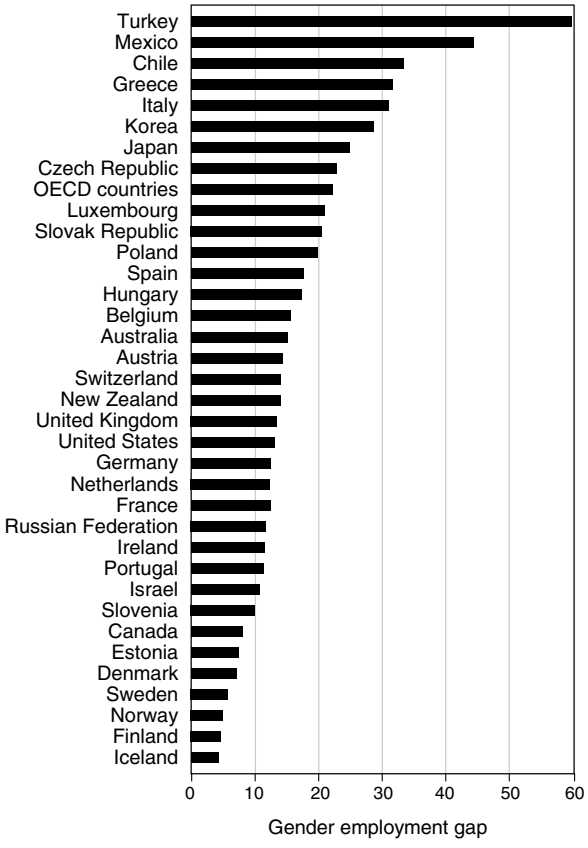


Figure 8.5

The gender employment gap in OECD countries in 2010 (percentage). The employment gap is defined as the difference between male and female employment rates as a percentage of the male employment rate.

Source: OECD Labor Force Statistics database.

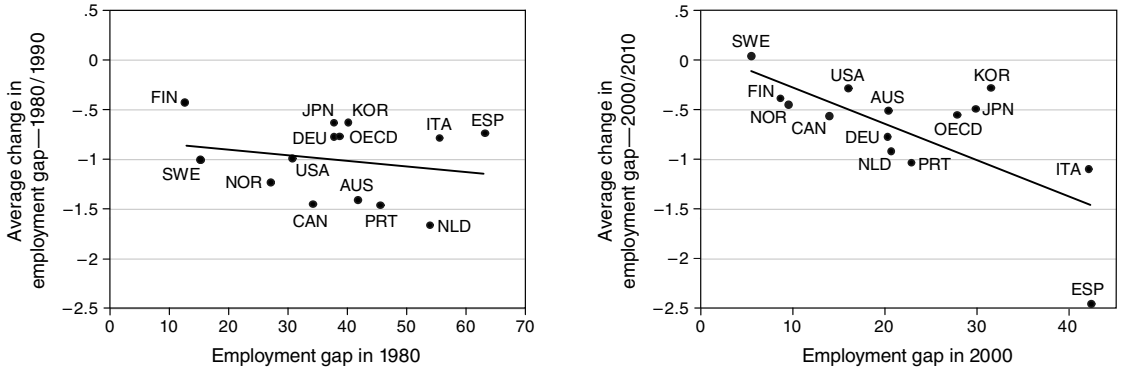


Figure 8.6

The evolution of gender employment gaps in OECD countries (percentage). The employment gap is defined as the difference between male and female employment rates as a percentage of the male employment rate.

Source: OECD Labor Force Statistics database.

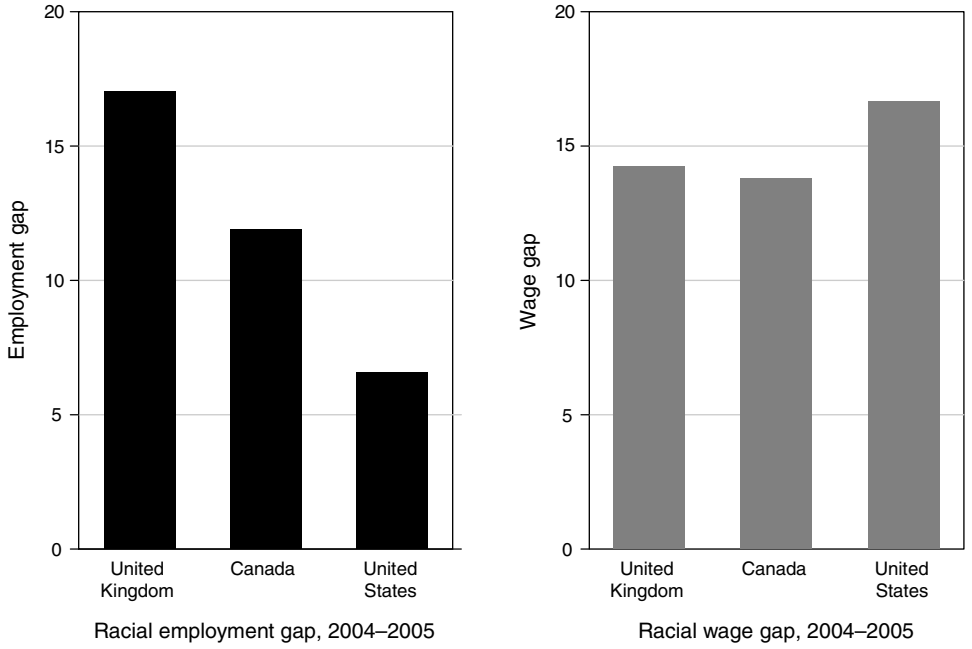


Figure 8.7
The racial wage gap and employment gap, 2004–2005 (percentage). The employment gap is defined as the difference between white and nonwhite employment rates as a percentage of the white employment rate. The wage gap is defined here as the difference between median white and nonwhite hourly wages as a percentage of the median white wage. Data refer to 2004 for Canada, and 2005 for the employment and wage gaps between white and nonwhite groups in the United Kingdom and the United States.

Source: OECD (2008, figure 3.4).

A closing of the racial wage gap has been observed in the United States since the civil rights laws prohibiting discrimination were passed in the 1960s (see figure 8.9), although at a slower pace since the beginning of the 2000s—a trend quite similar, viewed from a distance, to that observed across gender. The observed convergence in earnings does not necessarily mean that discrimination receded. We cannot rule out the possibility that wage convergence reflects changes in the distribution of who is employed within each racial group. Before concluding anything about potential discrimination, we would need to control for differences in skills across groups and over time (see section 3). Besides, the employment–unemployment gap did not recede in the same way. Since the 1960s, the unemployment rate of blacks has been more than twice as large as that of whites, and this was still the case in 2012 (see table 8.1). The unemployment duration of black men was roughly 30% longer than that of white men in the 2000s (Lang and Lehmann, 2012). Between the late 1980s and 2000, when there was strong wage convergence, the unemployment rate ratio between blacks and whites fluctuated around its mean. Considering all men over 20, the employment rate was 68.4% among whites in July 2013 compared with 59.2% for blacks (Bureau of Labor Statistics), approximately a 10-point difference that has been more or less stable over recent decades.

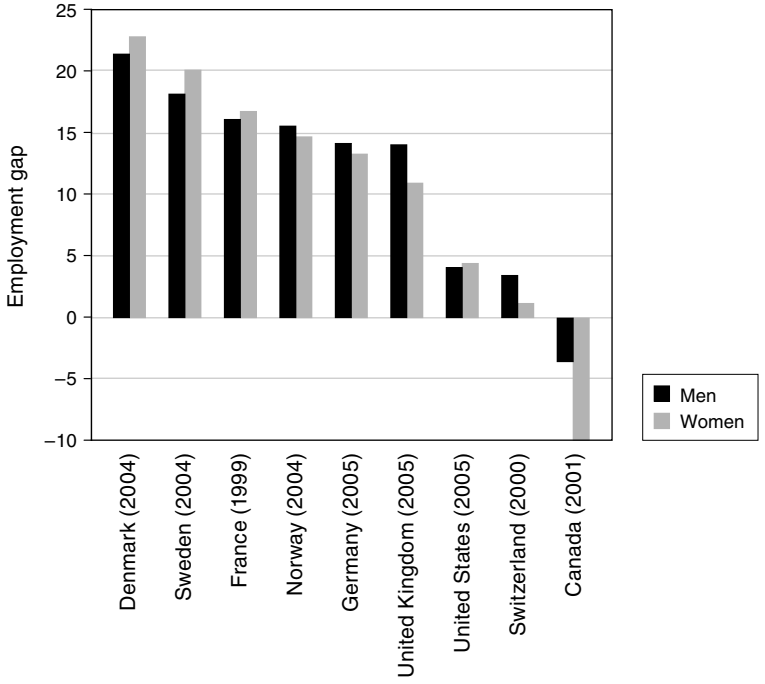


Figure 8.8
Employment rate gaps between the second-generation and native-born with no migration background, persons aged 20 to 29 years and not in education (percentage). The employment gap is defined as the difference between the native-born and second-generation employment rates as a percentage of the native-born employment rate.

Source: OECD (2008, figure 3.5).

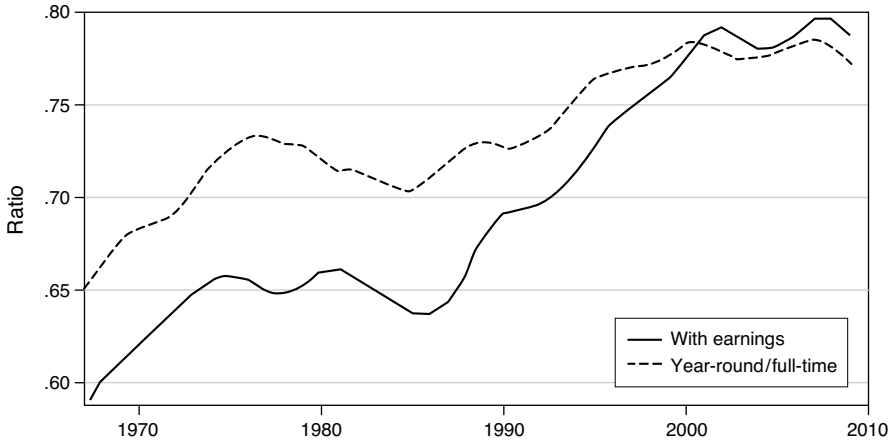


Figure 8.9
Ratio of median annual earnings—black men/white men, 1967–2009, based on the March Current Population Survey; men aged at least 20 working full-time/year-round, at least 35 hours per week and 50 weeks a year.

Source: Lang and Lehmann (2012, figure 1).

TABLE 8.1
Racial unemployment differences among men aged 16 and older in the United States in 2012.

	Unemployment rate	
	Annual average	Ratio with whites
Whites	7.4	1.0
Black or African American	15	2.0
Asian	5.8	0.8
Hispanic or Latino ethnicity	9.9	1.4

Source: Bureau of Labor Statistics, Labor Force Statistics from the Current Population Survey, annual averages, table 24.

2 THEORIES OF DISCRIMINATION

Becker (1957) pointed out that an aversion felt by employers, clients, or other workers toward persons belonging to certain groups may constitute a source of discrimination and lead to lower wages for discriminated workers. This model is often referred to as "taste discrimination." An alternative theory starts from the idea that employers think that membership in a given group sends a signal about the individual's productivity, a situation leading to "statistical discrimination" among groups. In all these theories, wage gaps no longer reflect compensating differentials.

2.1 TASTE DISCRIMINATION

Becker (1957) suggested that discrimination might arise from the fact that some employers feel a disinclination to hire workers belonging to certain groups. He presented this hypothesis in formal terms by assuming that the gains these employers derive from employing workers include the profit of the firm and some taste parameters. But such discrimination cannot persist under perfect competition, as employers with no preference will drive employers with discriminatory preferences out of the market for the discriminated employees by offering all workers equal wages. Hence, the presence of imperfect competition in the labor market might explain the persistence of discrimination.

2.1.1 PERFECT COMPETITION AND TASTE DISCRIMINATION

To illustrate the theory of taste discrimination, let us consider a labor market composed of workers who each produce a quantity y and who belong to two different groups, denoted A and B. Employers all have an aversion for workers of group A, even though they have the same productivity as workers of group B. For an employer, the gain derived from employing one of these workers is equal to $y - w_A - u$ where w_A is the wage received by workers in group A, and $u > 0$ is a parameter that represents the aversion employers feel toward workers of group A.

Under the free entry hypothesis, if the economy is composed solely of employers having an aversion toward workers of group A, the members of this group obtain a wage $w_A = y - u$, equal to their productivity y minus parameter u, which measures the aversion of employers for individuals in this group. Individuals in the nondiscriminated

group obtain a higher wage, equal to their productivity y. In this setting, if the labor supply of both groups rises with wages, individuals in the discriminated group are, all other things being equal, in employment less often. Taste discrimination thus leads to lower remuneration and less employment for the discriminated group. Besides, if the parameter u varies across firms, the latter tend to segregate: those for which $w_A + u > w_B$ prefer to hire workers of group B, and those for which $w_A + u < w_B$ prefer to hire workers of group A. The size of the wage differential between the two groups will be set by the marginal employer having some aversion $u > 0$ but still willing to hire a worker of group A (provided of course that at a wage $w_A = w_B$ there would not be enough employers to hire all workers of group A). Taste discrimination thus leads to lower remuneration, detrimental hiring policies on the part of more prejudiced firms, and less employment for the discriminated group.

If, however, there exist employers who experience no aversion for individuals in group A and if these employers can freely enter this labor market, the wage difference between the two groups vanishes. The null profit condition entails, on one hand, that the workers of groups A and B obtain the same wage equal to their productivity, and so $w_A = w_B = y$, and, on the other, that the employers experiencing an aversion towards workers of group A refrain from hiring these workers. Thus the preference certain employers have for discrimination results in segregation. The employers who discriminate employ only workers from group B, while the others employ workers from both groups indifferently.

Becker also discusses the case in which discrimination may arise out of the preferences of workers. In this situation, workers who belong to a majority group feel an aversion to working with members of a minority group, and employers must compensate the members of the majority group by paying them wages that exceed their productivity, financed by a levy on the wages of the minority workers, if they want the two types of workers to work together in the same firm. Clearly such a situation cannot arise under perfect competition, where the perfect mobility of workers must ensure that there is no firm employing members of both groups at the same time.

In sum, employer and employee discrimination resulting in persistent wage differences cannot occur in perfectly competitive markets, in which by definition all workers are paid according to their marginal productivity. Hence discrimination is necessarily linked to imperfect competition.

2.1.2 Imperfect Competition and Taste Discrimination

Limitations on personal mobility (geographical, or between kinds of employment) permit firms to exercise monopsonic power and to pay workers with identical productive abilities differently. This type of argument has been advanced to explain discrimination against women and certain ethnic minorities (see Gordon and Morton, 1974; Barth and Dale-Olsen, 2009). More generally, any employer enjoying some market power has, within limits, an opportunity to select workers according to her preferences. We will show that in this context, discrimination leads to lower wages and levels of employment for those who are its victims. These conclusions hold good in job search models.

Monopsony and Discrimination

To demonstrate this, let us suppose that a monopsonist is present in the market described above, composed of two groups, denoted A and B, of workers whose

productive abilities are strictly identical. Each individual supplies one unit of labor and the labor supply of individuals of group i is equal to $L^s(w_i) = G(w_i)$, $i = A, B$, where G designates the cumulative distribution function of the reservation wages.

If the entrepreneur feels a disinclination to employ workers of group A, his behavior is described by the following problem:

$$\max_{\{w_A, w_B\}} G(w_A)(y - w_A - u) + G(w_B)(y - w_B), \quad 0 < u < y$$

In this problem, w_A and w_B designate the wages that apply respectively to the members of groups A and B. Parameter u measures the loss which the employer feels in the presence of persons of group A.

Differentiating the criterion of the employer with respect to w_A and w_B, we find the values of the remunerations received by agents belonging to groups A and B. They are as follows:

$$w_i^M = \frac{\eta_w^L(w_i^M)}{1 + \eta_w^L(w_i^M)}(y - u_i), \quad \text{with } \eta_w^L(w_i) = \frac{w_i G'(w_i)}{G(w_i)} \geq 0 \quad \text{and} \quad u_i = \begin{cases} u \text{ if } i = A \\ 0 \text{ if } i = B \end{cases}$$

If the second-order condition is satisfied, we have $GG'' - 2(G')^2 < 0$, and we can verify that workers targeted for discrimination obtain a lower wage than that of the workers in group B. This result is easy to understand with the help of figure 8.10. The wage w_B^M obtained by workers in the group not targeted for discrimination corresponds to w^M, that is, to a tangency point between an isoprofit curve for the jobs in group B and the graph of the labor supply of this group. The slope of the isoprofit curve for the jobs in group A is given by $\mathrm{d}L/\mathrm{d}w = L/(y - w - u)$. It is greater than the slope of the

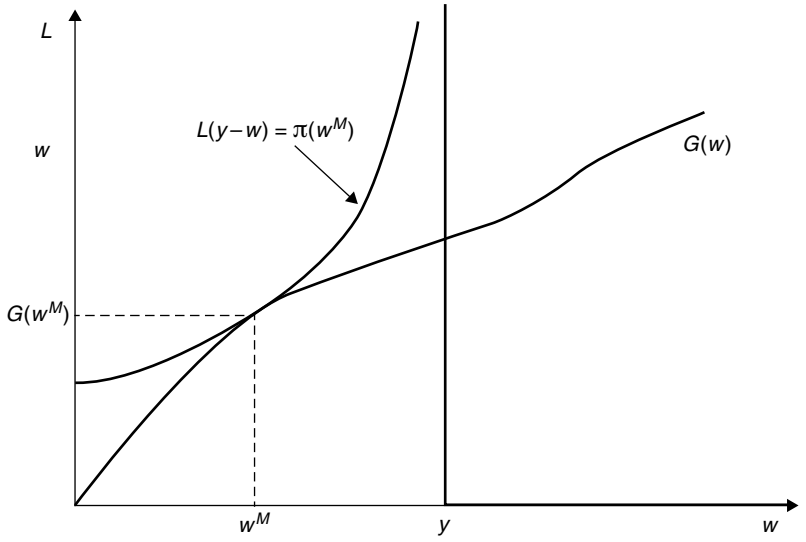


FIGURE 8.10
Discrimination in the monopsony model.

isoprofit curve for the jobs in group B for every pair (w, L). The tangency point between the isoprofit curve and the labor supply for group A is necessarily situated to the left of that for group B. Wages and employment are therefore both lower for group A.

Overall, the monopsony model shows that the taste for discrimination leads to lower wages and lower levels of employment for individual members of the discriminated groups, even though they have the same productivity. The disadvantages borne by individuals belonging to discriminated groups can persist as long as there is no possibility of forcing the employer who is discriminating to compete with other employers who experience no aversion for the discriminated groups.

Discrimination in Labor Markets with Frictions

The mechanisms brought to light in the monopsony model reappear in contexts in which the job search is costly. These search costs prevent workers from bringing the full weight of competition to bear on firms, and that confers some monopsony power on employers. Search models permit us to explain how discrimination arising out of the preferences of employers can persist in the market. This is shown by the model of Black (1995) and that of Bowlus and Eckstein (2002), which is closely related. With a job search model similar to that developed in chapter 5, Black (1995) has shown that discrimination based on the preferences of employers can in fact persist if workers are faced with costly job searches. Discrimination then takes the form of lower wages and longer spells of unemployment for the workers who are its victims. These models show that individuals belonging to discriminated groups receive lower wages not just with employers who have an aversion to the discriminated group but also with those who do not, since the presence of employers who discriminate reduces the employment and wage opportunities, and therefore the reservation wage, of persons in the discriminated group. Search models also show that when workers can choose where to search (directed search) they can self-select into low-wage jobs if they expect unfair competition with other types of workers in high-wage jobs (Lang et al., 2005).

2.2 Statistical Discrimination

Arrow (1973) and Phelps (1972) have shown that the unobservability of individual characteristics may provoke discriminatory behavior on the part of firms. The latter generally dispose of limited information about individual characteristics. They possess pieces of information like age, experience, education, and performance on hiring tests that may have been administered, but these elements are only correlated with productivity and so explain it only imperfectly. In order better to evaluate productivity, employers sometimes utilize supplementary information (or beliefs) on the *average* quality of one demographic group or another. A situation referred to as "statistical discrimination" may then arise. This expression signifies that individuals having identical abilities but belonging to different groups do not have equivalent career paths on account of the average quality, real or imagined, of the group to which they belong. We begin by showing how such a phenomenon may come about, and then focus on how statistical discrimination can become a source of *persistent inequality among groups* when the beliefs of employers influence the decisions agents make about education. These explanations of discrimination and inequalities throw valuable light on the consequences of quota policies of the kind that mandate, for example, the hiring of a given proportion of members of a certain group.

2.2.1 Statistical Discrimination as a Source of Individual Discrimination

Let us consider a labor market in which agents have zero opportunity cost of working and two different levels of productivity: a low level, $h^- = 0$, and a high one, $h^+ > 0$. Employers evaluate the performance of workers by using hiring tests or trial periods, the cost of which we take to be zero for the sake of simplicity. The test makes it possible to detect efficient workers (the h^+ type) with a probability equal to 1. The inefficient ones, however, have a probability $p \in [0,1]$ of passing the test and being wrongly taken for efficient. Moreover, employers estimate that the proportion of efficient workers in the demographic group considered is equal to $\pi \in [0,1]$. In these conditions, passing the test does not guarantee that the person hired will be efficient (the h^+ type) since an inefficient one (the h^- type) has a probability p of passing the test. The employer's first task is to assess quantitatively the reliability of the test in selecting efficient persons. In other words, he must calculate the a posteriori probability, denoted $\Pr\{h = h^+|\text{success}\}$, that a worker who passes the test will actually be of the h^+ type. By definition, this probability is given by the formula:

$$\Pr\{h = h^+|\text{success}\} = \frac{\Pr\{h = h^+ \text{ and success}\}}{\Pr\{\text{success}\}}$$

Since the test makes it possible to detect efficient workers infallibly, $\Pr\{h = h^+ \text{ and success}\}$ is equal to $\Pr\{h = h^+\}$. So we have:

$$\Pr\{h = h^+|\text{success}\} = \frac{\Pr\{h = h^+\}}{\Pr\{\text{success}\}} = \frac{\pi}{\Pr\{\text{success}\}} \tag{8.1}$$

The problem thus comes down to calculating the total probability $\Pr\{\text{success}\}$ of passing the test. The outcome {success} breaks down into two outcomes according to the equality:

$$\{\text{success}\} = \{\text{success and } h = h^+\} + \{\text{success and } h = h^-\} \tag{8.2}$$

From what has gone before, we know that the probability of outcome {success and $h = h^+$} is equal to π; as for the probability of outcome {success and $h = h^-$}, it is equal to the proportion $(1 - \pi)$ of inefficient workers times the probability p that one of them will pass the test. Taking the probabilities of both sides of relation (8.2), we find that $\Pr\{\text{success}\}$ is equal to $\pi + p(1 - \pi)$, and the equality (8.1) finally yields:[2]

$$\Pr\{h = h^+|\text{success}\} = \frac{\pi}{\pi + p(1 - \pi)}$$

For the employer, it turns out that the expected productivity of a person who passes the test is equal to $h^+\pi/[\pi + p(1 - \pi)]$. The condition of free entry then entails that this

[2] It would have been possible to obtain this equality directly by applying the Bayes formula:

$$\Pr\{h = h^+|\text{success}\} = \frac{\Pr\{\text{success}|h = h^+\} \cdot \Pr\{h = h^+\}}{\Pr\{\text{success}|h = h^+\} \cdot \Pr\{h = h^+\} + \Pr\{\text{success}|h = h^-\} \cdot \Pr\{h = h^-\}}$$

quantity also represents the wage of a worker who has passed the test. This wage applies to all workers of the h^+ type and to the proportion p of inefficient workers who pass the test (inefficient workers who fail it obtain a zero wage). It is increasing with the value π of the proportion of workers which employers estimate to be efficient. This constitutes a source of statistical discrimination, for the wage paid to efficient individuals is reduced by their membership in groups believed by employers to contain a high proportion of inefficient workers. The degree of precision in the tests is another source of statistical discrimination, for we can see that an increased probability p of failing the test has a negative impact on the wage. In this connection, Lang (1986) has pointed out that specific cultural and linguistic attributes of ethnic minorities may work to undermine the precision of their evaluation and for that reason constitute a source of statistical discrimination.

Statistical discrimination implies that individuals endowed with identical productive abilities may have different wages because they belong to different groups. Statistical discrimination may also appear in hiring decisions (and for that matter in areas outside the labor market, such as loan approval and insurance premium rate setting). Statistical discrimination does not, however, explain discrimination among groups. It does not allow us to understand why individuals belonging to different demographic groups *persistently* receive lower pay on average than their counterparts endowed with identical productive abilities. If individual performance is really independent of membership in a precise demographic group, repeated observation of this performance ought to cause employers to arrive sooner or later at an estimate of its true value, which is, by hypothesis, independent of membership in a group (Cain, 1986; Arrow, 1998).

2.2.2 Statistical Discrimination as a Source of Persistent Inequality Among Groups

Although statistical discrimination cannot persist, it is capable of creating inequalities, for the beliefs of employers and their capacity to make evaluations influence the behavior of workers. Let us assume that the efficiency of a worker depends in part on her investment in education. In a situation of statistical discrimination, the return to education is lower to the degree that employers believe that the proportion of inefficient workers in the group is substantial. This belief can act as an incentive for workers not to acquire education. Disincentive effects on educational investment can also arise when there is taste discrimination. However, in the situation of statistical discrimination, a self-fulfilling prophecy may come about: employers, anticipating that the proportion of efficient workers will be low, discourage efforts to acquire education and so do actually encounter fewer efficient workers (Lundberg and Startz, 1983; Coate and Loury, 1993; Loury, 2002).

A Model with Self-Fulfilling Prophecies

It is possible formally to illustrate this mechanism, in which beliefs lead to their own fulfillment, by slightly adapting the previous model. Let us now assume that workers can acquire education before starting their working lives. Their preferences are represented by a utility function $u(R, e) = R - e$, where R designates income, equal to wage w if they are employed, and 0 otherwise. The variable e represents the cost of the effort to acquire education. This cost may be equal to 1, which makes it possible to achieve efficiency of $h^+ > 1$, or to 0, in which case the worker has a productivity h^-, assumed to amount to zero. We represent decisions about education using a two-stage game. In the first stage, workers decide on educational effort e. In the second stage, there is free entry

into the labor market, and employers decide hires according to the process described in the previous model of statistical discrimination. At equilibrium, the beliefs of employers must be consistent, which means that their estimate of the proportion of efficient workers must be equal to the proportion actually observed.

We have shown that an educated worker obtains a wage $w^+ = h^+\pi/[\pi + p(1-\pi)]$ in the second stage, whereas an uneducated worker has an expected gain given by $\mathbb{E}(w^-) = pw^+$. An individual thus has an interest in acquiring education if $w^+ - 1 \geq \mathbb{E}(w^-)$, which is equivalent to:

$$\pi \geq \frac{p}{(1-p)(h^+ - 1)} \tag{8.3}$$

This condition indicates that workers only decide to acquire education if employers estimate that a sufficiently high proportion of the population to which they belong is efficient. In this sense, the beliefs of employers are indeed capable of influencing the behavior of workers.

Multiple Equilibria and Persistent Inequalities

The term $p/\left[(1-p)(h^+ - 1)\right]$ that appears in the right-hand side of (8.3) is greater than 1 if $p \geq (h^+ - 1)/h^+$. In this case, the inequality (8.3) is never satisfied, since the probability π must fall in the interval $[0,1]$. The frequency p with which inefficient workers pass the test is so high with respect to the gains won through education that there is no interest in acquiring education, whatever employers believe. Labor market equilibrium then corresponds to a situation in which no worker acquires education and in which the beliefs of employers must be such that $\pi = 0$ in order to be consistent with their observations. So all workers obtain a zero wage. The imprecision of the method of evaluation in this case represents an insurmountable source of statistical discrimination leading to deep inequalities, since the individuals who are victims of this discrimination decide not to acquire education.

If on the other hand $p \leq (h^+ - 1)/h^+$, there exist values of π capable of giving workers incentive to acquire education. Figure 8.11 shows that three equilibria, of which two are stable, are possible. In this figure, the curves u^+ and u^- represent the gains of workers in the plane (u, π). We see that for $\pi = 0$, workers prefer not to acquire education, since $u^- > u^+$. The value $\pi = 0$ thus represents an equilibrium at which no worker acquires education and where they all get a zero wage. But for $\pi = 1$, we necessarily have $u^+ > u^-$. The value $\pi = 1$ is then an equilibrium at which all workers become educated and thus obtain a wage equal to h^+. There also exists an equilibrium π_0 *strictly* comprised between 0 and 1. In this situation, workers are indifferent between acquiring education or remaining inefficient. But this equilibrium can be eliminated, for it is unstable: if a proportion $\pi_0 + \varepsilon$ (where ε is an arbitrarily small number) of workers get educated, all workers have an interest in getting educated for $\varepsilon > 0$ and none for $\varepsilon < 0$. A small deviation from equilibrium thus prevents a return to the initial position, which signifies that this equilibrium is unstable (the same line of reasoning will show that the other two equilibria are stable).

This very simple example shows that the influence of employers' beliefs may prevent groups from acquiring education and thus lead to persistent inequalities. If beliefs are unfavorable at the outset, $\pi < \pi_0$, it is possible that certain groups may be shackled to a low equilibrium ($\pi = 0$), while others, enjoying more favorable beliefs at the outset,

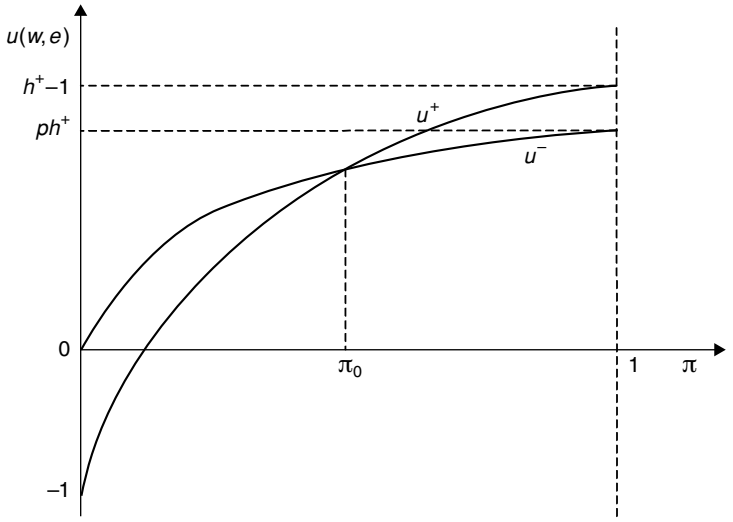


Figure 8.11
Statistical discrimination as a source of inequalities among groups.

$\pi > \pi_0$, may be coordinated at a high equilibrium ($\pi = 1$). In this respect, the weight of history becomes a significant source of discrimination, to the extent that beliefs are generally influenced by past experiences. Observation of poor performance by a group in the past is capable of influencing present beliefs and exerting a disincentive effect on the behavior of members of the group in question; the dynamic of self-fulfilling prophecies can engender persistent inequalities (Loury, 1998).

3 MEASURING WAGE DISCRIMINATION

The assessment of wage discrimination poses methodological problems linked essentially to the insufficiency of information on individual characteristics. Discrimination is rarely observed directly. Court cases are rare and it is often impossible to disentangle what factor in wage differences among people holding the same job reflects individual differences or deliberate discriminatory decisions by the employer. Wages in principle should be determined primarily by productivity, and wage differences should reflect productivity differences due to either job or personal characteristics. Whatever its origin (taste-based or statistical), discrimination could then ideally be identified as any remaining wage gaps across individuals with exactly the same productivity but belonging to different demographic groups (race, ethnicity, gender, etc.). Measuring discrimination well comes down first to measuring productivity well. Unfortunately productivity is generally not observed directly. What is observed are some key individual characteristics such as education, experience, and occupation that should have a strong impact on productivity. But some other important determinants of productivity—such as ability, noncognitive skills, or personal history—remain largely unobserved. These hidden characteristics can be linked to those observed (such as the years spent in education) or

to group membership. We present the econometric methods utilized in the estimation of wage discrimination within the two domains that have been studied most thoroughly, the racial and ethnic wage gap and the gender wage gap. We will see that the interpretation of estimates is particularly difficult and blunts the precision of this strategy. It should also be noted that even though many papers try to evaluate the implication of discrimination on wages and employment, very few papers achieve a convincing identification of the nature of discrimination or a clear distinction between taste-based and statistical discrimination.

In order to circumvent these difficulties and identify discrimination, numerous contributions have developed methods that try to estimate discrimination directly through field experiments (audit and correspondence studies), laboratory experiments, or a focus on particular situations, especially in the area of sports, where individual performance is directly observable.

3.1 Estimations of Wage Equations: The Case of the Black–White Wage Gap

The degree of labor market discrimination across races, and notably between blacks and whites in the United States, is among the most debated issues in the literature. A first approach to measuring discrimination across demographic groups in the labor market is to estimate wage equations that include relevant productivity-related factors and a dummy to identify the effect of group membership. The results of this type of estimation appear to be highly sensitive to the nature and the measure of the control variables, which are skills and educational level. These results are questionable as well in that they rest on strong hypotheses about the unobservable characteristics of individuals.

We illustrate this approach and its limitations utilizing the analyses of the black–white wage gap in the United States by Neal and Johnson (1996) and Lang and Manove (2011). The main results of these contributions are presented below and can be replicated with data and programs available at www.labor-economics.org. The question of the existence of a racial or ethnic wage gap has attracted continuing attention in the wake of the Civil Rights Act of 1964 and affirmative action policies. In a highly influential article, Neal and Johnson (1996) showed that the black–white wage differential, which until then was considered mostly due to discrimination, was dramatically reduced, and in some cases eliminated, by controlling for cognitive skills. Blacks earn less than whites mostly because they are less skilled, due to "premarket" factors. Moreover, Neal and Johnson showed that the return to skills in terms of earnings was as large for blacks as for whites, so that labor market discrimination could not have reduced incentives for blacks to invest in skills acquisition. Lang and Manove (2011) reopened this debate by showing that when controlling for years spent in education, and not just for skills level, blacks acquire more education than whites with the same skills but are not rewarded, leading to significant and unexplained wage differentials.

3.1.1 The Basic Equation

The standard approach consists of estimating an equation in which the logarithm of income is explained by a set of factors like the duration and quality of schooling, experience, and region, and by dummy variables representing ethnic origin and sex

(which are the principal sources of wage discrimination when it occurs). Let w_{it} be the income of individual i at time t, $\mathbf{x}_{it}$ the vector of individual characteristics and of the job held, and $\boldsymbol{\mu}_i$ a vector of dummy variables with a value of 1 if the individual belongs to groups potentially discriminated against, and 0 if not;[3] the estimated equation is written:

$$\ln w_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \boldsymbol{\mu}_i\boldsymbol{\alpha} + \varepsilon_{it} \tag{8.4}$$

In this equation, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are vectors of parameters to be estimated and the term ε_{it} represents a normally distributed disturbance term with zero mean. If the set of variables explaining the wage is sufficiently rich (but not too rich; see below), a negative value for one of the components of vector $\boldsymbol{\alpha}$ indicates that there is discrimination against the corresponding group with respect to the reference group. More exactly, each component of estimated vector $\hat{\boldsymbol{\alpha}}$ measures the average loss of income with respect to the reference group, evaluated in percentages, due to membership in the group to which this component relates, all other observable characteristics, such as age or education, being equal.

Identifying Assumptions and Interpretation

The estimation of equation (8.4) with ordinary least squares can be questionable, for unobservable individual characteristics, such as ability or social network, contained in ε_{it} can be correlated with membership $\boldsymbol{\mu}_i$. In other words, the estimation of $\boldsymbol{\alpha}$ is not biased under the standard hypothesis that the residues ε_{it} are not correlated to $\boldsymbol{\mu}_i$, conditionally upon observable characteristics $\mathbf{x}$, or formally, if:

$$Cov(\mu, \varepsilon \mid \mathbf{x}) = 0 \tag{8.5}$$

In this equation, μ represents any component whatever of vector $\boldsymbol{\mu}_i$ and ε designates the residue ε_{it}. In order to properly measure the weight of this condition, it is worth noting that it is equivalent, in this context, to the "conditional mean independence assumption," which is written, with the hypothesis that $\mathbb{E}(\varepsilon \mid \mathbf{x}) = 0$:[4]

$$\mathbb{E}(\varepsilon \mid \mu = 0, \mathbf{x}) = \mathbb{E}(\varepsilon \mid \mu = 1, \mathbf{x}) = \mathbb{E}(\varepsilon \mid \mathbf{x}) = 0 \tag{8.6}$$

[3] This assumes of course that group membership is something easy to define and identify, as gender clearly is. In many cases, however, it is not that clear. For instance, there are many shades of skin color, individuals may have multiple ethnic origins, etc. See Charles and Guryan (2011) for a presentation of this problem of taxonomy and its implications.

[4] To simplify the notation, we leave out the conditioning variable $\mathbf{x}$. Since $\mathbb{E}(\varepsilon) = 0$, we have:

$$Cov(\mu, \varepsilon) = \mathbb{E}(\mu\varepsilon) - \mathbb{E}(\mu)\mathbb{E}(\varepsilon) = \mathbb{E}(\mu\varepsilon)$$

with, since μ is a variable taking the value 0 or 1:

$$\begin{aligned}\mathbb{E}(\mu\varepsilon) &= 1.\Pr(\mu = 1)\mathbb{E}(\varepsilon \mid \mu = 1) + 0.\Pr(\mu = 0)\mathbb{E}(\varepsilon \mid \mu = 0)\\ &= \Pr(\mu = 1)\mathbb{E}(\varepsilon \mid \mu = 1)\end{aligned}$$

In consequence:

$$Cov(\mu, \varepsilon) = \Pr(\mu = 1)\mathbb{E}(\varepsilon \mid \mu = 1)$$

This equation shows that the covariance is null if and only if $\mathbb{E}(\varepsilon \mid \mu = 1) = 0$.

Since by definition:

$$\mathbb{E}(\varepsilon) = 0 = \Pr(\mu = 1)\mathbb{E}(\varepsilon \mid \mu = 1) + \Pr(\mu = 0)\mathbb{E}(\varepsilon \mid \mu = 0)$$

we have $\mathbb{E}(\varepsilon \mid \mu = 0) = 0$ when $\mathbb{E}(\varepsilon \mid \mu = 1) = 0$. In consequence $Cov(\varepsilon, \mu) = 0$ if and only if $\mathbb{E}(\varepsilon \mid \mu = 0) = \mathbb{E}(\varepsilon \mid \mu = 1) = 0$.

and which means that the mean value of the unobservables is identical across groups after controlling for observable characteristics.

This is of course impossible to test, since unobservables are by definition unobserved. For instance, children of minority groups might suffer from higher poverty rates which could impact the development of some noncognitive abilities which are not observed by the econometrician. So, in practice we may suspect that estimates of membership on wages are in general biased unless, after conditioning upon observable characteristics, there are good reasons to think that the determinants of wages that remain unobserved are the same across groups. This is unlikely in many cases.

Overall, it is difficult to define which wage determinants are related to group membership and which are not and should simply be controlled for. Still, many studies use equation (8.4) to try to account for the contribution of a range of control variables compared to group membership. Certain papers have extended the number of control variables to include the quality of education and also exogenous measures of ability, to better identify the effect of group membership independently of the effect of skills (see below).

3.1.2 Controlling for Skills

In their seminal paper, Neal and Johnson (1996) apply this method to race differences in wages, for blacks, whites, and Hispanics in the United States. Previous studies typically concluded that even though observable productive characteristics are important sources of black–white wage differentials, labor market discrimination accounts for at least one third to one half of the total gap. In their paper they address two types of problem typical of these studies: first, they try to avoid including productive factors that result from workers' choices and thus could be affected by labor market discrimination (and bias the estimate of the race dummies), such as postsecondary schooling (or more generally education beyond the compulsory age of education), part-time work, occupation, location, and even marital status; second, they include a measure of skills as the key productive characteristic, instead of number of years in education which typically overstates the relative "skills" of blacks since black children tend to demonstrate lower levels of achievement compared to white children in the same grade, which in turn is likely to overstate the role of discrimination. The number of years in education is only an indirect and noisy measure of skills acquired. Put differently, if skills are omitted but are truly a key determinant of wages in equation (8.4), the conditional mean independence assumption (8.6) will be violated because then there will be a relationship between the unobserved level of skills and group membership after controlling for observable characteristics.

To achieve these goals, Neal and Johnson (1996) use the National Longitudinal Survey of Youth (NLSY) of 1979 for the cohort born between 1957 and 1964, following 12,686 young people. The NLSY provides information on skills, education, family background, and work experience for the same individuals over a long span of years. The authors seek to make use of results from a test of skills *before* individuals entered the labor market (since postsecondary education and work experience can influence the results of the test and can themselves be influenced by discrimination, test scores might be influenced by discrimination after youths have entered the labor market). Since the test was taken by participants in 1980, when the cohorts in the sample were aged 15 to 23,

the authors restrict the sample to young people born after 1961 and who were at most 18 when they took the test, and for whom wages were observed in 1990 and 1991 when they were 26 to 29. This way, Neal and Johnson argue that skills measurement is premarket, not influenced by discrimination in the labor market (this point will be discussed below). The test is the Armed Forces Qualifying Test (AFQT) used for enlistment screening and job assignment by the military. It comprises scores for reading comprehension, word knowledge, arithmetic reasoning, and mathematics knowledge. Several studies have verified that the AFQT is a racially unbiased measure of cognitive skills, which is essential to the object of this analysis. Since the test was taken at different ages, Neal and Johnson adjust the test results for the age at the test date: the AFQT score is first regressed on age and then the score is adjusted by subtracting age times the coefficient obtained on age. The authors also normalize the score so that it has a mean of 0 and a standard deviation of 1. On this basis, black men have an age-adjusted average score of −.621, compared with −.284 for Hispanics and .422 for whites.

As the sample also comprises Hispanic young people, two race/ethnic dummies μ_{Bi} and μ_{Hi} are included in the vector $\boldsymbol{\mu}_i$ in equation (8.4), which can be rewritten:

$$\ln w_{it} = \sum_{k=1}^{K} x_{kit}\beta_k + \mu_{Bi}\alpha_B + \mu_{Hi}\alpha_H + \varepsilon_{it} \tag{8.7}$$

where K is the number of observable characteristics considered in the model, $\mu_{Bi} = 1$ if the individual is black and 0 otherwise, and $\mu_{Hi} = 1$ if the individual is Hispanic and 0 otherwise, and β_k, α_B, and α_H are parameters to be estimated.

The vector of productive characteristics $\mathbf{x}_{it}$ includes the age, the AFQT score, and its square (assuming the returns to skills might not be linear). Alternatively to skills, Neal and Johnson include a variable for schooling, with years of schooling achieved in 1991. Estimates are done separately for women and men because of sample selection issues that are typically more pronounced for women (see the next section), and include wages for the years 1990 and 1991. Results are presented in table 8.2.

The main result from these regressions is that, while the unadjusted male wage gap between black and white $\widehat{\alpha}_B$ is estimated at −.244 and the unadjusted female wage gap at −.185 units of log (columns (1) and (4) of table 8.2), these gaps go to −.072 and +.035 (not significantly different from zero) respectively once a control is introduced for skills before entering the labor market (columns (3) and (6) of table 8.2). The test score for skills explains three quarters of the wage differences with whites for young men and all the difference for young women. The wage gap is also totally explained by skill differences for young Hispanic men. This is a striking difference compared to the impact of schooling (columns (2) and (5) of table 8.2), which only reduces the unadjusted wage gap by one fifth for black men and one sixth for black women.

In order to verify that skills have the same return for whites and blacks, Neal and Johnson show that actually the realized effect of AFQT scores on wages is not different for black men and women than for whites. To do so they included interaction terms between the race dummy and the AFQT variables in equation (8.7): $\mu_{Bi}.AFQT_i$ and $\mu_{Bi}.AFQT_i^2$. The corresponding coefficients are not significant, except for highly skilled black men who fare relatively better at the end of the distribution.[5] Under these

[5] See tables 2 and 3 of Neal and Johnson (1996).

TABLE 8.2
Black–white wage gap among the younger generations of the NLSY 1970 cohort.

Dependent	Mean log wage						Median log wage	
	Men (participants)			Women (participants)			All men	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Black [$\widehat{\alpha}_B$ in eq. (8.7)]	−.244 (.026)	−.196 (.025)	−.072 (.027)	−.185 (.029)	−.155 (.027)	.035 (.031)	−.352 (.029)	−.134 (.035)
Hispanic [$\widehat{\alpha}_H$ in eq. (8.7)]	−.133 (.030)	−.045 (.029)	.005 (.030)	−.028 (.033)	.057 (.031)	.145 (.032)	−.180 (.034)	−.007 (.038)
Age	.048 (.014)	.046 (.013)	.040 (.013)	.010 (.015)	.009 (.014)	.023 (.015)	.067 (.015)	.055 (.017)
AFQT	. .	. .	.172 (.012)	. .	. .	.228 (.015)		.206 (.015)
$AFQT^2$	. .	. .	−.013 (.011)	. .	. .	.013 (.013)		−.010 (.014)
High grade by 1991	. .	.061 (.005)	. .	. .	.088 (.005)	. .		
Number of observations	1593	1593	1593	1446	1446	1446	1674	1674

Note: Based on wages observed in 1990 and 1991, individuals born after 1961. OLS regressions, standard errors in parentheses.

Source: Neal and Johnson (1996, tables 1 and 4).

conditions, table 8.2 suggests that blacks are paid less primarily because they are less skilled, not because of market discrimination. One explanation of this result could be that blacks underinvest in skill acquisition even when very young because they expect a lower return to skills later, once they have entered the labor market. Statistical discrimination models presented in section 2 would typically yield this type of lower investment as a best response to discrimination.

3.1.3 SELECTION BIAS

It should be noted that selection effects might seriously contaminate the results. Selection bias may be present since we only observe the wages of persons who work (see chapter 1, section 3.1.2). This is potentially an important source of error in the assessment of wage differentials between groups. The average wage of all the members of a group should actually depend not just on the (observed) wages of workers who have a job but also on the *potential* (and thus unobserved) wages of persons in this group who do not have a job. The distribution of observed wages therefore represents only a part of the distribution of the "offered wages" and it is necessary to know this last distribution in order to evaluate the wage differences between groups. The importance of this bias, to which Butler and Heckman (1977) drew attention, was illustrated by Brown (1984), Chandra (2000), and Heckman et al. (2000) in the analysis of wage gaps between blacks and whites in the United States. Two different approaches can be used to try to offset selection bias. The first approach imputes values for the missing wage data of nonparticipants and the second estimates wage and participation equations simultaneously.

Imputation of Values to Missing Data

To offset the selection bias, it is possible to impute values for the missing wage data of nonparticipants. This approach can be implemented when selection is based on both observable and time-invariant unobservable characteristics. For instance, Brown (1984) calculates the average wage of each demographic group on the assumption that the wage offered to a nonparticipating individual comes from a random draw below the median of the distribution of observed wages. On this basis, the movement of wage gaps turns out to be much less favorable to black men, for the members of this group situated at the low end of the wage distribution, who were excluded from the calculation of the average of observed wages, are now included, which contributes to bringing down the average wage of all black men. Along the same lines, Neal and Johnson (1996) assign an arbitrarily low wage (one cent per hour) to male nonparticipants, which ensures that they would not participate conditional on their characteristics. The results from this approach are shown in columns (7) and (8) of table 8.2: as expected, the estimated black–white log wage gap at the median is larger at −.352, and is reduced to −.134 when skills are controlled for with the AFQT scores, which is a higher measure of discrimination than when estimated at the mean of participants only. Still, 60% of the unadjusted wage gap is explained by skills.

There are other ways of calculating average wage. The "matching cell mean" method brings in the unobserved wages by creating categories using age and educational criteria and assigning each unemployed individual the average wage of the category to which he belongs. This method yields a movement of wage gaps close to that observed using the raw data.

Simultaneous Estimation of Wages and Participation

The second approach to dealing with selection bias is to estimate wages and participation simultaneously. This technique is especially appropriate when labor market participation is based on unobservable characteristics, but it is more demanding because of the identification requirements (see chapter 1, section 3.1.2). In this approach, based on Heckman (1979), a participation control variable—the inverse Mills ratio—obtained by fitting a probit model of the participation decision (see chapter 1, appendix 7.5 for a more complete presentation) is added as a regressor in equation (8.4). More precisely, if P_{it} denotes a dummy indicating participation in the market, the conditional probability of participating $\Pr[P_{it} = 1|\mathbf{z}] = \Phi(\mathbf{z}\boldsymbol{\gamma})$ must be estimated separately, where $\mathbf{z}$ is a vector of explanatory variables (including at least one variable excluded from the hours equation), $\boldsymbol{\gamma}$ is a vector of parameters, and Φ is the cdf of the standard normal distribution. The inverse Mills ratio is then $\lambda_{it} = \Phi'(\mathbf{z}\widehat{\boldsymbol{\gamma}})/\Phi(\mathbf{z}\widehat{\boldsymbol{\gamma}})$ and the wage equation becomes:

$$\ln w_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \boldsymbol{\mu}_i\boldsymbol{\alpha} + \lambda_{it}\delta + \varepsilon_{it} \qquad \textbf{(8.8)}$$

where δ is a parameter to be estimated. The "exclusion restriction" requires that an available instrument $\mathbf{z}$ explaining participation be included in the participation equations but not in the wage equation (see chapter 1, appendix 7.5). Finding an instrument that explains participation but has no influence on wages is not an easy task, and the variables used in the empirical literature are often debatable. For instance, Mulligan

and Rubinstein (2008) analyze the gender gap on the Current Population Survey cross sections since 1970. They control for selection bias, which may have changed over time, pulling more and more skilled women into the market. In their analysis, **x** includes educational attainment dummies, marital status, and a potential work experience quartic interacted with education. The vector **z** has the same elements, plus the number of children aged 0–6 interacted with marital status (the instruments). Selection bias is assumed to be negligible for men and thus the inverse Mills ratio is only estimated for women. They find that most of the apparent narrowing of the gender wage gap actually reflects changes in female workforce composition.

3.1.4 Controlling for Skills and Education

The study by Neal and Johnson reveals the importance of premarket factors in the analysis of wage differences. One question remains though: what are the determinants of skills and, more precisely, what drives the AFQT scores used in the regressions? Is AFQT a pure measure of ability or is it influenced by the family and school background as well as the years of education? If so, should these factors be included in equation (8.7)? The choice of these variables is crucial to the identification of discrimination, as the recent paper by Lang and Manove (2011) revealed.

Neal and Johnson analyzed the determinants of AFQT test scores and found that the black–white test score gap is reduced by half when account is taken of family background (education and professional status of parents, number of siblings, and reading material at home). The relation to school quality is less strong, maybe because several effects at stake cancel out. However, some racial gap remains even after accounting for these factors. Besides, the racial gap in test scores (measured at the same date for all participants) increases among older cohorts compared to younger cohorts, suggesting that this AFQT test is not a pure measure of innate ability but can be influenced by further education and work experience (which tend to diverge over time across blacks and whites). Indeed, an additional year of schooling increases test scores. The difficulty in measuring the relationship between test scores and schooling is that education may be influenced by ability measured by test scores, as much as test scores can be influenced by schooling. Using the quarter of birth as an instrument for years of education completed (as in Angrist and Krueger, 1991; see chapter 4), Neal and Johnson show that an additional year of education increases AFQT scores by about a quarter of standard deviation (the test gap between blacks and whites being -1 standard deviation—a typical result in the related literature).

The question remains of knowing whether or not to include these factors, and notably education, in the wage equation. Lang and Manove (2011) argue that it is inappropriate to control only for AFQT performance. They rely on the same data as Neal and Johnson (1996), and the same cohort, but observe the hourly wages for the years 1996, 1998, and 2000 instead of 1990–1991, and they focus on blacks and whites only. Education is measured by the highest grade completed as of 2000 (when wages are observed). It turns out that blacks get on average about three quarters of a year less education than do whites. But conditional on AFQT scores, black men get about 1.2 years more education than do white men with the same AFQT, and black women 1.3 years more. That means that blacks have on average longer durations of education to attain the same level of AFQT score as whites. One possible explanation is that AFQT is largely determined by

schooling, and since blacks often attend lower-quality schools, they gain fewer cognitive skills on average from a given level of education. On this view, blacks have more schooling for a given AFQT because it takes more schooling to reach a given level of cognitive skills. Another possibility, which would be at odds with the self-fulfilling mechanism presented in section 2.2.2, but favored by Lang and Manove, is that blacks have incentives to overinvest in education, compared to equivalently able whites, because they use education as a signaling device to counteract expected discrimination in the labor market. The similar earnings of blacks and whites when controlling only for AFQT suggest that blacks do not reap the expected reward for this supplement of education.

Neal and Johnson rejected education as a valid control variable, notably because education after the compulsory age is endogenous to discrimination; hence they did not add it to AFQT in their preferred specification. If schooling is negatively influenced by discrimination, their argument is valid: in that case blacks get less schooling than equivalent whites, and including educational achievements in equation (8.7) introduces a bias toward overstating the adjustment of the wage gap and thus biases downward the role of discrimination. In part, the impact of discrimination on wages would be accounted for already by education. But education might be influenced the other way around: in fact, blacks tend to obtain more education than whites, holding AFQT constant. Lang and Manove provide the following example: imagine that the market discriminates against blacks by paying them exactly what it would pay otherwise equally able whites with exactly one less year of education. Then, to a first approximation, all blacks should tend to get one year more education than otherwise equivalent whites to counteract this effect. Controlling only for a proxy of ability, such as AFQT, blacks and whites will have the same earnings—because blacks compensate with more education—but controlling for education as well as ability, blacks will earn less than whites by an amount equal to the return to one year of education. Hence excluding education from the equation can also bias downward the estimated effect of discrimination.

When controlling for both education and AFQT, wage differentials between blacks and whites are indeed substantially larger than when controlling for AFQT alone. Lang and Manove estimate equation (8.4) including in the set of productive characteristics $\mathbf{x}_{it}$, age, AFQT score and its square, and also family background, school input, and education (final educational attainment). Results are shown in table 8.3 for the younger cohort, all participants, based on the mean and median of log wages. The first row shows the very large differential that exists when we control only for age. The second row shows that, consistent with Neal and Johnson, AFQT (and its square) accounts for about three quarters of the unadjusted wage gap. The comparison of the second and third rows shows that controlling for educational attainment on top of AFQT test scores increases substantially the adjusted wage differential, as predicted by Lang and Manove. The last two rows show that controlling for family background and school input further reduces the adjusted wage gap, but that adding education again increases it. These results were also confirmed by Carneiro et al. (2005), who adjust the AFQT test score by the number of years of schooling at the time of the test, arguing that if test scores are influenced by education and education is potentially influenced by discrimination, then test scores should be adjusted first. These studies suggest that even though premarket factors are essential (Neal and Johnson's core results are not fully overturned), they cannot account for the full wage gap between blacks and whites: there seems to be a differential treatment in the labor market as well.

TABLE 8.3
Black–white wage gap among men in the NLSY 1979 cohort.

Black ($\widehat{\alpha}_B$ in eq. (8.7))	Mean log wage		Median log wage
Controlling for:	Young cohort	All men	All men
Age	−.37 (.04)	−.36 (.03)	−.42 (.03)
Age, AFQT	−.13 (.03)	−.09 (.02)	−.10 (.03)
Age, AFQT, education	−.18 (.03)	−.15 (.02)	−.18 (.03)
Number of observations	1634	3841	4055
Age, AFQT, family background, and school input	−.07 (.06)	−.06 (.04)	−.05 (.03)
Age, AFQT, education, family background, and school input	−.11 (.06)	−.11 (.04)	−.11 (.04)
Number of observations	731	1876	1955

Note: Based on wages observed in 1996, 1998, and 2000, Young cohort: individuals born after 1961. Note: OLS regressions, standard errors in parentheses.

Source: Lang and Manove (2011, table 5).

3.2 DECOMPOSITION METHODS: THE CASE OF THE GENDER WAGE GAP

Another approach to the measurement of discrimination is to estimate separately wage equations for the different groups and identify what part of the wage gap stems from individual characteristics and behaviors that can vary across groups, and what is left unexplained. These are called "decomposition methods." We will illustrate these methods for the gender wage differences in the United States as identified by O'Neill and O'Neill (2006).[6] The main results of this contribution are presented below and can be replicated with data and programs available at www.labor-economics.org. While there is no doubt about the existence of the persistent gender gap in the labor market despite a convergence of wages and employment rates over the last decades (OECD, 2012), the extent to which this remaining gap is due to discrimination is the object of intense debate. To sort out the various factors influencing wages, O'Neill and O'Neill base their analysis on the National Longitudinal Survey of Youth (NLSY), which provides information on skills and work experience for the same individuals over a large number of years. They show that the unexplained part of the gender gap, usually identified with discrimination, is significantly reduced when account is taken of the external nondiscriminatory factors that ought to reduce women's wages. But the conclusions of this type of analysis rely on a number of strong assumptions that are similar to those made above for the estimation of wage equations (see Fortin et al., 2011, on which our presentation is based).

[6] O'Neill and O'Neill (2006) also present a racial wage gap decomposition.

3.2.1 The Blinder-Oaxaca Method

When two groups are under consideration, the so-called Blinder-Oaxaca decomposition (Oaxaca, 1973; Blinder, 1974) aims at decomposing the mean wage across these groups: a reference group and another group to be compared to the reference group. The underlying wage-setting model ought to be linear with separate observed and unobserved characteristics. This method has become a standard part of the toolkit of labor economists. Decomposition methods can be applied to other types of outcomes and agents, such as gaps in test scores between schools. The number of groups can be larger than two. We illustrate this approach with the decomposition of the gender wage differences in the United States.

The Basic Decomposition

Let us assume that we have two mutually exclusive groups, that is, that individuals in the sample belong to either one group or the other but not both (in the case from O'Neill and O'Neill (2006) presented here, men and women). Let w_{Ai} and $\mathbf{x}_{Ai}$ be respectively the wage and the vector of observable characteristics of an individual i belonging to the reference group A. The wage equation relative to this group takes the form (we ignore the time dimension here for simplicity):

$$\ln w_{Ai} = \mathbf{x}_{Ai}\boldsymbol{\beta}_A + \varepsilon_{Ai} \tag{8.9}$$

In this equation $\boldsymbol{\beta}_A$ designates the vector of parameters to be estimated and ε_{Ai} represents a normally distributed disturbance term with zero mean. The term ε_{Ai} can also be interpreted as the effect of individual unobserved characteristics. In the same manner, the wage equation relative to a group B is written:

$$\ln w_{Bj} = \mathbf{x}_{Bj}\boldsymbol{\beta}_B + \varepsilon_{Bj} \tag{8.10}$$

Compared with equation (8.4), the last two equations allow us to estimate a divergent effect of control variables on wages for the two groups ($\boldsymbol{\beta}_A$ and $\boldsymbol{\beta}_B$, as opposed to $\boldsymbol{\beta}$). They correspond to two distinct wage structures. They also allow us to calculate the difference between the average values of the wage logarithms. Denoting this difference $\Delta = \mathbb{E}(\ln w_A) - \mathbb{E}(\ln w_B)$, we get:

$$\Delta = \mathbb{E}(\mathbf{x}_A)\boldsymbol{\beta}_A + \mathbb{E}(\varepsilon_A) - \mathbb{E}(\mathbf{x}_B)\boldsymbol{\beta}_B - \mathbb{E}(\varepsilon_B) \tag{8.11}$$

Since it is assumed that $\mathbb{E}(\varepsilon_A) = \mathbb{E}(\varepsilon_B) = 0$, we have:

$$\Delta = \mathbb{E}(\mathbf{x}_A)\boldsymbol{\beta}_A - \mathbb{E}(\mathbf{x}_B)\boldsymbol{\beta}_B$$

Replacing the expected values of covariates by their group means, the estimated decomposition becomes:

$$\hat{\Delta} = \overline{\mathbf{x}}_A\widehat{\boldsymbol{\beta}}_A - \overline{\mathbf{x}}_B\widehat{\boldsymbol{\beta}}_B \tag{8.12}$$

Adding and substracting the average counterfactual wage that group B workers would have earned under the wage structure of group A, $\overline{\mathbf{x}}_B\widehat{\boldsymbol{\beta}}_A$, to the previous equation yields the estimated decomposition:

$$\overline{\ln w_A} - \overline{\ln w_B} = (\overline{\mathbf{x}}_A - \overline{\mathbf{x}}_B)\widehat{\boldsymbol{\beta}}_A + \overline{\mathbf{x}}_B\left(\widehat{\boldsymbol{\beta}}_A - \widehat{\boldsymbol{\beta}}_B\right) \quad \textbf{(8.13)}$$

Here $\overline{\mathbf{x}}_A$ and $\overline{\mathbf{x}}_B$ designate the average values of the vectors of observed characteristics. The first term of the decomposition, $(\overline{\mathbf{x}}_A - \overline{\mathbf{x}}_B)\widehat{\boldsymbol{\beta}}_A$, represents the "explained" component of wage differences between groups, also called the composition effect. It concerns elements like education, experience, social milieu, and the nature of the jobs held. The second term, $\overline{\mathbf{x}}_B\left(\widehat{\boldsymbol{\beta}}_A - \widehat{\boldsymbol{\beta}}_B\right)$, represents the "unexplained" component, also called the wage structure effect. It measures, for group B, the differences of return to characteristics due to membership in this group. It builds a counterfactual: what would group B members be paid if they had the same returns to characteristics as group A members? The advantage of the Blinder-Oaxaca decomposition is that it does not demand that the coefficients linked to individual characteristics be identical. That notwithstanding, without further assumptions the unexplained component may capture the effects of characteristics not observable by the econometrician, on top of any possible discrimination effects.

Identifying Assumptions and Interpretation

The Blinder-Oaxaca decomposition method has been viewed critically as a mere accounting exercise based on correlations, with no causal interpretations of the underlying parameters. As suggested by Fortin et al. (2011), this is because most papers using this approach do not discuss their identification strategy first: What do we want to estimate? What assumptions are needed to interpret these estimates as sample counterparts of the parameters in the model? What is the best procedure to recover these parameters? However, under the following identifying assumptions, the interpretation of explained and unexplained components of the decomposition is easier:

- The first of these assumptions is the existence of a common support: the set of variables potentially influencing the wage level, $\mathbf{x}$ and ε, are the same across groups. This might not be the case if, for instance, group B was made up of immigrants for whom the age of arrival in the country can influence the wage, compared to native-born individuals. In these cases, the decomposition across groups might be harder to interpret.
- The second assumption is the conditional mean independence assumption, as stated in the previous section in equation (8.5), which is necessary to distinguish differences associated with returns to observable characteristics from differences stemming from unobservable characteristics. Indeed, if the distribution of the unobservable characteristics is independent of the group membership conditional on the observable characteristics, then the composition effect based on differences in individual characteristics across groups really can be identified separately from the effect of the return to observables (what we call the wage structure effect). Again, note that the conditional mean independence assumption is a strong assumption which may not hold in many cases, due for instance in the case of women to differences in labor market

participation based on unobservables, which might influence observed wages. Also, in some cases, group membership, such as belonging to a union, might be endogenous based on unobservables.

- The third assumption is the invariance of the conditional distributions, by which the conditional wage distribution based on characteristics of the individuals of the reference group would remain valid if members of the other group were paid like them. Were all women paid the same wage as men, would wages of men remain the same in the sample? Put differently, this assumption amounts to excluding the possibility of equilibrium effects and self-selection into groups based on unobservables. If all women were paid as men are, however, the equilibrium effects would probably be non-negligible.

Should these three assumptions be valid, we would have assurance that the explained component of a simple Blinder-Oaxaca decomposition between two mutually exclusive groups does reflect only the effect of the differences in the distribution of observable characteristics between the two groups, while the unexplained component reflects solely the difference between the underlying structural wage functions including the effect of discrimination. In practice, however, these very restrictive assumptions are likely to remain unsatisfied in many cases.

Decomposing the Gender Wage Gap

Based on this approach, and keeping these limitations in mind, O'Neill and O'Neill (2006) exploit the NLSY cohort, which was first interviewed in 1979 (at ages 14–22) and then each year through 1994 and every other year since then until 2000 (at ages 35–43). The sample includes 5,600 wage and salary workers. The NLSY contains actual labor market histories, which makes it possible to measure accurately work experience (for an important feature of the data set for estimating gender wage differences, see Regan and Oaxaca, 2009).[7] The sample includes detailed information on education and many other individual characteristics and behaviors that can influence labor market outcomes. It also includes, for all participants, the Armed Forces Qualifying Test (AFQT) score, which is viewed as a nonbiased measure of cognitive skills, reflecting ability (but also differences in educational attainment; see the discussion above about the racial wage gap). The NLSY is a good data set for the study of the gender wage gap, primarily because it contains more information than comparable sets on lifetime patterns of work, labor market participation, and family.

However, education and skills ought to be less important factors in the case of gender differences, notably compared with work experience. Table 8.4 details the contributions of education and work experience to the log of hourly wage for men and women, and it shows how the corresponding composition effects are calculated (data and calculations are provided by Fortin et al., 2011). The first two columns show the average value of the various education dummies and work experience variables for both

[7] Work experience varies widely across gender, for reasons often unrelated to discrimination in the labor market, such as family obligations. Failure to account for work experience in a gender wage gap decomposition or using a poor proxy such as the number of years since the end of education in case this variable is missing from the data, would lead to underestimating the explained portion of a gender wage gap and overestimating the role of discrimination.

TABLE 8.4

Gender wage gap among the NLSY cohort, ages 35–43 in 2000. No diploma or GED is the reference group. All female coefficients are significant at the 10% level.

Dependent: Log hourly wage	Mean of variable $\bar{\mathbf{x}}$		Male coefficient	Female coefficient	Decomposition of $(\overline{\ln w_A} - \overline{\ln w_B})$
	Male ($\bar{\mathbf{x}}_A$)	Female ($\bar{\mathbf{x}}_B$)	$\widehat{\boldsymbol{\beta}}_A$ in eq. (8.9)	$\widehat{\boldsymbol{\beta}}_B$ in eq. (8.10)	"Explained" $(\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B)\,\hat{\boldsymbol{\beta}}_A$
Education	—	—	—	—	−.012
< 10 yr	.053	.032	−.027 (.043)	−.089 (.050)	−.001
10–12 yr (no diploma or GED)	.124	.104	—	—	—
HS grad (diploma)	.326	.298	−.013 (.028)	−.002 (.029)	−.000
HS grad (GED)	.056	.045	.032 (.042)	−.012 (.044)	.000
Some college	.231	.307	.164 (.031)	.101 (.030)	−.012
BA or equiv. degree	.155	.153	.380 (.037)	.282 (.036)	.001
MA or equiv. degree	.041	.054	.575 (.052)	.399 (.046)	−.007
PhD or prof. degree	.015	.007	.862 (.077)	.763 (.100)	.007
Lifetime work experience	—	—	—	—	.137
Years worked civilian	17.160	15.559	.038 (.003)	.030 (.002)	.061
Years worked military	0.578	0.060	.024 (.005)	.042 (.013)	.012
Part-time work	0.049	0.135	−.749 (.099)	−.197 (.049)	.064
Number of observations	2655	2654			

Note: All dependent variables are dummy variables equal either to 1 or to 0, except for lifetime work experience, years worked civilian, and years military. The first two columns display the average value of the corresponding variable. By definition, the average value of a dummy variable equals the share of the population for which the dummy variable equals 1. OLS regressions, also including age, race, region, city, AFQT, sectors, and nonparticipation due to family (coefficients not shown here). Standard errors in parentheses. For education, "10–12 years, no diploma or GED" is the reference group.

Source: O'Neill and O'Neill (2006, table 10), and Fortin et al. (2011, table 2).

men and women. The level of education across gender is rather close, with a slightly higher level of upper education for women. The difference in work experience is more drastic: women work on average 1.5 years less than men at these ages, while the incidence of part-time is three times greater. The third and fourth columns then show the male and female coefficients respectively associated with these variables in the regressions of equations (8.9) and (8.10), which also include variables such as age, location, sector of occupation, and AFQT score. These coefficients differ across gender, with slightly lower returns education for women. The estimates are then used to compute the difference of the log of hourly wage between men and women, using men as the reference group, as explained by the different levels in education and work experience, if women had the same returns to work and education as men. This is shown in the last column. For example, if k is an index number referring to one level of education—corresponding to one variable in the equation (8.9)—the term $\left(\overline{x}_{\widehat{A}k} - \overline{x}_{\widehat{B}k}\right)\widehat{\beta}_{Ak}$ measures

the difference in wages due to differences in level of education, assuming that the returns to education are identical for men and women. If there are K education variables in the model, each representing different levels of education, the total contribution of education is the sum of the contribution of all dummy variables $\sum_{k=1}^{K} \left(\overline{x}_{\widehat{A}k} - \overline{x}_{\widehat{B}k}\right) \hat{\beta}_{Ak}$. This assumes of course that the underlying wage structures are additively separable functions of the workers' observable characteristics. This sum is equal to –0.012 log points for the whole set of education variables (women being slightly more educated), but to 0.137 log points for the whole set of lifetime work experience variables (women having less experience), which is 10 times larger. Note that the same type of detailed decomposition could be done for the wage structure effect $\overline{\mathbf{x}}_B \left(\hat{\boldsymbol{\beta}}_A - \hat{\boldsymbol{\beta}}_B\right)$ for each variable, after running equation (8.10) for females and then using the male and female regression coefficients.

This type of decomposition is then done for all variables used in the model to estimate the explained and unexplained components of the wage gap. The wage gap between men and women, the target of the analysis, amounts to a bit more than 23%[8] in the sample (which corresponds approximately to 0.233 log point). Results are shown in column (1) of table 8.5. Work experience is by far the largest contributor to the explanation of the wage gap, representing about 60% of the wage gap and 70% of the part that can be explained by individual characteristics. The AFQT score does not explain much more in absolute terms than education factors do. The same holds for industrial sector dummies, which might be influenced by discrimination anyway (and could thus be eliminated from these regressions). Using males as the reference group, about 20 percentage points of the wage gap is explained by composition effects, and only about 3.6 percentage points are left unexplained, including the possibility of discrimination.

TABLE 8.5

Decomposition of the gender wage gap among the NLSY cohort, ages 35–43 in 2000. All coefficients are significant at the 10% level.

Decomposition of the wage gap $\overline{\ln w_A} - \overline{\ln w_B}$	Using male	Using female	Weighted	Pooled
	(1)	(2)	(3)	(4)
Unadjusted mean log wage gap	.233	.233	.233	.233
Composition effect, controlling for:				
Age, city, region, race	.012	.009	.011	.010
Education	−.012	−.008	−.010	−.010
AFQT	.011	.011	.011	.011
L.T. withdrawal due to family responsibilities	.033	.035	.034	.028
Lifetime work experience	.137	.087	.112	.092
Industrial sectors	.017	.003	.010	.009
Total "explained" by model	.197	.136	.167	.142
Total "unexplained" by model (incl. cst)	.036	.097	.066	.092

Note: OLS regressions. L.T. = Long Term.

Source: O'Neill and O'Neill (2006, table 11) and Fortin et al. (2011, table 3).

[8] The unadjusted mean log wage gap between males and females is 0.233 log point in table 8.5, which is approximately 23%.

Changing the Reference Group Is Not Neutral

Utilization of the Blinder-Oaxaca method raises a serious problem, to the extent that the portion explained by discrimination using this method depends on the reference group chosen to build the counterfactual.[9] If women (group B) is the reference group instead of men (group A), adding and subtracting the average counterfactual wage that group A workers would have earned under the wage structure of group B, $\overline{\mathbf{x}}_A\widehat{\boldsymbol{\beta}}_B$, to equation (8.12) gives:

$$\hat{\Delta} = (\overline{\mathbf{x}}_A - \overline{\mathbf{x}}_B)\,\widehat{\boldsymbol{\beta}}_B + \overline{\mathbf{x}}_A\left(\widehat{\boldsymbol{\beta}}_A - \widehat{\boldsymbol{\beta}}_B\right)$$

The explained part of the differences in average values of the wage logarithms is no longer $(\overline{\mathbf{x}}_A - \overline{\mathbf{x}}_B)\widehat{\boldsymbol{\beta}}_A$, but $(\overline{\mathbf{x}}_A - \overline{\mathbf{x}}_B)\widehat{\boldsymbol{\beta}}_B$. Obviously, this part (as well as the unexplained part) depends on the reference group chosen. Hence if the returns to individual characteristics (notably experience, seniority, profession) of men are higher, and if this group is also endowed with better characteristics on average, the explained part $(\overline{\mathbf{x}}_A - \overline{\mathbf{x}}_B)\widehat{\boldsymbol{\beta}}_A$ is greater than $(\overline{\mathbf{x}}_A - \overline{\mathbf{x}}_B)\widehat{\boldsymbol{\beta}}_B$, and the extent of discrimination against women gauged by taking men as a reference is weaker than if the other group is taken. This is exactly what we observe in the sample used by O'Neill and O'Neill. Column (2) of table 8.5 reproduces the decomposition exercise, this time taking women as the reference group. The wage gap left unexplained by the model now amounts to about 10% and, as expected, work experience has less weight in the composition effects.

Several studies have tried to solve this problem by proposing a more general form of the Blinder-Oaxaca decomposition. The idea is no longer to take any one group as reference but to assign each group an arbitrary weight. Let us continue to denote $\hat{\Delta} = \overline{\ln w_A} - \overline{\ln w_B}$, the decomposition is then written:

$$\hat{\Delta} = (\overline{\mathbf{x}}_A - \overline{\mathbf{x}}_B)\,\hat{\boldsymbol{\beta}} + \overline{\mathbf{x}}_A\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_A\right) + \overline{\mathbf{x}}_B\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_B\right), \qquad \hat{\boldsymbol{\beta}} = \lambda\hat{\boldsymbol{\beta}}_A + (1-\lambda)\hat{\boldsymbol{\beta}}_B$$

In this equation, $\lambda \in [0,1]$ designates the relative weight of group A (here, men) in the definition of the reference group, and $\hat{\boldsymbol{\beta}}$ is interpreted as the vector of the returns to observable variables, like education or work experience, in a competitive market. The weight λ is clearly very hard to define, and a large area of arbitrariness always subsists (see Cotton, 1988; Oaxaca and Ransom, 1994; and the summary of Kunze, 2008). A popular choice is to use the share of the two groups in the population.[10] The results of this method are shown in column (3) of table 8.5 (using a relative weight $\lambda \simeq 0.5$). As expected, the results are in between those obtained using solely males or solely females as the reference group.

An alternative measure of the unexplained wage gap can be arrived at by wage equations of the type studied in the previous section, like (8.4), based on the pooled

[9] In principle, other counterfactuals could be based on hypothetical states of the world: what would be the wage of type B workers according to some nondiscriminatory wage structure, or what woud be the mean wage if there were no type B workers? These counterfactuals involve general equilibrium effects. They are not simple counterfactuals on which standard Blinder-Oaxaca decompositions can rely.

[10] Another possibility is to use the variation in the observable characteristics of the two groups in the population (see Fortin et al., 2011, p. 47).

sample of men and women. Bearing in mind that in equation (8.4) the dummy is equal to 1 for women and 0 for men, the difference between the average values of the wage logarithms obtained on the basis of this equation is written:

$$\Delta = \mathbb{E}(\ln w_A) - \mathbb{E}(\ln w_B) = [\mathbb{E}(\mathbf{x} \mid \mu = 1) - \mathbb{E}(\mathbf{x} \mid \mu = 0)]\boldsymbol{\beta} + \alpha$$

Once estimated using the pooled data of men and women, the difference would decompose as:

$$\hat{\Delta} = (\overline{\mathbf{x}}_B - \overline{\mathbf{x}}_A)\widehat{\boldsymbol{\beta}} + \widehat{\alpha} \tag{8.14}$$

The partial regression coefficient $\widehat{\alpha}$ is then interpreted as reflecting the wage differential between women and men. The characteristics included in the regression are the same as for the Blinder-Oaxaca decomposition. But this method assumes that the effect of observable characteristics (other than gender) on wages can be approximated by the average effect for the two groups. As shown by the last column of table 8.5, once adjusted for observable characteristics, the wage gap comes down to less than 10% ($\widehat{\alpha}$ is 0.092; see the last line of the table), a figure higher than the one obtained by the Blinder-Oaxaca decomposition using males as the reference group, and close to that using women as the reference group (but this is not necessarily always the case).

Selection Biases

The selection bias due to the fact that we only observe the wages of persons who work is still an issue in the decomposition methods. The question is whether those excluded from the sample are different from those included to the point where estimates would be biased. In the case of O'Neill and O'Neill (2006), out of the entire cohort of men, 74% of white men were included in the sample compared to 68% of black men and 73% of Hispanic men. The proportion of women included in the analysis was 66% for white women, 68% for black women, and 63% for Hispanic women. So a larger proportion of women was excluded from the analysis. Women excluded because they had no reported wage in the last two years were almost as large a group as those who reported wages, contrary to men, who most often reported a wage in the last two years. This means that exclusion probably relates to different causes in the case of men and women. However, O'Neill and O'Neill do not correct the selection bias in their study.

We have seen that Neal and Johnson (1996) use a simple method assigning almost zero hourly wages to nonparticipant black men. This approach would not be appropriate for women because the causes of nonparticipation might be unrelated to skill levels. Blau and Kahn (2006) have proposed a method, using panel data, that proceeds in several steps: they first recover past earnings of nonparticipants when available for the most recent year, thanks to the panel dimension of their data, and then they assume that individuals with at least a college degree and at least eight years of actual full-time labor market experience had wage offers above the median for their gender, and that those with less than a high school degree and less than eight years of actual full-time labor market experience had wage offers below-median for their gender. This way they estimate a wage gap.

To control for selection bias, it is also possible to estimate wages and participation simultaneously by adding the inverse Mills ratio obtained by fitting a probit model of the participation decision as a regressor in equations (8.9) and (8.10) (see section 3.1.3).

The "exclusion restriction" still requires that an available instrument Z explaining participation should be included in the participation equations but not in the wage equations (see chapter 1, appendix 7.5). The control variables obtained for each group (giving the probability to participate in the labor market for each individual), denoted $\lambda_A(X_A, Z_A)$ and $\lambda_B(X_B, Z_B)$, are added to the decomposition in the following manner:

$$\overline{\ln w_A} - \overline{\ln w_B} = (\overline{\mathbf{x}}_A - \overline{\mathbf{x}}_B)\widehat{\boldsymbol{\beta}}_A + \overline{\mathbf{x}}_B\left(\widehat{\boldsymbol{\beta}}_A - \widehat{\boldsymbol{\beta}}_B\right) + \overline{\lambda}_B(\widehat{\sigma}_A - \widehat{\sigma}_B) + \widehat{\sigma}_A(\overline{\lambda}_A - \overline{\lambda}_B)$$

where $\widehat{\sigma}_A$ and $\widehat{\sigma}_B$ are the estimated coefficients for λ_A and λ_B included in equations (8.9) and (8.10) respectively.

In sum, the different methods of decomposing the wage gaps between demographic groups can give different results when applied to the same sample, so it is important to identify and clearly define the hypotheses of every empirical study, in order to be able to interpret, and eventually compare, assessments of discrimination. For a detailed presentation of the decomposition methods, notably for other distributional statistics than the mean (median, variance, quantiles), see Fortin et al. (2011).

3.2.2 How to Estimate Changes in Discrimination

In the United States, wage inequalities between men and women have had a tendency to shrink during the 1980s and 1990s (Blau and Kahn, 1997; Fortin and Lemieux, 1998). This fact is surprising, for if one ponders the overall distribution of wages, inequalities have mounted sharply over the same period. Less skilled workers in particular have undergone relative losses of purchasing power. Why has the relative position of women improved, when on average they hold less skilled jobs than men? Is it the consequence of reduced discrimination or the result of an improvement in their relative productivity over time? In order to understand the dynamics of wage inequalities and the role of discrimination, several studies have utilized the decomposition of the evolution of wage differences *over time* introduced by Juhn et al. (1993), which makes it possible to separate the between- and within-group components. We begin by explaining the principles of this decomposition and then go on to emphasize the detrimental consequences of selection biases in this type of research.

Decomposing the Impact of Both Observables and Unobservables: The Method of Juhn, Murphy, and Pierce

Juhn et al. (1993) begin by estimating the wage equation (8.9) for a demographic reference group A (men, for example) at date t. For an individual i of group A at date t, this equation takes the form:

$$\ln w_{Ait} = \mathbf{x}_{Ait}\boldsymbol{\beta}_{At} + \varepsilon_{Ait} \qquad \textbf{(8.15)}$$

Juhn et al. (1993) decompose the statistical residual by assuming constant returns to unobservables: $\varepsilon_{Ait} = \sigma_{At}\theta_{Ait}$, where $\sigma_{At} = \sqrt{var(\theta_{Ait})}$ is the standard error of the residuals of the distribution of wage logarithms at date t of the members of group A. The error term θ_{Ait} is interpreted as a standardized residual with zero mean and unitary variance. The estimation of equation (8.15) by ordinary least squares for the members of group A gives the estimated values $\hat{\boldsymbol{\beta}}_{At}$, $\hat{\theta}_{Ait}$, and $\hat{\sigma}_{At}$. Let θ_{At} and $\mathbf{x}_{At}$ be respectively the average

of the θ_{Ait} and of the $\mathbf{x}_{Ait}$; the average of the wage logarithms of group A, denoted $\overline{\ln w_{At}}$, is then defined by the equality:

$$\overline{\ln w_{At}} = \overline{\mathbf{x}}_{At}\hat{\boldsymbol{\beta}}_{At} + \hat{\sigma}_{At}\hat{\theta}_{At}$$

Juhn et al. (1993) then assume that the coefficients $\hat{\boldsymbol{\beta}}_{At}$ and the variance of the residuals $\hat{\sigma}_{At}$ are identical for the two groups, or $\hat{\boldsymbol{\beta}}_{At} = \hat{\boldsymbol{\beta}}_{Bt} = \hat{\boldsymbol{\beta}}_t$ and $\hat{\sigma}_{At} = \hat{\sigma}_{Bt} = \hat{\sigma}_t$. The latter is a strong assumption (see below). Let us introduce the difference operator Δ, defined by $\Delta y_t = y_{At} - y_{Bt}$; the difference of the average values of the wage logarithms between groups at date t is written thus:

$$\Delta\overline{\ln w_t} = \overline{\ln w_{At}} - \overline{\ln w_{Bt}} = \Delta\overline{\mathbf{x}}_t\hat{\boldsymbol{\beta}}_t + \hat{\sigma}_t\Delta\hat{\theta}_t \qquad \textbf{(8.16)}$$

This equation indicates that the wage differential between the two groups includes a $\boldsymbol{\beta}$ component arising from the differences of observable characteristics, $\Delta\overline{\mathbf{x}}_t\hat{\boldsymbol{\beta}}_t$, and a component that results from the differences in the standardized residuals, $\Delta\hat{\theta}_t$, between the members of the two groups. The term $\hat{\sigma}_t\Delta\hat{\theta}_t$ is interpreted as the differences unexplained by the observable variables and which can therefore be attributed to discrimination. Equation (8.16) then gives the difference observed at dates t and s between the intergroup wage differentials. The result is:

$$\Delta\overline{\ln w_t} - \Delta\overline{\ln w_s} = (\Delta\overline{\mathbf{x}}_t - \Delta\overline{\mathbf{x}}_s)\,\hat{\boldsymbol{\beta}}_t + \Delta\overline{\mathbf{x}}_t\left(\hat{\boldsymbol{\beta}}_t - \hat{\boldsymbol{\beta}}_s\right) + \left(\Delta\hat{\theta}_t - \Delta\hat{\theta}_s\right)\hat{\sigma}_t + \Delta\hat{\theta}_t\,(\hat{\sigma}_t - \hat{\sigma}_s) \qquad \textbf{(8.17)}$$

The first term of the right-hand side measures "the effect of changes in observed characteristics" and represents the contribution of changes in the averages of observable characteristics of the members of the two groups between dates s and t. The second term is "the observed price effect" and represents the contribution of differences in returns to characteristics observed at date t. The two last terms account for changes in the unexplained gender pay gap. The third term measures "the effect of changes in unobserved characteristics" and corresponds to the change in the average relative position of the members of the two groups in the distribution of wages that is not due to observed variables. Finally, the fourth term is an "unobserved price effect."

This method allows us to pinpoint the contributions of the different components of the changes in the gaps in average wage between demographic groups. It has played an important role in the inequality literature, ever since Juhn et al. (1993) concluded that most of the growth in inequality from the 1960s to the 1980s was linked to the residual component $\hat{\sigma}_t\Delta\hat{\theta}_t$, reflecting increased returns to components of skills other than those observed (years of schooling and work experience). Based on the same method but controlling additionally for selection bias, Blau and Kahn (2006) find that the shrinkage of the gender pay gap in the 1990s compared to the 1980s is not due to observable factors (human capital) but mostly to a reduction of the unexplained factors, including discrimination (see table 8.8 for detailed results). Like all decompositions, it rests on arbitrary hypotheses of which we need to be aware in order to gauge its significance. First, as we have already pointed out in studying the Blinder-Oaxaca method, the distinction between observed characteristics and unobserved characteristics does not capture exclusively phenomena linked to discrimination. It also reflects, among other things,

measurement errors, specification problems, and the existence of omitted variables. Moreover, the choice of reference group is as critical as in the Blinder-Oaxaca decomposition (Fortin and Lemieux, 1998). Finally, the hypothesis according to which the variation in residuals of the two demographic groups is identical (i.e., $\sigma_{At} = \sigma_{Bt} = \sigma_t$), which is indispensable to be able to distinguish "the effect of changes in unobserved characteristics" from "the unobserved price effect," is highly debatable (see Suen, 1997; the assessment of Blau and Kahn, 2003; Fortin et al., 2011). It means notably that the residuals are totally independent of the observed characteristics of individuals, which differ across groups (a stronger assumption than the conditional independence assumption set forth above).

3.3 Direct Assessment of Discrimination

Assessing the extent of discrimination by estimating wage equations poses problems that are difficult to overcome. The inevitable existence of omitted variables would point to the conclusion that the results obtained always *overestimate* discrimination. Another problem has to do with the influence of discrimination on explanatory variables. Our theoretical analysis in section 2.2.2 has shown that discrimination may discourage education, and more generally, any investment leading to increased incomes. In consequence, the observation of a lower level of education may be caused by discrimination. In this case, the results obtained through estimating wage equations may *underestimate* discrimination. These limitations justify the use of alternative methods that aim to estimate discrimination directly. We give the broad outlines of three of these approaches: correspondence and audit studies, experimentation, and a method based on the comparison of productivity differences and wage differences.

3.3.1 Audit Studies

Audit studies consist of having individuals with fictitious resumes who are identical except for race, sex, or ethnicity apply for jobs and go to interviews. The main interest of these methods is to ensure that the productivity revealed to employers is identical among individuals belonging to different groups, so that there is no need to control for productivity-related factors in equation (8.4). Hence the coefficient associated with the group dummy in this equation would reflect only discrimination. Most often, the outcome of interest in this type of study is the chances of getting a callback, or the chances of getting a job offer, rather than the wage offered. Beyond discrimination, it has been applied as a test to various economic theories in the recent past (see Levitt and List, 2009, for an overview). Despite its popularity, this approach also has a number of drawbacks and relies on strong identifying assumptions.

Audit studies consist of setting up experiments in order to compare the performance in the labor market of individuals who are identical, except for their membership in a clearly specified group. To that end, the investigator pairs off individuals who belong to different groups but who have the same individual characteristics in terms of education and social origin and who go about their job search in exactly the same way (for a summary presentation of the relevant work, see Darity and Mason, 1998; Altonji and Blank, 1999; OECD, 2008). For instance, potential employers, selected at random, are sent resumes that are identical in every respect (that could signal individual

productivity) except for an indication of group membership, such as the spelling of the name, the gender, or the residential address. As noted, the measured outcome is most often not the wage but the probability of getting an interview or being hired.

Suppose that productivity y of an individual on a given job depends on a set of individual observable and unobservable characteristics $(\mathbf{x}, \varepsilon)$ and on the characteristics of the firm, represented by the scalar f, where she is employed. For simplicity let us assume that $y = \mathbf{x}\boldsymbol{\beta} + \varepsilon + f$. Let μ be a dummy for group membership, which has value $\mu = 0$ if the individual belongs to group A and $\mu = 1$ if the individual belongs to group B. Now, discrimination is identified if individuals with equal productivity receive different treatment. Since $y = \mathbf{x}\boldsymbol{\beta} + \varepsilon + f$, to ensure that individuals have equal productivities, we need to assume that the expected unobservable content of productivity is the same across groups, which means that $\mathbb{E}(\varepsilon \mid \mu = 1) = \mathbb{E}(\varepsilon \mid \mu = 0)$. This is the key identifying assumption in conventional audit studies. Let us designate by T_i the treatment of individual i by the firm (getting a job offer, for instance); the degree α of discrimination can be obtained by regressing the outcome on a constant and the group dummy μ using the OLS:

$$T_i = \gamma + \mu_i \alpha + \varepsilon_i \quad \text{with } \mathbb{E}(\varepsilon) = 0 \tag{8.18}$$

But the assumption that $\mathbb{E}(\varepsilon \mid \mu = 1) = \mathbb{E}(\varepsilon \mid \mu = 0)$ is a strong one, as Heckman and Siegelman (1993) and Heckman (1998) argued. Despite the experimenter's best efforts, it is difficult to ensure that randomly assigned individuals act in exactly the same way with the various employers they have to meet or that employers perceive them in exactly the same way. Actually, this can lead to large biases if individuals are standardized on the other observable characteristics related to productivity. Moreover, it is possible that individuals with identical observable characteristics belonging to different demographic groups do have different productivities.

In general, investigations carried out in the United States following this method find that whites are more frequently given the opportunity to take hiring tests than blacks or Hispanics, and also receive more job offers. Moreover, whites have access to better jobs than blacks or Hispanics. The study by Goldin and Rouse (2000), which looks at the effects on hires of a change in recruitment policy by major symphony orchestras, follows a very similar approach. In order to guarantee the impartiality of the judging panels, the musicians audition behind an opaque screen. This type of audition was introduced by the Boston Symphony Orchestra in 1952 and has been adopted by many orchestras in the 1970s and 1980s. Goldin and Rouse find that the presence of the screen significantly increased the number of women hired and that it explains almost one quarter of the increased presence of women in symphony orchestras in the 1970s and 1980s.

3.3.2 Correspondence Studies

To circumvent some of the critiques by Heckman and Siegelman (1993) described above, some researchers avoid sending applicants to interviews. Instead, they send applications at random featuring similar productive characteristics but indicating different group membership. Then they compute the simple difference in the number of callbacks across groups, using group averages or equivalently a regression similar to (8.18). For instance, Bertrand and Mullainathan (2004) rely on written applications only, with

names that signal race without explicitly stating it. They assign the same resume at random with different names, some that sound white (such as Emily Walsh or Greg Baker) and others that sound African American (such as Lakisha Washington or Jamal Jones) in response to newspaper ads in Chicago and Boston. The quality of the resumes used also varies at random in response to a given ad (a little more labor market experience, fewer holes in their employment history, a degree completed, etc.). Four resumes are sent in response to each ad: two of higher quality and two of lower quality. One of the higher- and one of the lower-quality resumes is assigned at random to a name that sounds African American. In total, 1,300 employment ads in sales, administrative support, clerical, and customer services were answered and nearly 5,000 resumes were sent. The results are striking: whereas applicants with "white" names need to send about 10 resumes to get one callback, African American names need to send around 15 resumes. This 50% gap in callback rates is statistically very significant and stable across industries and occupations. Race also affects the reward to having a better resume: whites with higher-quality resumes receive 30% more callbacks than whites with lower-quality resumes, but the impact of a high-quality resume is much smaller for African Americans.

Unfortunately, such correspondence studies are not exempt from limitations either. First, as with audit studies, it is possible that individuals with identical observable characteristics belonging to different demographic groups do have different productivities. Second, parameter α in (8.18) does not distinguish between taste discrimination and statistical discrimination. Indeed, the answers (or the absence of answers) from potential employers might well reflect differences in the productivity they expect from different groups, despite the absence of differences in characteristics displayed in the applications. A third limitation of the correspondence approach is that it is difficult to generalize the experimentally measured differences in outcomes. If, for instance, members of the group discriminated against react to discrimination by sending more resumes to compensate for a lower response rate by employers, or if they select themselves across firms to minimize their contacts with discriminatory employers, then the impact of discrimination in the market in equilibrium ought to be different—higher or lower—from the differential outcome measured in the audit experiment.

Another limitation is that even if groups have the same observed and unobserved characteristics on average ($\mathbb{E}(\varepsilon \mid \mu = 1) = \mathbb{E}(\varepsilon \mid \mu = 0)$), a correspondence study (but also an audit study, if we consider that ε is not fully revealed during an interview) could in some cases generate spurious evidence of discrimination if the distributions of unobserved characteristics are not exactly identical across groups. To grasp this, let us take the example of a treatment which is not a continuous function of productivity, such as the hiring decision (as opposed to the wage). Imagine a jumping contest with a bar set at level c. Assume that the performance of jumpers depends on both their height x, which is easily observable, and their technique ε, which cannot be observed ex ante. Suppose that individuals in both groups are of exactly equal height, and also of equal technique on average, but that the technique among jumpers in group B is considered to have more variance than that among jumpers in group A. In that case, if the bar is set at a low level, one will prefer to choose a jumper from group A because, given the low variance of technique in this group, the individual picked will be more likely to pass the easy bar. But if the bar is set at a high level, then one will prefer jumpers of group B, where variance in technique is larger, hoping for good luck in picking a jumper with a technique good enough to pass the difficult bar (see Heckman, 1998). Thus, depending

on the distribution of unobserved characteristics for each group and the standardization of the study (i.e., the individuals presented in the resumes are highly, or only moderately, qualified for the type of jobs considered), the method can yield positive, negative, or no difference in the measured outcome, even if there is no discrimination in the market. Neumark (2012) presents a method for identifying discrimination in a setting where the variations in the productivities of the two groups are different, and where the resumes contain observable characteristics that vary within each group. Applied to Bertrand and Mullainathan's (2004) data, this method leads to even stronger evidence of race discrimination that adversely affects blacks.

Despite their limitations, correspondence studies are highly developed at present and have brought to light important differences in the probabilities of the responses received by groups differentiated by gender, race, ethnicity, and religion (Adida et al., 2010), by caste (Banerjee et al., 2009), or by sexual orientation (Tilcsik, 2011), as we will see in section 4. In sum, field experiments make it possible to throw into relief hiring behaviors that treat different demographic groups differently. Such behavior may arise out of either taste discrimination or statistical discrimination: the question remains open, inasmuch as neither the distributions of productivity within two contrasting groups nor the beliefs of employers about these distributions are known to the econometrician.

3.3.3 Laboratory Experiments

Another method, one little used at present, is based on setting up laboratory experiments. Fershtman and Gneezy (2001) have used it to study ethnic discrimination in Israel, bringing Ashkenazic students (descendants of European and American Jewish communities) and Oriental Jews (descendants of Jewish communities in Asia and Africa) together to participate in the *trust game* and the *dictator game*.

The trust game is a game with two players, in which player A holds a sum of money and must decide how much to hand over to player B. The experimenter triples the amount of the transfer and gives it to player B, who can decide to make a gift in return to player A. Within these rules, the efficient outcome—the one, that is, that gives both players the maximum of resources—dictates that player A should transfer the whole sum he holds to player B, so that B will receive the maximum from the experimenter. But on the assumption that both players are rational egoists, running the game in a noncooperative context makes it impossible to reach this efficient outcome. Player A, foreseeing that player B has no interest in returning anything at all to him in the final stage of the game, lacks any motive to give B a positive amount in the first stage. So the solution to this noncooperative game is a zero transfer or, in more technical terms, the zero transfer is the only subgame perfect equilibrium.[11] The experiments carried out by Fershtman and Gneezy (2001) reveal, however, that player A does give a positive amount and that player B frequently gives back a greater amount. To be precise, Fershtman and Gneezy (2001) organize their experiment this way: the role of player A is assigned to students whose ethnic origin is not specified, and the role of player B to other students whose names sound either Ashkenazic or Oriental. Fershtman and Gneezy find that individuals with Oriental backgrounds receive lower transfers than others.

[11] See chapter 7, section 2.3.1, for a definition of this concept.

This result seems not to flow from statistical discrimination, inasmuch as the behavior of players B during the running of the game is no different, whether they are Ashkenazic or Oriental. Fershtman and Gneezy use the dictator game to show that their result does not flow from a taste for discrimination either. In the dictator game, player A decides on a transfer to player B, who is unable to give anything back to player A. The amount of player A's gift is always tripled by the experimenter, which can only benefit player B, who has no strategic role in this game. On average, the experiment shows that those playing B obtain the same transfers whether they are Ashkenazic or Oriental. Hence it appears that the reduced gains of those of Oriental descent in the trust game do in fact result from a problem of trust on the part of player A and not a taste for discrimination. Fershtman and Gneezy conclude that the discrimination springs from the existence of groundless stereotypes. What is more, when this experiment is run on a female population, women do not engage in discriminatory behavior; therefore, they do not subscribe to the same stereotypes as men. The particular interest of this approach is that it sheds precise light on the origins of certain forms of discrimination.

3.3.4 Field Experiments

Experiments can also be used to complement and interpret results obtained from the field. List (2004) recruited buyers and sellers at a baseball card show. Buyers and sellers were asked to trade a highly valued card, namely the "1989 Upper Deck Ken Griffey Jr. PSA graded '9' baseball card," for a small monetary reward. Buyers interested in this card were asked to purchase it from dealers for the lowest possible price below a predefined reservation price. Sellers, who were not dealers and who possessed this card, were asked to sell it to dealers at the highest possible price above a predefined reservation price. Unlike audit studies, participants were not informed, so as to avoid potential bias in behavior, that the purpose of the experiment was to detect discrimination (Heckman, 1998). After the trading took place, List observed 240 outcomes for buyers and 300 outcomes for sellers, recording both the initial and final offers. List controlled for subjects' experience in the card market, as well as their age (age 20–30 or 60 and older), ethnic origin (whites and nonwhites), gender, income and education, and also the trading time. He finds that there is a significant price differential in the initial offer if seller or buyer belonged to a minority, with greater discrimination among sellers. For instance, females and older buyers received initial offers that were more than 10% higher than those received by white males aged 20–30 (the majority of participants), while nonwhites, females, and older sellers would receive initial offers 30% lower. Final transactions, after the buyer and seller had negotiated for a while, revealed less (but still significant) discrimination: 6%–8% for minority buyers and 18%–20% for minority sellers. Regardless of whether dealers were buying or selling, these differences are more prevalent among experienced dealers.

Now, this field experiment alone cannot identify what drives the measured price differences. Three potential explanations may be invoked: (1) prejudice against minorities (taste-based discrimination), (2) unobserved differences in bargaining ability, or (3) statistical discrimination (stemming from the lack of information on reservation prices). To sort this problem out, List recruited about half of the dealers who participated in the previous field experiment and asked them to play a $5 dictator game with nondealers belonging to various groups. In this game, dealers simply state what the split of the $5 will be and the responder has no veto power. But dealers are informed about

the group membership of their partner (sex, age, and race). Thus any observed difference in the final split among groups should stem from pure prejudice. List finds no statistically significant differences: dealers do not exhibit tastes for discrimination that would systematically favor the majority group.

To gain more insight into bargaining differences across groups, List also ran an experiment in which buyers and sellers negotiated over the trade price. Under this setup, dealers know that they are part of an experiment and either know that buyers' reservation prices are drawn randomly (first type of treatment) or are given no such information about the buyers' reservation prices (second type of treatment). He finds that when dealers were told that reservation prices were randomly assigned, the final prices offered were unrelated to minority status. By contrast, when dealers were not given this information and were thus unsure about buyers' reservation prices, they tended to offer lower prices to minorities, like in the initial field experiment, and despite the fact that dealers knew they were being observed. This shows that the dealers' behavior is not driven by their belief that minorities are less effective bargainers, but it may reflect their beliefs about the distribution of the reservation prices. Indeed, if dealers know that the variance in the reservation prices is larger in minority groups, they may attempt to secure deals with high (low) reservation value agents when selling (buying) their cards. List used a price auction to elicit buyers' reservation prices for the same "Upper Deck Ken Griffey Jr. PSA" card and finds that minority reservation price distributions are indeed more disparate than those of the majority. Yet it is dealers' perceptions of these distributions that drive behavior. Thus List asked 60 dealers to determine to which group the reservation price distributions they were shown belonged. It turned out that dealers generally matched these correctly, with the experienced dealers being more informed about the disparities. Overall, using experiments, List provides strong evidence that at least for some agents, in some markets, transactions seem to be based on statistical inference about how reservation prices vary among groups, rather than on prejudice.

3.3.5 Productivity Differences and Wage Differences

Another method used to evaluate discrimination consists in evaluating differences in actual productivity and comparing them to wage differences. Data available in the field of sports have made it possible to study wage discrimination between athletes of different ethnic origin (see Kahn, 1991, 2000, for summaries). Discrimination of this kind against blacks was brought to light in the National Basketball Association. For instance, Price and Wolfers (2010) find that NBA referees, who observe the performance of players and award fouls and points, have preferences for players of their own race, despite being closely scrutinized by observers. The authors control for the performance of players, as well as the racial composition of the referee teams, which is set at random. They then compare the number of fouls, the number of minutes played, and the number of points accrued by white and black players respectively, and they study how these differences vary with the racial composition of the refereeing crew. They find systematic evidence of an own-race bias by referees: players earn up to 4% fewer fouls or score up to 2.5% more points when they benefit from a positive own-race bias, rather than a negative opposite-race effect. Results at the team level are similar, with the racial composition of the refereeing crew increasing the probability of a team winning.

Hellerstein et al. (1999) have tried to apply methods of this kind to industry by estimating the marginal productivity differentials of workers belonging to different ethnic groups. They use surveys that match data about firms and individuals, and compare them with estimated wage differentials. The results they obtain tend to confirm assessments based on the estimation of wage equations: differences of income between ethnic groups generally correspond to differences in productivity, except for women, who on average receive wages lower than their productivity.

Another setting makes it possible to control for the "productivity" of agents in a quasi-experimental setting: game shows. A couple of papers have investigated discrimination in *The Weakest Link* (Levitt, 2004; Antonovics et al., 2005). On this show, players answer general knowledge questions. The objective is to create a chain of consecutive correct answers and so earn an increasing amount of money within a specific time limit. An incorrect answer breaks the chain and the money accumulated up to that point is lost. At the end of each round, contestants must vote one player out of the game. In early rounds, strategic incentives encourage voting for the weakest competitors but build up the amount of the reward. In later rounds, the incentives reverse, and the strongest competitors become the logical target because each contestant wants to win the jackpot. In this setting, productivity is easily controlled by measuring the number of right answers of the various players. With other characteristics controlled for, if taste-based discrimination were the key driver of votes, minorities would continue to experience excess votes in later rounds of the game, while if statistical discrimination were at stake, these votes should be fewer by the end of the game as the performance of players is revealed. In this setting, there is little evidence of discrimination against women or blacks. Using a logit model to analyze the probability that a player votes against other players, given their characteristics and controlling for the percentage of correct answers during the game, Antonovics et al. (2005) find no evidence of discrimination, except for women against men. The data are consistent with taste-based discrimination: women would simply prefer to play with other women.

4 EMPIRICAL RESULTS REGARDING DISCRIMINATION

Research on discrimination by economists has by now touched on many areas. The most important of these remain discrimination among persons of different race, gender, or ethnicity. More recently, research has widened the range of possible discrimination to take in groups whose sexual orientation diverges from a heterosexual norm and has tried to assess whether physical appearance, in other words "beauty" or size, may constitute a factor generating inequality in career paths.

4.1 Race- and Ethnicity-Related Discrimination

Racial and ethnic inequality has attracted much attention, notably in the United States. Economic research in this domain has focused principally on wage discrimination, but there are also contributions that bear on non-wage discrimination, primarily of the sort that may occur during the hiring process.

4.1.1 Wage Discrimination

Fryer (2011) recalls that in the United States "blacks earn 24% less than whites, live five fewer years, and are six times more likely to be incarcerated on a given day." As shown by figure 8.9, the unadjusted wage gap has decreased over time between blacks and whites from 35% to about 25% in 2009 for those working full-time and year-round. Identifying what in the remaining wage gap flows from discrimination in the labor market versus other factors is a challenge, as we lack credible sources of identification. Readers are reminded that wage regressions and decomposition methods, as reviewed above, arbitrarily assign unexplained differences across groups to discrimination after controlling for a variety of factors. But these results cannot be interpreted as causal. Figure 8.12 suggests however that the observed decrease in the racial wage gap went hand in hand with a decline of prejudiced attitudes. This further suggests that prejudice cannot be the only factor at stake when it comes to explaining the remaining 25% difference.

Fryer (2011) replicates the Neal and Johnson (1996) method described above, using the same NLSY data, but for the year 2006 and extending the set to include,

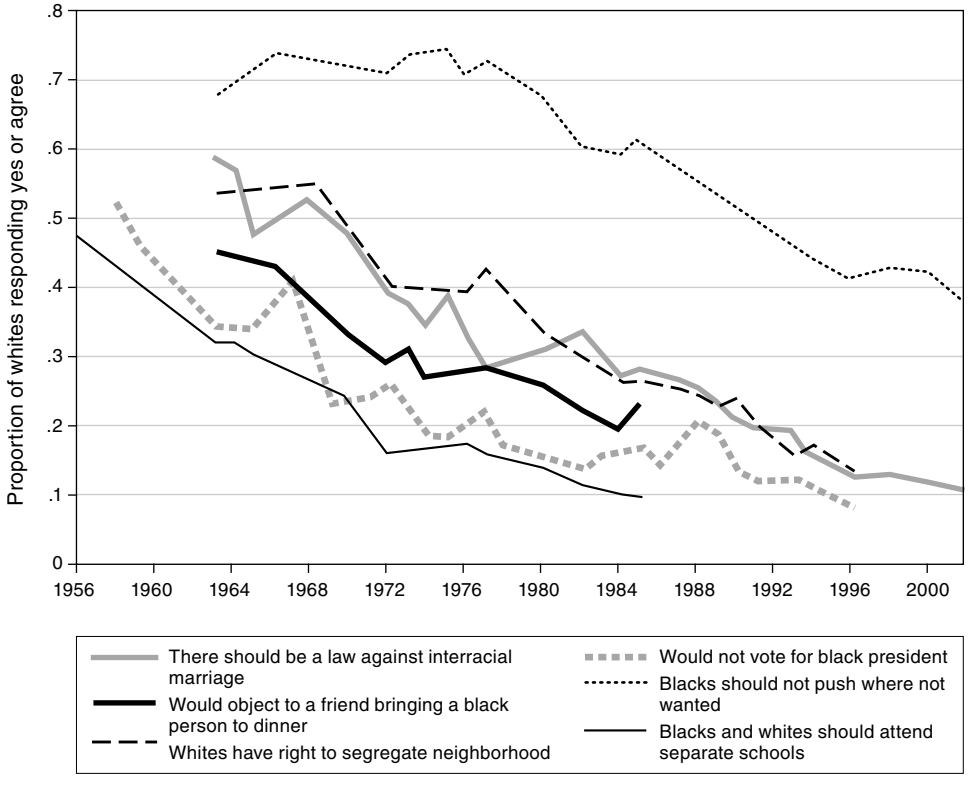


Figure 8.12
Trends in prejudiced attitudes toward blacks in the United States.

Source: Lang and Lehmann (2012, figure 3).

TABLE 8.6

Black–white wage and unemployment gap among the younger generations of the NLSY 1999 cohort (aged 42–44) and the NLSY 1999 cohort (aged 21–27) in 2006.

	Mean log wage				Unemployment			
Dependent:	Men aged 42–44		Men aged 21–27		Men aged 42–44		Men aged 21–27	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Black	−.394 (.043)	−.109 (.046)	−.179 (.023)	−.109 (.024)	2.312 (.642)	1.332 (.384)	2.848 (.377)	2.085 (.298)
Hispanic	−.148 (.049)	.039 (.047)	−.065 (.023)	−.014 (.024)	2.170 (.691)	1.529 (.485)	1.250 (.205)	.994 (.170)
Mixed race			.007 (.143)	.009 (.145)			3.268 (1.661)	3.216 (1.618)
Controlling for:								
Age	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
AFQT, $AFQT^2$	No	Yes	No	Yes	No	Yes	No	Yes
Number of observations	1167	1167	3278	3278	1315	1315	3294	3294

Note: Wages in 2006 dollars. OLS regressions, standard deviations in parentheses.

Source: Fryer (2011, tables 1 and 2).

notably, unemployment. Table 8.6 shows wage and unemployment differences for black and Hispanic men of the same generation as those studied by Neal and Johnson (but aged 42 to 44 in 2006) and for a younger generation (aged 21 to 27).

- The unadjusted wage gap between blacks and whites is 39% for the older generation but 18% for the younger generation. Taking skills into account reduces the black–white gap by three quarters, as in the original study. Taking skills into account reduces the wage gap to insignificant levels for Hispanics of both generations. The same holds good for black and Hispanic women of the younger generation, while women of the older generation have higher wages controlling for skills (not shown here). The decline of raw wage differences across generations can stem from either real gains made by blacks and Hispanics over time (notably at entry into the labor market) or a steeper wage trajectory for white males.
- Black and Hispanic men are more than twice as likely to be unemployed among those aged 42 to 44, but 30% and 50% respectively more likely after controlling for skills. Among the younger generation, black men are three times more likely to be unemployed than whites, and twice as likely after controlling for skills. For both generations, controlling for skills has less influence on employment probabilities than on wage levels. This confirms Ritter and Taylor (2011) and earlier results by Johnson and Neal (1998), who found large unexplained annual earnings differences, mostly due to disparities in hours worked.

Overall, this work and comparable studies (see Lang and Lehmann, 2012, for a review) suggest that there is a 10% wage difference between black and white men of the same age and skill level in the United States. This average hides differences across skills

and occupations though, with no difference among men with high education and AFQT scores (Lang and Manove, 2011) or among white collars with the same skills (Bjerk, 2007). Knowledge regarding women's wages is less reliable due to selection effects that affect the wage gap.

4.1.2 Non-wage Discrimination

Wage discrimination may not be the primary form of discrimination. In particular, experiments carried out to compare the performances of workers through field experiments indicate that discrimination may occur during the hiring process. On the other hand, Eckstein and Wolpin (1999) have shown, using a job search model, that discrimination in terms of *offered* wages at the time of hiring can lead to much weaker discrimination in terms of wages *accepted*. According to their estimates for high school graduates, the differences among ethnic groups in wages offered are three times greater than the differences in those accepted. These results suggest that it is not enough to focus on wages in order to detect the presence of discrimination in the labor market, and they also reveal the limits of estimations of wage equations in this area. It is necessary to take into account the histories of individuals, including their unemployment spells, the manner in which they conduct job searches, and the kinds of jobs that they wind up holding. On that basis, evidence from Fryer (2011) and Lang and Lehmann (2012) suggests that the risk of nonemployment remains much greater for blacks than for whites and has not declined over time; skills can only account for 50% of the observed difference. Most of this difference is due to nonparticipation (including incarceration, rates of which are much higher for blacks) and unemployment durations that are 30% longer for blacks.

This finding is consistent with other studies done in other countries on the relation between labor market outcomes and different ethnic origins. For instance, Aeberhardt et al. (2010) find that in France the types of jobs taken up by individuals (conditional upon their experience, background, and education) is more important in explaining wage differences than wage discrimination itself (as identified by the unexplained part of wage differences in the Blinder-Oaxaca decomposition). This suggests the existence of occupational segregation. Aeberhardt et al. use a survey containing information on professional mobility, initial education, vocational training, social origin, and earnings, and control for potential selection effect.

Correspondence studies typically identify strong discrimination in the hiring process, based on the spelling of names in the resumes. Identical resumes are sent to potential employers with names that sound like they belong either to a majority or to a minority group and the numbers of callbacks are compared. Table 8.7 reports results from some of these studies done in different countries and for different years. Two alternative measures of the discrimination rate are reported in the table. They differ with respect to the way the event of no callback for both types of applicants is treated. According to Heckman (1998), the event of no callback is equivalent to evidence of equal treatment and must be included in the denominator. By contrast, Riach and Rich (2002) argue that it provides no information and should be excluded. Whatever the measure, resumes with a minority-group sound typically get fewer callbacks. The difference between callbacks to majority and minority group resumes as a percentage of the number of jobs applied for is typically of the order of 10%.

TABLE 8.7
Results of correspondence studies of discrimination by ethnic origin, 1991–2007.

	Country / year / region	Group-identifying characteristics (minority group tested)	Net rate of discrimination % points	
			Heckman's definition[a]	Riach and Rich's definition[b]
Carlsson and Rooth (2007)	SWE / 2005–2006 Stockholm and Gothenburg	Name (Middle Eastern)	9.7***	28.9***
Bertrand and Mullainathan (2004)	USA / 2001–2002 Chicago and Boston	Name (African American)	4.9***	29.5***
Esmail and Everington (1997)	GBR / 1997 England	Name (Asian)	16.0*	27.6*
Riach and Rich (1991)	AUS / 1984–1988 Melbourne	Name (Greek and Vietnamese)	8.9***	17.7***

Notes: (a) Difference in the number of callbacks between majority and minority groups as a percentage of the number of jobs applied for (jobs for which no callback is registered are treated as providing evidence of equal treatment). (b) Difference in the number of callbacks as a percentage of jobs applied for with at least one observed callback (jobs for which no callback is registered are excluded from the sample). *, ***: statistically significant at the 10% and 1% level of confidence.

Source: OECD (2008, table 3.1).

4.2 GENDER DISCRIMINATION

Blau and Kahn (2003) find that taking education and experience into account explains only a part of the wage difference observed between genders in many countries (see section 1.1). More recently, the focus has shifted to explaining why the gender wage gap has been closing at a slower pace in recent decades. Is this due to some interruption in the evolution of productive characteristics, such as skills improvement, among females, or to a slower reduction of discrimination, or to other factors not linked to the labor market?

The study of the movement of the relative wages of women and men during the 1970s and 1980s in the United States has also drawn much attention. The wage gap between men and women remained stable between World War II and the end of the 1970s, but it shrank noticeably during the 1980s and also in the 1990s but at a slower pace, as figure 8.13 shows. Figure 8.14 also shows that this movement went along with an increase in inequalities assessed over the total distribution of wages. It may therefore seem surprising that the relative position of women, whose performance in the labor market is traditionally inferior to that of men, should have improved over this period.

The decomposition of wage gaps during the decades of the 1980s and the 1990s suggests that the overall shift in a direction favoring women results from the combined working of opposing movements. Table 8.8 portrays the decomposition carried out by Blau and Kahn (2006) on the basis of equation (8.17), based on the Michigan Panel Study of Income Dynamics (PSID) for full-time employed workers. It shows that the reduction in the wage gap between men and women over the two periods results primarily from an improvement in the observed characteristics of women, like experience and

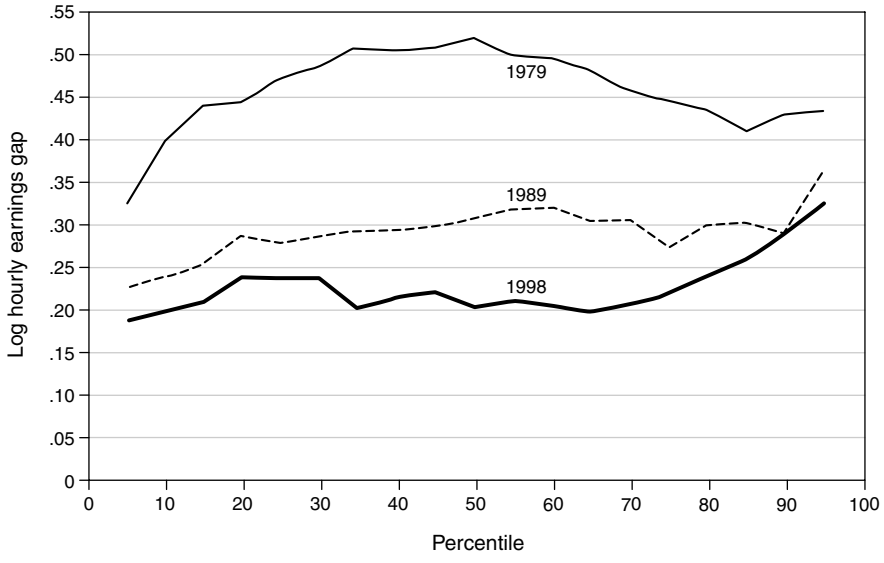


Figure 8.13
Ratio of hourly wages of women to those of men by decile in the United States.

Source: Blau and Kahn (2006, figure 1).

Figure 8.14
The figure graphs time series of (a) the log of the ratio of the wage of the median working woman to that of the median working man (left scale, black line), and (b) the log of the ratio of the wage of a man at the 90th percentile of the male wage distribution to that of a man at the 10th percentile (right scale, gray line). Data from the Current Population Survey for white persons aged 25–54, without trimming of outliers.

Source: Mulligan and Rubinstein (2008, figure 1).

the types of jobs held, as well as the trend in unionization, which was more favorable to women than to men over this period. Nonetheless, the price effects observed have contributed to an *increase* in the wage gap between the genders in the 1980s: women, who always hold less skilled jobs, the kind to which relative returns have fallen, have been disadvantaged by this phenomenon. This was no longer the case in the 1990s. The decomposition of terms corresponding to unobserved characteristics comes to the same type of conclusion. Unobserved price effects have contributed to increasing the wage gap, while changes in unobserved characteristics have pushed the wage gap the other way. Hence the combined trend in characteristics, observed and unobserved, of women, represented in the penultimate line of table 8.8, turns out to be favorable to women in both periods. In contrast, the total contribution of price effects, which includes the effect of discrimination, reported in the last line, has run counter to the reduction of wage gaps in the 1980s but not in the 1990s. Finally, the estimates of Blau and Kahn (2006) also show that the unobserved part of the wage gap between men and women declined substantially in the 1980s—the decline is given by the sum of terms (3) and (4)

TABLE 8.8

Decomposition of the movement in the gender wage gap, 1979–1989, and 1989–1998. Annual changes times 10. The wage gap D_t is defined as the difference between the mean of log of wages of men and women, that is, $D_t = \ln(w_{mt}) - \ln(w_{ft})$. Standard errors in parentheses.

Dependent: log hourly wage	$(D_{89}-D_{79})$	$(D_{98}-D_{89})$	$(D_{98}-D_{89})-(D_{81}-D_{79})$
Change in differential $(D_1 - D_0)$	−.164	−.075	.089
$(\Delta\bar{\mathbf{x}}_1 - \Delta\bar{\mathbf{x}}_0)\,\hat{\beta}_1$ (1)	−.092 (.005)	−.068 (.005)	.025 (.007)
Education	−.001	−.029	−.028
Experience	−.046	−.017	.029
Occupation	−.029	−.019	.010
Collective bargaining	−.022	−.007	.014
Industry	.003	.005	.002
$\Delta\bar{\mathbf{x}}_1\left(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0\right)$ (2)	.038 (.023)	−.006 (.024)	−.044 (.033)
Education	.004	.002	−.002
Experience	.022	−.027	−.048
Occupation	−.018	.054	.072
Collective bargaining	.007	.000	−.008
Industry	.022	−.036	−.058
$\left(\Delta\hat{\theta}_1 - \Delta\hat{\theta}_0\right)\hat{\sigma}_1$ (3)	−.128 (.014)	−.007 (.0016)	.121 (.021)
$\Delta\hat{\theta}_1\,(\hat{\sigma}_1 - \hat{\sigma}_0)$ (4)	.019 (.004)	.007 (.002)	−.012 (.005)
Sum gender characteristics ((1) + (3))	−.219	−.074	.145
Sum wage structure ((2) + (4))	.057	.001	−.056

Source: Blau and Kahn (2006, table 2b).

in table 8.8—which points to the conclusion either that discrimination against women has declined or that their unobserved characteristics have moved closer to those of men. But this trend was no longer in play in the 1990s.

Overall these results indicate that the primary reason for the slowdown of the convergence of wages in the 1990s between women and men stems from a gap in unmeasured characteristics which has narrowed at a slower pace in the 1990s, a slower reduction in discrimination, or supply and demand conditions for women having changed more favorably in the 1980s than in the 1990s. Removing the occupation and industry-sector variables, which could be endogenous to discrimination, from the controls does not qualitatively change this conclusion. Blau and Kahn (2006) also control for selection bias, which can change over time and differently across gender and which could explain part of this result. To do so, they impute wages above or below the median wage to nonparticipants based on observable characteristics and then compute the log of hourly wage differentials between the median of the hourly wages of each group. This procedure brings about a greater increase in the number of women in the sample compared to the increase in the number of men. It also tends to reduce the relative wage offers of women. The results of Blau and Kahn (2006) show that while accounting for sample selection does not qualitatively change the previous results, selection did tend to overstate women's convergence, notably in the 1980s, because the female labor force growth in that period was positively selected. The selection bias was smaller in the 1990s due to a slower growth of the female labor force and a negatively selected growth this time (more relatively low-skilled women entering the market). Overall, selection could explain 25% of the slowdown in the narrowing of the unexplained gender pay gap (due to unobservable characteristics or prices).

Mulligan and Rubinstein (2008) confirm that selection effects can account for a large part, in fact for most, of the apparent changes of wages across gender among whites aged 25 to 54. They conducted a decomposition based on the Current Population Survey cross sections over the period 1975–1999, controlling for selection bias based on Heckman's (1979) two-step procedure. In practice, they regress equations (8.10) and (8.9) for men and women but add an inverse Mills ratio in the female wage equation (not in the male one since they have no good exclusion restriction for men). Their control variables include educational attainment dummies, marital status, a potential work experience term interacted with education dummies, and region. The probit model used to estimate the inverse Mills ratio includes the same variables plus the number of children aged 0–6 interacted with marital status. Table 8.9 shows that after controlling for selection, the adjusted wage gap has not decreased in the 1970s compared to the 1990s. One possible explanation of this phenomenon is that the growing wage inequality within genders observed since the 1980s could indicate a shift in the demand for higher skills. In response, women with less human capital may have dropped out of the workforce, and those with more human capital may have entered. Women, especially the more able ones, may also have increased their human capital investment. This is evidenced by the observed increase in skill proxies—such as schooling—of the females participating in the labor market relative to the female population as a whole. This led to an increase in women's measured wages conditional on their observed characteristics because skills are imperfectly proxied by education. In sum, since the 1970s wage inequality drove changes in the composition of the female workforce, which appeared to speed up the convergence of female and male wages.

TABLE 8.9
Correcting the gender wage gap using the Heckman two-step method, 1975–1979 and 1995–1999.

Period	OLS	Heckman two-step	Bias
	(1)	(2)	(1)–(2)
1975–1979 (D_0)	−0.414 (0.003)	−0.337 (0.014)	−0.077 (0.015)
1995–1999 (D_1)	−0.254 (0.003)	−0.339 (0.014)	0.085 (0.015)
Change in differential ($D_1 - D_0$)	0.160 (0.005)	−0.002 (0.020)	0.162 (0.021)

Note: The entries are female minus male log wages. The regressions control for demographics interacted with gender and use CPS wage sample of white persons aged 25–54. Standard errors in parentheses.

Source: Mulligan and Rubinstein (2008, table 1).

Selection effects are also important in explaining the gender difference in high-earnings occupations and competitive settings. Gobillon et al. (2012) show that based on French administrative data collected at the firm level in the private sector and for full-time executives aged 40–45, the gender difference in the probability of getting a job conditional on a number of observable characteristics increases progressively along the wage ladder from 9% to 50%. Females thus have significantly reduced access to high-paid jobs than to low-paid jobs. Differences in the survival rates in these jobs might also be part of the explanation. Using a large data set of executives in North American firms over the period 1991–2006, Gayle et al. (2012) find that controlling for executive rank and background, women earn higher compensation than men in these positions and are promoted more quickly conditional on survival as an executive. Female executives, however, have a higher exit rate than men and the probability of a female executive becoming CEO is less than half that of male executives at every age. Hence, the unadjusted gender pay gap and job-rank differences are primarily attributable to female executives exiting at higher rates than men in highly competitive environments.

4.3 SEXUAL ORIENTATION AND DISCRIMINATION

We have focused so far on differences in labor market outcomes across relatively easily identifiable demographic groups. Other individual characteristics, sometimes more difficult to track down in statistics, can also have detrimental effects on hiring probabilities and wages. Sexual orientation is one of them. Gay men tend to have lower wages than heterosexual men, and both gays and lesbians tend to have fewer opportunities than heterosexuals. Most studies related to this type of discrimination are based on field experiments since sexual orientation is usually not recorded in surveys. They typically find reductions in wage offers of about 5% to 15% and lower rates of callbacks to their applications.

The World Values Survey (WVS) reveals that the prejudice against homosexuals is significant in many countries. One question is especially revealing of distaste for homosexuals: "On this list are various groups of people. Could you please mention

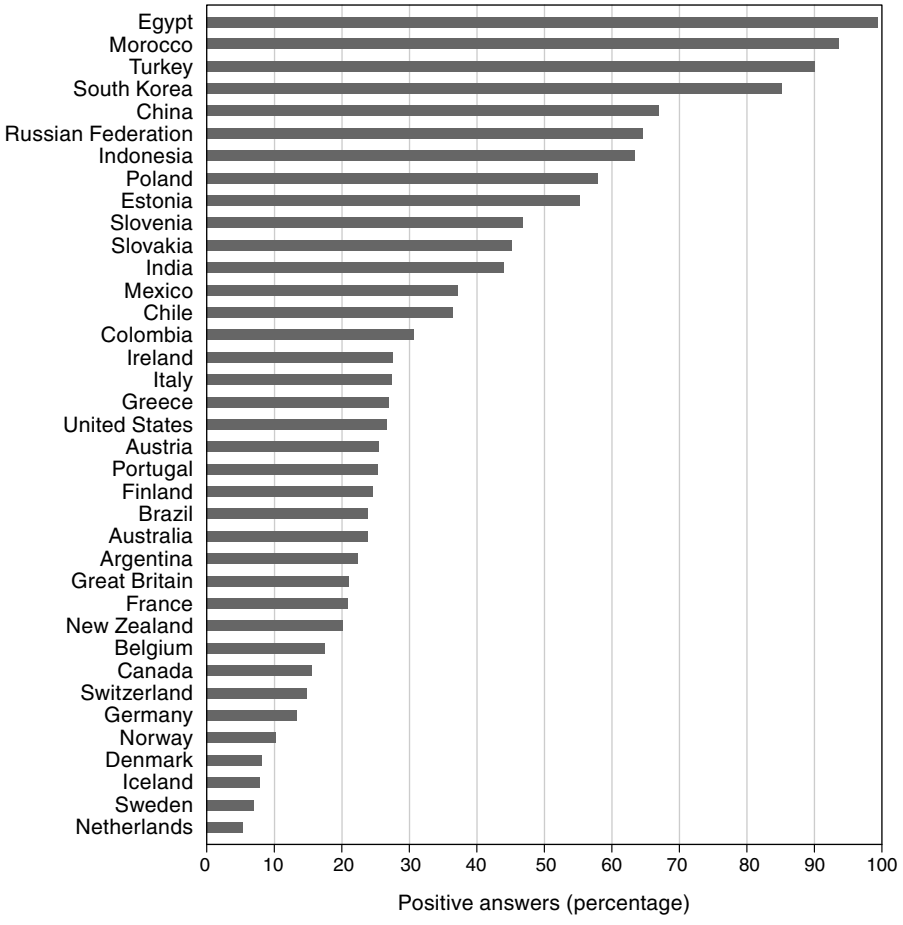


FIGURE 8.15
Percentage of population who would not like to have homosexuals as neighbors.

Source: World Value Surveys, waves 1995 and following.

any that you would not like to have as neighbors?" Figure 8.15 shows that about 48% of the respondents in the 90 countries where this question has been asked since 1995 mention homosexuals. Intolerance seems most severe in Turkey, the Middle East, and Asia (80%–90%), while tolerance is high in Northern Europe and Germany (10%–15%). In the United States about 27% answered that they would not like to have homosexuals as neighbors.

Drydakis (2012) sent resumes to potential employers (advertising on websites) in the Republic of Cyprus in which the "applicant's" sexual orientation was disclosed at random through a reference in his/her resume to work as a volunteer for a gay association. The applicants were identical in all employment-relevant characteristics except sexual orientation and all had Greek Cypriot names. The occupations covered a

wide spectrum of work environments: office jobs, industry jobs, café and restaurant services, and shop sales. The results highlight clear discrimination in access to jobs: while applications with the control signal elicited a callback in 40% of the cases, less than 14% of the gay-labeled applications did so. Results are similar for lesbians. Following the interviews, wage offers were lower by 4% for gays and 3% for lesbians.

Tilcsik (2011) obtains similar findings for the United States for gay men. He sent pairs of fictitious resumes in response to 1,769 job postings (administrative assistant, analyst, customer service representative, manager, sales representative) in seven states. One resume in each pair was randomly assigned experience in a gay campus organization and the other resume was assigned activities in a control organization. Table 8.10 shows the results in terms of callbacks. The first line shows that on average, nongay resumes get 1.6 times more callbacks than gay resumes. Interestingly, some traditionally liberal states such as New York, Pennsylvania, and California were significantly less likely to treat gay job applicants unfavorably, compared to Florida, Ohio, Nevada, and Texas. In general, gay resumes sent to employers in states, counties, or cities where there are laws that prohibit sexual orientation discrimination had more chances of getting callbacks than those sent to employers where there is not a similar prohibition.

Gay and lesbian workers also tend to sort themselves into tolerant occupations. Plug et al. (2014) use the Australian Twin Registers, which contain detailed information on a large sample of identical (monozygotic) and fraternal (dizygotic) twins. For the year 1992 it included a sex survey in which twins were asked about their sexual orientation, the sexual orientation of their twin sibling (which permits an estimate of lower and upper bounds on the sexual orientation effect linked to measurement errors in the report of sexual orientation; some people might lie), their opinion on sexual orientation

TABLE 8.10
Callback rates by sexual orientation.

		% Callback		
Sample	Number of ads	Not gay	Gay	Difference (P-value)
Total sample	1769	11.5	7.2	4.3(.000)
California	337	11.0	9.2	1.8(.443)
Nevada	131	12.2	6.1	6.1(.087)
New York	236	10.2	11.4	−1.2(.656)
Pennsylvania	201	12.9	9.4	3.5(.268)
Ohio	219	14.1	5.5	8.6(.002)
Florida	347	9.5	5.5	4.0(.044)
Texas	298	12.0	3.7	8.3(.000)
Employers subject to a city, county, or state law that prohibits sexual orientation discrimination:				
Yes	983	11.6	8.7	2.9(.037)
No	786	11.3	5.3	6.0(.000)

Source: Tilcsik (2011, table 5).

(reflecting distaste or, on the contrary, indifference[12]), and the type of occupation in which they were employed. The advantage of working with twins is that apart from their sexual orientation, which actually may differ, their family background and other important characteristics difficult to observe are more likely to be identical or closely similar. Of the 4,835 twins who responded to the sexual orientation question, 215 of them were gay, lesbian, or bisexual. Plug et al. include in their analysis twin fixed effects and therefore control for all observed and time-invariant unobserved characteristics that twins share which could drive occupational choices. The majority of the workforce surveyed turned out to be prejudiced against homosexuals: among straight twins only, about 75% are prejudiced. The observed fraction of prejudiced straight workers by occupation serves as a measure of intolerance by occupation. The most intolerant occupations can be found among plant operators and tradespersons including carpenters, motor mechanics, printing machinists, vehicle and building tradespersons, and gardeners.

Plug et al. then analyze the exposure to prejudice in occupations across workers who are identical twins. But obviously identical twins with different sexual orientations are not fully identical. The key identification assumption of prejudice-based segregation is then that those unobservable twin differences in productivity and taste factors are unrelated to observable twin differences in sexual orientation. Otherwise the estimated coefficient for sexual orientation would be spurious. The authors thus control for a number of measures of educational achievement and personality traits that could be correlated with sexual orientation and with skills and occupational choices. Their results indicate that, indeed, gay and lesbian workers choose to work in less prejudiced occupations: considering only identical twins, for which characteristics are the closest, gay and lesbian workers have on average about 6 percentage points fewer prejudiced colleagues, which corresponds to a 50% of a one-standard-deviation decrease in the fraction of intolerant colleagues compared with straight workers. Including fraternal twins in the sample does not change the results significantly. The authors show that this segregation cannot be explained by reverse causation (where workplace contact raises tolerance among straight workers).

As we have seen in the taste-based model, this type of segregation can in principle lead to lower earnings. A number of studies confirm this effect, notably for males. Plug and Berkhout (2004) find that in the Netherlands among highly educated young workers, gay males earn about 3% less than comparable heterosexuals. Similarly, Laurent and Mihoubi (2011) find that in France the magnitude of the unexplained wage gap for gay males is about 6% and that the wage gap is higher for skilled workers than for the unskilled. Neither study identifies a wage gap for lesbian workers.

Overall, this literature identifies significant effects of sexual orientation on employment and wages. International surveys of opinion regarding gays and lesbians, as well as studies of their occupational choices, suggest that these differences are taste-based, arising out of the preferences of employers or other employees. Yet we do not dispose of elements that would permit us to exclude with certainty a statistical origin for this type

[12]Here, prejudice against homosexuals is measured through agreement or disagreement with statements such as "Homosexuality is a social corruption and can cause the downfall of civilization" and "Homosexuality is merely a different kind of sexuality and is not immoral." Several statements of each kind were submitted to approval in the questionnaire.

of discrimination: employers might think, for example, that the lifestyle of homosexuals tends to make them less assiduous at work or more inclined to change jobs.

4.4 The Premium for Beauty

We round off this review of empirical studies with a form of discrimination that is perhaps one of the most widespread and natural, to the point that few people remark on it. One of the obvious ways personal characteristics could influence labor market outcomes is how people look. If the popular saying that "first impressions count" is true, then physical attributes could have an impact on hiring decisions and wage offers, even if they are independent of a person's productivity-related features. In their seminal work, Hamermesh and Biddle (1994) do reveal the existence of a beauty premium on earnings. Of course, if there were no shared agreement on what constitutes beauty, then looking for a beauty effect on labor market outcomes would make no sense. Though standards of beauty may vary across cultures, there is evidence of persistent standards within cultures, based for instance on the answers of respondents of various ages who are asked to rank the appearance of people depicted in photographs. Hamermesh and Biddle used several surveys carried out in Canada and the United States where the interviewers were pointedly asked to evaluate the "physical appearance" of respondents on a 5-point scale ranging from strikingly handsome or beautiful to homely. They found that plain people earn less than average-looking people, who earn less than the good-looking, even after controlling for other characteristics. The plainness penalty is 5% to 10%, slightly greater than the beauty premium (5% more). Effects for men are at least as great as for women. Beauty also impacts positively women's labor force participation rates. Size too has a positive impact on earnings. Overall the impact of individuals' looks seems largely independent of occupation, which is interpreted by the authors as evidence of the existence of pure employer discrimination (if beauty were linked to productivity, it should matter only in occupations where attractiveness is economically important). Actually this interpretation is debatable, if beauty alters characteristics that play on productivity such as confidence or personal network.

Indeed, factors other than discrimination could be in play. For instance, attractiveness could be correlated with unobservable productive attributes such as health, education, and other types of human capital that could explain wage differences. Beauty could also increase confidence and improve social skills. If these characteristics are improperly measured or omitted from regression analysis, the impact of beauty on wages may be overestimated.

Beauty can improve productivity through a variety of channels. Biddle and Hamermesh (1998) studied the legal sector, where contacts between lawyers and clients are frequent and important. They used longitudinal data on graduates from a law school and measured beauty by rating matriculation photographs. Better-looking attorneys who graduated in the 1970s earned more than others after five years of practice, an effect that grew with experience. The premium existed in all areas of expertise, including among those self-employed. One plausible explanation is taste-based discrimination by clients. A higher demand on the part of clients for handsome or beautiful lawyers could drive up their earnings by bringing them both more cases and higher fees. Arunachalam and Shah (2012) estimate the earnings premium for beauty in another profession where beauty can have an important impact on earnings: prostitution. They use two representative surveys of female sex workers in Mexico and Ecuador at the beginning of the

2000s. They find a beauty premium of about 20%, slightly higher than that estimated for women in general. The beauty premium stems from both the ability to charge a higher price for each transaction and the ability to work more over a given day and increase the number of transactions.

Mobius and Rosenblat (2006) decomposed the beauty premium that comes into play during the wage negotiation process between employers and workers in an experiment conducted in 2002 and 2003. "Workers" stationed at computers have to perform true skill tests (mazes) that are unaffected by physical attractiveness, while "employers" estimate the productivity of workers, including through interviews, and set wages accordingly. Before taking the test, participants are asked for an estimate of how many mazes for each level of difficulty they expect to complete. This information provides a measure of worker confidence. The beauty of workers was assessed independently. It turns out that beauty has no impact on actual performance during tests but that physically attractive participants are substantially more confident: a one-standard-deviation increase in beauty raises confidence by about 13%. Division into several groups allowed the researchers to control for the impact of beauty on wages: in one group, employers only have access to the resume (without photo), in another group they see the photo, in a third group they can talk with workers, in a fourth group they can both talk with and see photos of the workers, and in a last group they can conduct full face-to-face interviews. Regressing the wage on beauty, actual performance, and other controls for each group separately, the authors find a wage premium for beauty that is sizable: about 12% to 13%, and even 17% in the face-to-face interviews for a one-standard-deviation increase in beauty, but no beauty effect in the baseline group where employers only have access to resumes. The authors then decompose the beauty premium. They first add the measure of confidence to the wage equation and find that a 1% increase in confidence increases wages by about 0.2%, while reducing the size of the coefficient of the beauty measure, but only in the groups where oral communication is possible. Interestingly, there is a remaining beauty premium in the group where only oral communication is allowed even after controlling for confidence, suggesting that physically attractive participants have oral skills (such as fluency and social ease) that can improve their wage independently of confidence. Since beauty increases confidence, the increase in wages of a one-standard-deviation increase in beauty transmitted through the confidence channel is 13 times 0.2%, or 2.6%. Then the authors pool the data from all groups and make a combined assessment of the impact of confidence on wages, as well as that of beauty, which can work through the channels of oral or visual communication depending on the type of treatment of participants. They find that about 15% to 20% of the total beauty premium stems from confidence and that about 40% stems from each of the interaction channels, visual and oral.

Discrimination based on beauty is probably also more difficult to detect as compared to race or gender discrimination because people are less aware of its prevalence and discriminators might be less subject to social disapproval. Belot et al. (2012) revealed that discrimination can arise in fully public settings, actually in front of millions of people. They study the outcomes of a television game show, to be precise 69 episodes of the game show "Does (S)he Share or Not?" broadcast in the Netherlands in 2002, with 345 contestants in total. In this game the performance of contestants is clear-cut and the stakes are high. The game takes place over three rounds, in which players accumulate "earnings" by answering quiz questions. Their earnings depend on the accuracy of their answers, on how quickly they press the buzzer, and on their "investment

decisions." Earnings therefore depend on ability as well as a player's confidence (stakes are pretty high, with a median of €1,683). At the end of each round, the lead player—the one with the highest earnings—decides which one of the remaining players to eliminate. The authors find that unattractive players are more often eliminated than attractive ones. Players can only win positive earnings by making it to the final stage of the game show. But only 27% of the least attractive players make it to the final round as against 49% of the most attractive ones, and this difference is not linked to performance. Actually, the least attractive players are eliminated by the lead players even when they have a higher score than others.

The distaste of employers for some physical characteristics can also be measured directly, but it takes more than an explicit questionnaire to identify discriminatory attitudes. Agerstrom and Rooth (2011) focus on obesity and combine data from a correspondence approach (testing for employer discrimination against obese job applicants) with "implicit association test" (IAT) data derived from the employers who received the correspondence resumes. In a first stage, all resumes were matched to job vacancies based on credentials and were equivalent except for the applicant's weight, which was indicated by attaching a facial photograph of either an obese or a normal-weight person. Discriminatory behavior was then quantified by the extent to which the hiring managers invited normal-weight versus obese applicants to a job interview. In a second stage, several months after these data were obtained, the hiring managers were asked to pair photographs of obese and normal-weight individuals with words denoting performance (e.g., effective, hardworking, ambitious vs. ineffective, incompetent, slow, etc.). Agerstrom and Rooth find that employers holding implicit prejudicial attitudes against the obese were also those less likely to call back obese applicants for an interview, suggesting that automatic stereotypes do have an impact in actual hiring situations. Explicit (self-reported) hiring preferences based on weight were also collected through a questionnaire but, contrary to implicit measures, they failed to predict hiring decisions.

There is now clear evidence that beauty has an impact on labor market outcomes. But the task of pinning down a discrimination phenomenon is rendered difficult by the fact that beauty may influence characteristics that are hard to observe, like self-confidence and communicative capacity, and that may exert effects on productivity and wages. Besides, the origin of this discrimination is also hard to pin down: certain employers may find it more agreeable to work with good-looking personnel (taste discrimination) while at the same time believing, rightly or wrongly, that good-looking personnel will on average be more productive in their line of business (statistical discrimination).

5 HOW TO REDUCE INEQUALITY AMONG DEMOGRAPHIC GROUPS

Analysis of the performance differentials of individuals belonging to different demographic groups indicates that a large proportion of the variation that obtains is generally explainable by differences in observable characteristics that might tend to influence their productivity. This suggests that policies to combat discrimination on the labor market can reduce some but not all of the inequalities in performance between demographic

groups. In this section we begin by studying the consequences of affirmative action, which lays down legal rules to offset the disadvantages experienced by groups that are faring worse than others. We then review the importance of such premarket factors as initial cognitive competence and present policies that might modify them. Last, we see that other lines of research on the boundary between economics and psychology have examined the role of psychological attributes such as attitudes towards risk and competition, or the role played by social norms that dictate what women ought or ought not to be engaged in, as alternative or complementary explanations of the different trajectories of men and women on the labor market. This research shows that in order to counteract inequalities among demographic groups, it is indispensable to account for premarket factors that are influenced by education, psychological attributes, and social norms.

5.1 Affirmative Action

Affirmative action imposes legal constraints (e.g., hiring quotas or minimum wage levels) benefiting certain minority groups in order to offset their unfavorable situation on the labor market. It will be helpful to start by looking at the theoretical consequences of this kind of intervention before going on to examine what we can learn from empirical work in this area.

5.1.1 Theoretical Considerations

We saw in section 2.2 that observation of poor performance by a group in the past is capable of influencing present beliefs and exerting a disincentive effect on the behavior of members of the group in question. The dynamic of self-fulfilling prophecies can engender persistent inequalities (Loury, 1998) that have to be combated through suitable policies. Affirmative action forces employers to treat persons belonging to disadvantaged groups in the same way they treat those belonging to more favored ones. The imposition of quotas privileging the hiring of workers belonging to groups that are a priori disadvantaged by the functioning of the labor market forms part of the toolkit of affirmative action measures. Coate and Loury (1993) have pointed out that hiring quotas risk having a disincentive impact on the investments in education of persons belonging to groups that benefit from affirmative action and thus turn out to be inefficient in the end. We can easily grasp this using the two-stage game set forth above (section 2.2) and adding the assumption that the public authorities oblige employers to hire a minimum proportion π_g of workers at a minimum wage of $w^+ = h^+\pi_g/[\pi_g + p(1-\pi_g)]$. This wage corresponds to the equilibrium wage of efficient workers when such workers do in fact represent a part π_g of their group (if the hiring test is reliable enough for the inequality $p \leq (h^+ - 1)/h^+$ to be satisfied). In these circumstances, the government can implement affirmative action for the purpose of combating the perverse effects of statistical discrimination that lock the labor market into a suboptimal situation. Let us assume that the government is aware of the model developed just now and is striving to reach the high equilibrium of figure 8.11 by imposing $\pi_g = 1$. Employers are thus obliged to hire all workers at wage h^+. The return to education becomes systematically negative, since an educated worker obtains $h^+ - 1$ while an uneducated one obtains h^+. The existence of the quota discourages education and leads ultimately to a highly inefficient situation in which firms make negative profits by being forced to hire workers who have no

incentive to improve their productivity. These considerations suggest that affirmative action can have detrimental consequences that lead to efficiency losses.

To combat the potential negative effects of statistical discrimination, Coate and Loury (1993) recommend instead the use of subsidies targeted so as to raise the returns to education. In figure 8.11, a subsidy equal to the cost of educational effort, financed by a lump sum tax, gives agents an incentive to get educated and leads to coordination at the good equilibrium by shifting the curve u^+ upward. Neumark (1999) and Altonji and Blank (1999) also suggest that giving employers incentive to improve the procedures they use to evaluate job applicants would constitute an effective means of combating statistical discrimination.

5.1.2 Empirical Results

Affirmative action has been in place in the United States since the beginning of the 1960s. It has led to a large number of decisions in the wake of *Kennedy Executive Order 10925* in 1961, which requires that firms contracting with the government "take affirmative action to ensure that applicants are employed and employees are treated during employment without regard to their race, creed, color, or national origin." Following this decision, the policy of affirmative action underwent a number of developments. In 1965 *Johnson Executive Order 11246* reiterated *Kennedy Executive Order 10925*. In 1967 *Johnson Executive Order 11375* stated that *Executive Order 11246* applied to women as well. In 1968 *Department of Labor Regulations Governing Executive Orders 11266* and *11375* made it mandatory for firms contracting with the federal government, and which had more than 50 employees or a contract worth more than $50,000, to establish the degree to which women and minorities were underrepresented in their workforce, then set out corrective goals and a timetable for achieving them. In 1979, in *United Steelworkers of America v. Weber*, a case concerning a training program in which 50% of the places were reserved for blacks, the Supreme Court decided that the Civil Rights Act of 1964 "does not prohibit such race-conscious affirmative action plans." The judgment states clearly that such plans, which aim to remedy entrenched phenomena of segregation, are legitimate.

This series of decisions shows that demographic groups which have lower labor market performance obtain the benefit of programs which give firms contracting with the government an incentive to hire them. Empirical studies dedicated to the consequences of affirmative action have sought, in the first place, to detect its impact on wages. The evidence is that it very probably favored blacks and women during the 1960s and 1970s, although to assess the extent of this effect with precision is a very tricky business (see Donohue and Heckman, 1991). Other studies focused on the distorting effects of these programs, attempting to discover whether they drove firms to recruit underperforming workers. In their overview of the literature, Holzer and Neumark (2000, 2001) emphasize that there is nothing to point to the conclusion that women who benefited from affirmative action had lower levels of education or experience than those of men for comparable types of jobs. Their performance in the labor market was likewise comparable. On the other hand, the levels of education and experience of ethnic minorities who benefited from affirmative action programs are frequently lower than those of their white colleagues. Their performance in the labor market is, however, very close to that of whites. Holzer and Neumark (2000) show that this result arises from the fact that employers practicing affirmative action select members of ethnic minorities with

greater rigor at the hiring stage. Overall, Holzer and Neumark (2000) estimate that losses in productivity owing to affirmative action appear to be limited.

Affirmative action does not necessarily favor its beneficiaries. For instance, Sander (2004) shows that affirmative action in the area of tertiary education may actually harm some of its beneficiaries. Using a national cohort of 27,000 law school students in the United States, Sander finds that half of black first-year students fall into the lowest decile of the overall grade distribution, while lower first-year grades are associated with higher rates of dropping out and lower chances of becoming a lawyer. A related consequence of affirmative action stressed by Arcidiacono, Aucejo, and Spenner (2012) is that it induces changes in course selection that result from black and white students having very different persistence rates in the natural sciences, engineering, and economics. While, conditional on sex, black students have stronger initial preferences than whites for majoring in the natural sciences, engineering, or economics, they are significantly less likely to choose one of these majors for their final major. Arcidiacono, Aucejo, and Spenner (2012) show that these differences in persistence rates are fully explained by differences in academic background. Courses in the natural sciences, engineering, and economics are rated more difficult, are associated with longer study times, and have harsher grade distributions than those in the humanities and social sciences. The differences in difficulty levels across course types then works to dissuade individuals with relatively worse academic backgrounds from persisting in natural science, engineering, or economics majors. From this perspective, affirmative action may be working to increase the number of nonscience majors at top schools at the expense of science majors at less-selective schools and therefore be contributing to reduce the share of minority representation in the sciences.

But, in the absence of affirmative action, would black students succeed in higher proportions because they would enroll in greater numbers at less-selective schools where exams are less difficult to pass, instead of enrolling at top schools? This question cannot be easily answered. Selection based on race or ethnic origin in American universities yields de facto significantly different selection rates across groups. Espenshade and Chung (2005) simulated that for the 1997 entering class in elite universities, eliminating affirmative action would reduce acceptance rates for African American and Hispanic applicants by as much as one half to two thirds. This type of policy was secured but also constrained by recent U.S. Supreme Court decisions. In 2003 the Court ruled that race could be used as one of several factors (with no quota) in professional school admissions without necessarily violating the Constitution (*Grutter v. Bollinger*), but it also ruled that an undergraduate admissions system that granted extra "points" to minorities based on race was unconstitutional because it is too mechanical and not sufficiently reflective of actual individual merits (*Gratz v. Bollinger*). Fryer and Loury (2005) present a review of these debates.

5.2 The Importance of Premarket Factors

It is by now well established that capacities, whether cognitive or noncognitive, acquired well before entry onto the labor market explain the largest part of the inequalities among demographic groups as defined by race, ethnicity, gender, and geographic origin. To significantly reduce these inequalities, it is thus imperative to act on these premarket factors.

5.2.1 EDUCATION

Fryer (2011) studied the cognitive performance of young children in the United States according to their ethnic origin before they started school, on the basis of individual data. While at the age of 8 to 9 months all children perform at the same level whatever their origin, the distribution of performance by whites shifts to the right from the age of two years, signifying performances better on average than those of blacks and Hispanics. By four years the gap widens, especially where mathematics are concerned, and continues to do so thereafter.

These differences may flow from different individual characteristics. Controlling for demographic factors such as sex, age (expressed in months) at the time of testing, the region of residence, and also for environmental factors like socioeconomic status, parents' educational level and age, family structure, and health-related data like birth weight or premature birth, a measure of the effect of racial or ethnic origin is obtained by regressing an equation of type (8.7), where the dependent variable is no longer the log of wages but the results of cognitive tests. Table 8.11 shows that if the introduction of controls cancels any differences in performance before the age of 1 year among ethnic groups, it reduces them by half or a little more but does not erase them entirely from the age of 4 years onward (or even from the age of 2 years, according to another test). Complementary data on performance in school suggest that the performance gap persists as the years pass, or even widens slowly until high school, in all areas (mathematics, reading, etc.). The level of competence of mothers (measured by the AFQT) plays an important part in the explanation of these results. The same holds good for the quality of schools and teachers, where that can be controlled.

These results suggest that family environment plays an important part in the development of infant skills from the earliest age. They also show that even with numerous controls, it is not easy to fully specify the factors that contribute to the formation of the premarket racial gap. The quality of education, not taken directly into account here, no doubt plays an essential part as well. A certain number of initiatives have targeted infant skills in the United States, such as the Perry Preschool Program, which offers support to children from disadvantaged backgrounds from the age of 3 years and

TABLE 8.11
Black–Hispanic–white mental test score gap.

Dependent:	Mental function composite score					
	Less than 1 year		4 years		7 years	
	(1)	(2)	(3)	(4)	(7)	(8)
Black	−.096 (.012)	.024 (.017)	−.785 (.011)	−.296 (.016)	−.854 (.010)	−.348 (.016)
Hispanic	.183 (.034)	−.039 (.040)	−.895 (.032)	−.542 (.039)	−.846 (.031)	−.545 (.038)
Controls	No	Yes	No	Yes	No	Yes
Number of observations	31116	31116	31116	31116	31116	31116

Note: OLS regressions, standard deviations in parentheses, based on the Collaborative Perinatal Project (CPP) data. Reference group: whites. Scores are normalized to have a mean of zero and a standard deviation of one.

Source: Fryer (2011, table 5).

aims to develop their cognitive and noncognitive capacities (see chapter 4). Similarly, structures like the federal Head Start program (Puma et al., 2010), which has already been extended to 900,000 children between 3 and 5 years old and living in families below the poverty threshold, or the Nurse Family Partnership (Olds et al., 2002), which provides prenatal and postnatal home visits by nurses until the age of 2, have shown a certain degree of effectiveness (see Fryer, 2011, for a survey of this type of program). Subsequent interventions during elementary school tend to be less effective on average than these earlier ones. There are several exceptions to that assessment, such as Success For All, where the main goal is to remedy reading deficiencies (Borman et al., 2007), or Mastery Learning, a group-based approach that adjusts the rhythm of learning until a certain objective is reached before moving on to the next (Guskey and Pigott, 1988).

While improvements in schooling are essential for narrowing the premarket racial gap, it is also the case that the environment of children raised in disadvantaged circumstances blighted by poverty contributes to this gap. Hence combating poverty and developing activities that allow children to make constructive use of their time outside the classroom are equally essential complements to courses of action pursued strictly in the school setting. Still, the Harlem Children's Zone program, which exemplifies this logic, is not entirely conclusive on this point. This program combines innovative charter schools (schools that receive public money while benefiting from greater flexibility than traditional public schools) with a network of community services (including youth programs and housing and health programs) in an effort to make the social environment outside the classroom positive and supportive. The charter schools select the pupils from their catchment area by lottery, which makes it possible to assess the impact of the school independently of the impact of the community services. Dobbie and Fryer (2009) find that the charter schools are effective at increasing the achievement of the poorest minority children, but they find no clear correlation between achievement at school and participation in community programs. Reducing poverty in the community is probably not enough to improve academic achievement.

Recent research thus suggests the need to intervene very early in childhood, especially through intensive and innovative programs of education that bring in the parents, if we are to improve the chances of reducing the differences that persist between racial and ethnic minorities and whites on the labor market (see Fryer, 2011; Heckman, 2011; and Lang and Lehmann, 2012, for recent reviews of findings on racial discrimination and premarket factors).

5.2.2 The Role of Psychological Attributes and Social Norms

Psychological attributes, for instance attitudes toward risk and competition or prevailing social norms, can influence labor market performance across demographic groups. If we consider the situation of women for instance, psychological specificities can help to explain why women select themselves or are selected for specific types of jobs, as well as their career paths. These specificities can also shed light on some of the "unexplained" factors that may underlie wage differences without necessarily arising out of naked discrimination (see Bertrand, 2011, for a review of the new perspectives on gender). Many of these results are based on experimental methods and are situated at the frontier between psychology and economics.

Risk and Competitions Preferences

Risk preference might be an important factor determining wages, since risk-averse individuals tend to prefer jobs with more stable wages, which also tend to pay less on average. Using the German Socio-Economic Panel, which contains a question about willingness to take risks that has been shown to be a behaviorally valid measure of risk aversion, Bonin et al. (2007) show that individuals with low willingness to take risks are more likely to work in occupations with low earnings risk. Dohmen et al. (2011) show, using the same survey complemented with a lab experiment with real-stake lotteries, that gender is the primary factor for explaining attitudes toward risk: women tend to be less willing to assume risk in general and in various areas of activity, even after controlling for many observable characteristics.

Attitudes towards competition also vary across gender and impact earnings. For instance, in many high-earnings occupations the work is done in highly competitive settings—environments in which women tend to underperfom. Niederle and Vesterland (2007) asked participants in a laboratory experiment to solve a real task (of an arithmetical kind). While men and women perform equally in both a noncompetitive and a competitive setting, when given the choice 73% of the men select a tournament-style competitive setting, and only 35% of the women make this choice. This gender gap in tournament entry is not explained by performance or by factors such as the presence of risk or greater aversion to negative feedback, which play only a negligible role. Men choose the tournament primarily because they are more overconfident than women and have a preference for competition. Other research suggests that women perform better when competing among themselves than with men.

These attitudes toward risk and competition also have some correlates in social preferences and behaviors: women tend to be more altruistic than men. There is a body of evidence that women tend to be more favorable to more redistribution than men. For instance, Alesina and Giuliano (2010a) find, on the basis of values surveys, that even after controlling for a wide range of characteristics, women tend more often to favor pro-redistributive policies than men. There is also evidence that women perform better when negotiating for others than for themselves, while performance is unchanged for men (Bowles et al., 2005).

Why Do Attitudes Toward Risk and Competition Vary?

Booth and Nolen (2012) show, in a controlled experiment, that women may differ in their propensity to choose a risky outcome not because of innate preferences but rather because they are influenced by the pressure to conform to gender stereotypes. The authors asked students from eight publicly funded single-sex and coeducational schools in the United Kingdom to choose between a real-stakes lottery and a sure bet. They found, comparing students with similar performances, that girls in an all-girls group or attending a single-sex school are more likely than their counterparts in mixed schools to choose a real-stakes gamble. They are more likely to choose competition (tournament) schemes as compared to piece-rate settings as well. Other research contributions have also pointed to biological factors. In particular, much research has focused on the role of testosterone levels in attitudes to risk, which differ between genders but also within each gender. Individuals with higher blood testosterone levels have more positive attitudes towards competition and dominance (see Bertrand, 2011, for a detailed review). The possibility exists that prenatal exposure to higher levels of testosterone may lead to

greater willingness to take financial risk. Dreber and Hoffman (2007) proxy the level of prenatal testosterone exposure by the ratio of the length of the second digit (the index finger) to that of the fourth digit (the ring finger). This so-called 2D:4D ratio has in fact been shown to be negatively correlated with prenatal testosterone exposure. They find that a higher ratio predicts more risk aversion. Similarly, Coates et al. (2009) find that male traders in the City of London with low 2D:4D ratios experience both higher profitability and increased longevity in the financial markets.

Women's participation in the labor market and their occupational choices can also be driven by social norms, by what men and women think the role of women should be in society. Fortin (2005) used the World Values Survey to show that the employment status of women in 25 OECD countries over a 10-year period (1990–1999) is negatively correlated with the social representation of women as homemakers and men as breadwinners (using the prevalence of statements such as "scarce jobs should go to men first" or "being a housewife is fulfilling"). In the same vein, Algan and Cahuc (2007) estimate that family values explain a non-negligible part of low female employment rates in southern European countries. More generally, Alesina and Giuliano (2010b) use questions from the World Values Survey regarding the role of the family and the love and respect that children are expected to have for their parents in 81 countries. They find that with strong family ties, home production is higher, families larger, and the labor force participation of women lower. To assess causality, they look at the behavior of second-generation immigrants and estimate a significant influence of the strength of family ties on economic outcomes. The gender pay gap is also influenced by attitudes: for instance Fortin (2005) shows that when 1% more men than women think that "scarce jobs should go to men first," the pay gap increases by 0.5%. The declining prevalence of this attitude across cohorts and over time seems consistent with a decline in the role of discrimination. The slowdown in the closing of the gender gap since the mid-1990s in the United States can also be linked to a shift in attitudes toward women. Using data from the 1977–2006 General Social Survey, Fortin (2009) shows that the evolution of gender role attitudes over time coincides with the evolution of female participation. In the 1970s and 1980s more women disagreed with the notion that husbands should be the breadwinners and wives should be the homemakers; they became more egalitarian until the mid-1990s, at which point these trends reversed. Gender role attitudes are found to explain at least a third of the recent leveling-off in the participation of women in the labor force.

This line of research on premarket factors suggests that some of the observed wage and occupational differences between men and women might be due to deeply rooted social norms, as well as psychological and even biological determinants, that are difficult to counteract. These results do not downplay the importance of discrimination towards women, but they do suggest that certain key factors, the influence of which is felt very early in life, may leave their imprint on labor market performance later.

6 SUMMARY AND CONCLUSION

- Taste discrimination refers to the aversion of employers, but also sometimes of clients and other workers, for some groups of workers which may lead

to lower wages for them. Such discrimination cannot persist under perfect competition, as employers with no preference will drive employers with discriminatory preferences out of the market, offering all workers equal wages. Monopsonistic competition, as well as search frictions, might explain why this type of discrimination can persist over time.

- The term "statistical discrimination" applies to a situation in which individuals with identical abilities but membership in different demographic groups have divergent career paths because of the average productivity, real or imagined, of agents belonging to their group. In this case, the beliefs of employers concerning the average quality of a demographic group can become self-fulfilling prophecies and provoke the appearance and persistence of productivity differences between groups, to the detriment of the ones discriminated against in the first place. A quota policy, providing for the hiring of a given proportion of members of a certain group, may turn out to be ineffective if it discourages efforts to acquire education.
- The difficulty of evaluating discrimination lies mainly in the assessment of the weight of unobserved individual characteristics. If we use sufficient care in controlling for the characteristics of individuals and of jobs offered, the proportion of wage differences attributable to discrimination declines. The majority of studies, though, conclude that in the United States blacks, Hispanics, and women are the victims of significant wage discrimination. Sexual orientation and beauty are other examples of individual characteristics that can give rise to wage differences unexplained by productivity factors.
- The empirical literature has recently insisted on the importance of premarket factors, such as early education, social norms, and psychological determinants, as complements to discrimination in the analysis of labor market performances. Female attitudes towards competition or the acceptance of norms about the role of women in society, which are acquired very early in life, can determine choices of occupation yielding lower earnings in adulthood. Family backgrounds and conditions of education for blacks and ethnic minorities in early childhood can in turn explain lower skill levels, yielding lower earnings in adulthood.

7 RELATED TOPICS IN THE BOOK

- Chapter 3, section 2: Compensating wage differentials and the hedonic theory of wages
- Chapter 4, section 2: The theory of human capital
- Chapter 4, section 3: Education as a signaling device
- Chapter 5, section 4.1: Empirical facts about wage differentials
- Chapter 6, section 5: Social preferences
- Chapter 11, section 3: Migrations

8 FURTHER READINGS

Altonji, J., & Blank, R. (1999). Race and gender in the labor market. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 3C, chap. 48, pp. 3143–3259). Amsterdam: Elsevier Science.

Arrow, K. (1998). What has economics to say about racial discrimination? *Journal of Economic Perspectives*, *12*, 91–100.

Bertrand, M. (2011). New perspectives on gender. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 4B, pp. 1545–1592). Amsterdam: Elsevier Science.

Becker, G. (1957). *The economics of discrimination*. Chicago, IL: University of Chicago Press.

Fortin, N., Lemieux, T., & Firpo, S. (2011). Decomposition methods. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 4A, pp. 2–102). Amsterdam: Elsevier Science.

Fryer, R., Jr. (2011). Racial inequality in the 21st century: The declining significance of discrimination. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 4B, pp. 875–971). Amsterdam: Elsevier Science.

Lang, K., & Lehmann, J.-Y. (2012). Racial discrimination in the labor market: Theory and empirics. *Journal of Economic Literature*, *50*(4), 959–1006.

REFERENCES

Adida, C., Laitin, D., & Valfort, M. A. (2010). Identifying barriers to Muslim integration in France. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(52), 384–390.

Aeberhardt, R., Fougère, D., Pouget, J., & Rathelot, R. (2010). Wages and employment of French workers with African origin. *Journal of Population Economics*, *23*, 881–905.

Agerstrom, J., & Rooth, D.-O. (2011). The role of automatic obesity stereotypes in real hiring discrimination. *Journal of Applied Psychology*, *96*(4), 790–805.

Alesina, A., & Giuliano, P. (2010a). Preferences for redistribution. In J. Benhabib, M. Jackson, & A. Bisin (Eds.), *Handbook of social economics* (vol. 1A, pp. 93–131). Amsterdam: North-Holland.

Alesina, A., & Giuliano, P. (2010b). The power of the family (with Alberto Alesina). *Journal of Economic Growth*, *15*(2), 93–125.

Algan, Y., & Cahuc, P. (2007). The roots of low European employment: Family culture? In J. Frankel & C. Pissarides (Eds.), *NBER International Seminar on Macroeconomics 2005* (pp. 65–109). Cambridge, MA: MIT Press.

Altonji, J., & Blank, R. (1999). Race and gender in the labor market. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 3C, chap. 48, pp. 3143–3259). Amsterdam: Elsevier Science.

Angrist, J., & Krueger, D. (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, *106*, 976–1014.

Antonovics, K., Arcidiacono, P., & Walsh, R. (2005). Games and discrimination: Lessons from "The Weakest Link." *Journal of Human Resources*, *40*, 918–947.

Arcidiacono, P., Aucejo, E., & Spenner, K. (2012). What happens after enrollment? An analysis of the time path of racial differences in GPA and major choice. *IZA Journal of Labor Economics*, *1*(5).

Arcidiacono, P., Beauchamp, A., Hull, M., & Sanders, S. (2012). Isolating mechanisms for the racial divide in education and the labor market: Evidence from interracial families. Mimeo, Boston College.

Arrow, K. (1973). The theory of discrimination. In O. Ashenfelter & A. Rees (Eds.), *Discrimination in labor markets* (pp. 3–33). Princeton, NJ: Princeton University Press.

Arrow, K. (1998). What has economics to say about racial discrimination? *Journal of Economic Perspectives*, *12*, 91–100.

Arunachalam, R., & Shah, M. (2012). The prostitute's allure: Examining returns to beauty, productivity and discrimination. *The B.E. Journal of Economic Analysis & Policy*, *12*(1), 1–25.

Banerjee, A., Bertrand, M., & Mullainathan, S. (2009). Labor market discrimination in Delhi: Evidence from a field experiment. *Journal of Comparative Economics*, *37*, 14–27.

Barth, E., & Dale-Olsen, H. (2009). Monopsonistic discrimination and the gender-wage gap. *Labour Economics*, *16*(5), 589–597.

Becker, G. (1957). *The economics of discrimination*. Chicago, IL: University of Chicago Press.

Belot, M., Bhaskar, V., & van de Ven, J. (2012). Beauty and the sources of discrimination. *Journal of Human Resources*, *47*(3), 851–872.

Bertrand, M. (2011). New perspectives on gender. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 4B, pp. 1545–1592). Amsterdam: Elsevier Science.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, *94*(4), 991–1013.

Biddle, J., & Hamermesh, D. (1998). Beauty, productivity and discrimination: Lawyers' looks and lucre. *Journal of Labor Economics*, *16*, 172–201.

Bjerk, D. (2007). The differing nature of black-white wage inequality across occupational sectors. *Journal of Human Resources*, *42*(2), 398–434.

Black, D. (1995). Discrimination in an equilibrium search model. *Journal of Labor Economics*, *13*, 309–334.

Blau, F., & Kahn, L. (1996). Wage structure and gender differentials: An international comparison. *Economica*, *63*, S29–S62.

Blau, F., & Kahn, L. (1997). Swimming upstream: Trends in the gender wage differential in the 1980s. *Journal of Labor Economics*, *15*, 1–42.

Blau, F., & Kahn, L. (2003). Understanding international differences in the gender pay gap. *Journal of Labor Economics*, *21*(1), 106–144.

Blau, F., & Kahn, L. (2006). The US gender pay gap in the 1990s: Slowing convergence. *Industrial and Labor Relations Review*, *60*(1), 45–66.

Blinder, A. (1974). *Toward an economic theory of income distribution*. Cambridge, MA: MIT Press.

Bonin, H., Dohmen, T., Falk, A., Huffman, D., & Sunde, U. (2007). Cross-sectional earnings risk and occupational sorting: The role of risk attitudes. *Labour Economics*, *14*(6), 926–937.

Booth, A., & Nolen, P. (2012). Gender differences in risk behaviour: Does nurture matter? *Economic Journal*, *122*, F56–F78.

Borman, G., Slavin, R., Cheung, A., Chamberlain, A., Nancy, A., & Chambers, B. (2007). Final reading outcomes of the national randomized field trial of Success For All. *American Educational Research Journal*, *44*(3), 701–731.

Bowles, H., Babcock, L., & McGinn, K. (2005). Constraints and triggers: Situational mechanics of gender in negotiation. *Journal of Personality and Social Psychology*, *89*, 951–965.

Bowlus, A., & Eckstein, Z. (2002). Discrimination and skill differences in an equilibrium search model. *International Economic Review*, *43*(4), 1309–1345.

Brown, C. (1984). Black-white earnings ratios since the Civil Rights Act of 1964: The importance of labor market dropouts. *Quarterly Journal of Economics*, *99*, 31–44.

Butler, R., & Heckman, J. (1977). The government's impact on the labor market status of black Americans: A critical review. In B. Farrell (Ed.), *Equal rights and industrial relations* (pp. 235–281). Madison, WI: Industrial Relations Research Association.

Cain, G. (1986). The economic analysis of labor market discrimination: A survey. In O. Ashenfelter & R. Layard (Eds.), *Handbook of labor economics* (vol. 1, chap. 13, pp. 693–785). Amsterdam: Elsevier Science.

Carlsson, M., & Rooth, D.-O. (2007). Evidence of ethnic discrimination in the Swedish labor market using experimental data. *Labour Economics*, *14*, 716–729.

Carneiro, P., Heckman, J., & Masterov, D. (2005). Labor market discrimination and racial differences in premarket factors. *Journal of Law and Economics*, *48*(1), 1–39.

Chandra, A. (2000). Labor market dropouts and the racial wage gap: 1960–1990. *American Economic Review*, *90*, 333–338.

Charles, K., & Guryan, J. (2011). Studying discrimination: Fundamental challenges and recent progress. *Annual Review of Economics*, *3*, 479–511.

Coate, S., & Loury, G. (1993). Will affirmative action eliminate negative stereotypes? *American Economic Review*, *83*, 1220–1240.

Coates, J., Gurnell, M., & Rustichini, A. (2009). Second-to-fourth digit ratio predicts success among high-frequency financial traders. *Proceedings of the National Academy of Sciences*, *106*(2), 623–628.

Cotton, J. (1988). On the decomposition of wage differentials. *Review of Economics and Statistics*, *70*, 236–243.

Darity, W., & Mason, P. (1998). Evidence on discrimination unemployment: Code of colors, codes of genders. *Journal of Economic Perspectives*, *12*, 63–90.

Dobbie, W., & Fryer, R., Jr. (2009). Are high quality schools enough to close the achievement gap? Evidence from a social experiment in Harlem (Working Paper No. 15473). NBER, Cambridge, MA.

Dohmen, T., Falk, A., Huffman, D., Schupp, J., Sunde, U., & Wagner, G. (2011). Individual risk attitudes: Measurement, determinants and behavioral consequences. *Journal of the European Economic Association*, *9*(3), 522–550.

Donohue, J., & Heckman, J. (1991). Continuous versus episodic change: The impact of civil rights policy on the economic status of blacks. *Journal of Economic Literature*, *29*, 1603–1643.

Dreber, A., & Hoffman, M. (2007). Portfolio selection in utero (Working Paper). University of Chicago, Chicago, IL.

Drydakis, N. (2012). Sexual orientation discrimination in the Cypriot labour market: Distastes or uncertainty? (IZA Discussion Paper No. 6684).

Eckstein, Z., & Wolpin, K. (1999). Estimating the effect of racial discrimination on first job offers. *Review of Economics and Statistics*, *81*, 384–392.

Esmail, A., & Everington, S. (1997). Asian doctors are still being discriminated against. *British Medical Journal*, *314*, 1619.

Espenshade, T., & Chung, C. (2005). The opportunity cost of admission preferences at elite universities. *Social Science Quarterly*, *86*(2), 293–305.

Fershtman, C., & Gneezy, U. (2001). Discrimination in a segmented society: An experimental approach. *Quarterly Journal of Economics*, *116*, 351–377.

Fortin, N. (2005). Gender role attitudes and women's labour market outcomes across OECD countries. *Oxford Review of Economic Policy*, *21*(3), 416–438.

Fortin, N. (2009). Gender role attitudes and women's labor market participation: Opting out, AIDS, and the persistent appeal of housewifery (Working Paper). University of British Columbia.

Fortin, N., & Lemieux, T. (1998). Rank regressions, wage distributions and the gender gap. *Journal of Human Resources*, *33*, 610–643.

Fortin, N., Lemieux, T., & Firpo, S. (2011). Decomposition methods. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 4A, pp. 2–102). Amsterdam: Elsevier Science.

Fryer, R., Jr. (2011). Racial inequality in the 21st century: The declining significance of discrimination. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 4B, pp. 875–971). Amsterdam: Elsevier Science.

Fryer, R., Jr. & Loury, G. (2005). Affirmative action and its mythology. *Journal of Economic Perspectives*, *19*(3), 147–162.

Gayle, G.-L., Golan, L., & Miller, R. (2012). Gender differences in executive compensation and job mobility. *Journal of Labor Economics*, *30*(4), 829–872.

Gobillon, L., Meurs, D., & Roux., S. (2012). Estimating gender differences in access to jobs (IZA Discussion Paper No. 6928).

Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of blind auditions on female musicians. *American Economic Review*, *90*, 715–736.

Gordon, N., & Morton, T. (1974). A low mobility model of wage discrimination with special reference to sex differential. *Journal of Economic Theory*, *7*, 241–253.

Guskey, T., & Pigott, T. (1988). Research on group-based mastery learning programs: A meta-analysis. *Journal of Educational Research*, *81*(4), 197–216.

Hamermesh, D., & Biddle, J. (1994). Beauty and the labor market. *American Economic Review*, *84*(5), 1174–1194.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*, 153–162.

Heckman, J. (1998). Detecting discrimination. *Journal of Economic Perspectives*, *12*, 101–116.

Heckman, J. (2011). The American family in black and white: A post-racial strategy for improving skills to promote equality (IZA Discussion Paper No. 5495).

Heckman, J., Lyons, T., & Todd, P. (2000). Understanding black-white wage differentials, 1960–1990. *American Economic Review*, *Papers and Proceedings*, *90*, 344–349.

Heckman, J., & Siegelman, P. (1993). The Urban Institute audit studies: Their methods and findings. In M. Fix and R. Struyk (Eds.), *Clear and convincing evidence: Measurement of discrimination in America* (pp. 187–258). Washington, DC: The Urban Institute Press.

Hellerstein, J., Neumark, D., & Troske, K. (1999). Wages, productivity and workers' characteristics: Evidence from plant-level production functions and wage equations. *Journal of Labor Economics*, *17*, 409–446.

Holzer, H., & Neumark, D. (2000). Assessing affirmative action. *Journal of Economic Literature*, *38*(3), 483–568.

Holzer, H., & Neumark, D. (2001). What does affirmative action do? *Industrial and Labor Relations Review*, *53*, 240–271.

Johnson, W., & Neal, D. (1998). Basic skills and the black-white earnings gap. In C. Jencks & M. Philips (Eds.), *The black-white test score gap*. Washington, DC: Brookings Institution.

Juhn, C., Murphy, K., & Pierce, B. (1993). Wage inequality and the rise in returns to skill. *Journal of Political Economy*, *101*, 410–442.

Kahn, L. (1991). Discrimination in professional sports: A survey of the literature. *Industrial and Labor Relations Review*, *44*, 395–418.

Kahn, L. (2000). The sports business as a labor market laboratory. *Journal of Economic Perspectives*, *14*, 75–94.

Kunze, A. (2008). The determination of wages and the gender wage gap: A survey. *Empirical Economics*, *35*, 63–76.

Lang, K. (1986). A language theory of discrimination. *Quarterly Journal of Economics*, *10*, 363–382.

Lang, K., & Lehmann, J.-Y. (2012). Racial discrimination in the labor market: Theory and empirics. *Journal of Economic Literature*, *50*(4), 959–1006.

Lang, K., & Manove, M. (2011). Education and labor market discrimination. *American Economic Review*, *101*, 1467–1496.

Lang, K., Manove, M., & Dickens, W. (2005). Racial discrimination in labor markets with posted wage offers. *American Economic Review*, *95*(4), 1327–1340.

Laurent, T., & Mihoubi, F. (2011). Sexual orientation and wage discrimination in France: The hidden side of the rainbow (MPRA Paper 33723). University Library of Munich, Germany.

Levitt, S. (2004). Testing theories of discrimination: Evidence from weakest link. *Journal of Law and Economics*, *47*, 431–452.

Levitt, S., & List, J. (2009). Field experiments in economics: The past, the present, and the future. *European Economic Review*, *53*, 1–18.

List, J. (2004). The nature and extent of discrimination in the marketplace: Evidence from the field. *Quarterly Journal of Economics*, *119*, 49–89.

Loury, G. (1998). Discrimination in the post–civil rights era: Beyond market interactions. *Journal of Economic Perspectives*, *12*, 117–126.

Loury, G. (2002). *The anatomy of racial inequality*. Cambridge, MA: Harvard University Press.

Lundberg, S., & Startz, R. (1983). Private discrimination and social intervention in competitive labor markets. *American Economic Review*, *73*, 340–347.

Mobius, M., & Rosenblat, T. (2006). Why beauty matters. *American Economic Review*, *96*(1), 222–235.

Mulligan, C., & Rubinstein, Y. (2008). Selection, investment, and women's relative wages over time. *Quarterly Journal of Economics*, *123*, 1061–1110.

Neal, D., & Johnson, W. (1996). The role of premarket factors in black-white wage differences. *Journal of Political Economy*, *104*, 869–895.

Neumark, D. (1999). Wage differentials by race and sex: The roles of taste discrimination and labor market information. *Industrial Relations*, *38*(3), 414–445.

Neumark, D. (2012). Detecting discrimination in audit and correspondence studies. *The Journal of Human Resources*, *47*(4), 1128–1157.

Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *Quarterly Journal of Economics*, *122*(3), 1067–1101.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International Economic Review*, *14*, 693–709.

Oaxaca, R., & Ransom, M. (1994). On discrimination and the decomposition of wage differentials. *Journal of Econometrics*, *61*, 5–21.

OECD. (2008). The price of prejudice: Labour market discrimination on the ground of gender and ethnicity. In *OECD employment outlook*, chap. 3.

OECD. (2012). *Closing the gender gap: Act now*. Paris: OECD Publishing.

Olds, D. L., Robinson, J. A., O'Brien, R., Luckey, D. W., Pettitt, L. M., Henderson, C. R., Ng, R. K., Sheff, K. L., Korfmacher, J., Hiatt, S., & Talmi, A. (2002). Home visiting by paraprofessionals and by nurses: A randomized, controlled trial. *Pediatrics*, *110*(3), 486–496.

O'Neill, J., & O'Neill, D. (2006). What do wage differentials tell us about labor market discrimination? In S. Polackek, C. Chiswick, & H. Rapoport (Eds.), *The economics of immigration and social diversity* (vol. 24, pp. 293–357). Bingly, U.K.: Emerald Group Publishing.

Phelps, E. (1972). The statistical theory of racism and sexism. *American Economic Review*, *62*, 639–651.

Plug, E., & Berkhout, P. (2004). Effects of sexual preferences on earnings in the Netherlands. *Journal of Population Economics*, *17*(1), 117–131.

Plug, E., Webbink, H., & Martin, N. (2014). Sexual orientation, prejudice and segregation. *Journal of Labor Economics*, *32*(1), 123–159.

Price, J., & Wolfers, J. (2010). Racial discrimination among NBA referees. *Quarterly Journal of Economics*, *125*(4), 1859–1886.

Puma, M., Bell, S., Cook, R., & Heid, C. (2010). Head Start impact study: Final report. Washington, DC: U.S. Department of Health and Human Services.

Regan, T., & Oaxaca, R. (2009). Work experience as a source of specification error in earnings models: Implications for gender wage decompositions. *Journal of Population Economics*, *22*, 463–499.

Riach, P., & Rich, J. (1991). Testing for racial discrimination in the labour market. *Cambridge Journal of Economics*, *15*, 239–256.

Riach, P., & Rich, J. (2002). Field experiments of discrimination in the market place. *Economic Journal*, *112*, F480–F518.

Ritter, J., & Taylor, L. (2011). Racial disparity in unemployment. *Review of Economics and Statistics*, *93*, 30–42.

Sander, R. (2004). A systematic analysis of affimative action in American law schools. *Stanford Law Review*, *57*, 367–483.

Suen, W. (1997). Decomposing wage residuals. *Journal of Labor Economics*, *15*, 555–566.

Tilcsik, A. (2011). Pride and prejudice: Employment discrimination against openly gay men in the United States. *American Journal of Sociology*, *117*(2), 586–626.

PART THREE

Job Creation, Job Destruction, and Unemployment

CHAPTER 9

Equilibrium Unemployment

In this chapter we will:

- See that during the last 40 years, the industrialized countries have evolved in very different directions with respect to unemployment
- Observe the magnitude of job creation, job destruction, and worker flows
- Discover the meaning and the importance of the Beveridge curve
- Analyze the functioning of the labor market as a matching process between vacant jobs and job seekers
- Think about the efficiency of a labor market with trading externalities
- Analyze the dynamics of vacancies and unemployment

INTRODUCTION

All developed economies are affected by unemployment. But they are affected in ways that vary across a strikingly wide spectrum. In the first quarter of 2013, Austria posted an unemployment rate of 4.8%, while in the United States 7.7% of the workforce was unemployed, and in Spain the figure reached 26.5%. The variation in the dynamic of unemployment is also striking. Between 1960 and 1994 Japan experienced a very stable rate of unemployment, but then it rose steadily until 2001. Meanwhile the American unemployment rate was undergoing ceaseless fluctuation. The duration of spells of unemployment also varies widely from one country to another. Long-term unemployment is a phenomenon proper to certain countries of continental Europe like Greece, Italy, Belgium, and France, while durations are much shorter in the United States and Japan.

To understand unemployment, one must bear in mind that in every developed country, jobs and manpower experience movements of considerable magnitude. Differences in the profile of unemployment from one country to another are largely explainable by different approaches to managing these ebbs and flows. In all OECD countries, workers' mobility among the different possible states in the labor market (from one job

to another, from holding a job to looking for one, from unemployment to nonparticipation, etc.) is a phenomenon of major dimensions. Every month from 1996 to 2003 in the United States, for example, 2.6% of wage earners changed jobs, 0.8% became unemployed, and 2.7% ceased to participate in the labor market at all.

For a worker, the search for a job that fits her requirements and skills is a process that often takes a lot of time. Likewise, when a firm wants to recruit new workers, it often chooses to devote substantial resources (with a corresponding cost in time) to the selection of suitable individuals. There are imperfections in the information available in the labor market; the result is the simultaneous presence of unemployed persons and vacant jobs. This is the origin of *frictional* unemployment.

The intensity of the processes of job destruction and creation has an effect on the level of frictional unemployment. When the economy is restructured, job rotation increases workers' mobility and thus pushes up frictional unemployment. But frictional unemployment also depends on factors of a more institutional kind, like the amount of unemployment benefit, which determines how long the unemployed person can wait, or the level of hiring and firing costs, which influences the behavior of firms. The first dynamic analyses of the labor market date from the 1960s. They were based principally on the job search behavior of workers and explained frictional unemployment by the fact that the unemployed reject job offers that pay wages they consider too low, in the hope of subsequently receiving more attractive offers. We saw in chapter 5 that the main determinants of unemployment duration are the unemployment benefits, the arrival rate of job offers, and the characteristics of the distribution of possible wages.

In this chapter we study a complementary approach, which brings in the behavior of firms when faced with a costly hiring process. This approach envisages the hiring process as a phenomenon of *matches* between employers and workers. In this framework, the probability for every unemployed person to receive a job offer suited to her abilities depends on the *tightness* prevailing in the labor market, that is, the ratio of the number of vacant jobs to the number of unemployed persons. If this ratio is high (many vacant jobs, few job seekers), every unemployed person has a high probability of finding a job. Symmetrically, each person's probability of finding and filling a vacant job has to decrease when this ratio decreases (few vacant jobs, many job seekers). This representation of the process of matching up jobs and workers, developed especially by Hall (1979a, 1979b), Bowden (1980), and Pissarides (1979, 2000), makes it possible to analyze the determinants of unemployment in a setting that takes into explicit consideration the transaction costs linked to labor mobility and the imperfection of information in the labor market. In particular, it allows us to grasp the determinants of unemployment in a dynamic environment where jobs are created and destroyed continually and in which there are transaction costs attached to reallocating employment.

The first section lays out the main facts about unemployment, manpower mobility, and the processes of job creation and destruction as they emerge from empirical studies. Section 2 presents the determination of the equilibrium of a labor market with labor adjustment costs when there is perfect information and highlights the limitations of such a competitive model. Then section 3 presents the basic matching model. This model takes the flow of jobs into consideration and is grounded in an imperfectly competitive mode of wage formation. Section 4 discusses the efficiency of the equilibrium resulting from such a model. Section 5 introduces capital explicitly in order to focus on the relationships between investment, the interest rate, and unemployment. Section 6

analyzes the dynamics of vacancies and unemployment and discusses the ability of the matching model to reproduce some stylized facts about these dynamics.

1 FACTS

To grasp the determinants of unemployment, we need a dynamic perspective that takes in the reallocation of jobs and movements in manpower. This section adopts such a perspective to describe the progression of unemployment rates, job creation and destruction, and movements of manpower in OECD countries.

1.1 Unemployment, Employment, and Participation

There exists great diversity in the progression of unemployment in the most developed economies. Certain European countries, in particular those on the continent, have experienced a rise in unemployment since the 1970s that has proved difficult to halt. The economic crisis of 2008–2009 made the situation worse. It is generally observed as well that countries with high rates of employment and participation have lower rates of unemployment than others. Hence the growth of unemployment results more from a lack of job creation than from an increase in the active population (those in work or seeking work). Over the long term, certain countries fail to create enough jobs to absorb the growth in their active populations while others succeed in doing so.

1.1.1 Different Experiences of Unemployment

According to the standardized definition based on the 13th Conference of Labour Statisticians—generally referred to as the International Labor Office (ILO) guidelines—unemployment comprises all persons who, during the reference period (1) were without work, that is, were not in paid employment or self-employment during the reference period; (2) were currently available for work, that is, were available for paid employment or self-employment during the reference period; (3) were seeking work, that is, had taken specific steps in a specified recent period to seek paid employment or self-employment.

Table 9.1 shows average rates of unemployment, labor market participation, and employment in 25 OECD countries in 2000–2011. We see that unemployment affects all OECD countries but in very different proportions. Some countries, like the United States, Japan, Norway, the Netherlands, Switzerland, and the United Kingdom, have an unemployment rate at or below 6%. But other countries, like France, Greece, and Spain, display an unemployment rate at or above 9%. For the European Union as a whole, the average unemployment rate over the first decade of the century is in the neighborhood of 9%.

The second column of table 9.1 reports the employment rates—the ratio of the number of persons employed to the number of persons in the population who are of working age (from 15 to 64 years old). This indicator is a useful complement to the data on unemployment, given that the difference between unemployment and inactivity, which may be fuzzy (as stressed in chapter 5, section 2.1.3), can be influenced

TABLE 9.1
Average rates of unemployment, participation, and employment in 25 OECD countries, the European Union, Brazil, and Russia, 2000–2011.

	Employment rate (%)	Participation rate (%)	Unemployment rate (%)
Australia	71	75	5.5
Austria	70	73	4.4
Belgium	61	66	7.7
Canada	72	78	7.2
Chile	56	61	8.6
Denmark	76	80	5.2
Finland	69	75	8.3
France	64	70	8.8
Germany	68	74	8.6
Greece	59	66	10.7
Italy	57	62	8.2
Ireland	65	70	6.9
Israel	58	63	8.5
Japan	70	73	4.9
Korea	63	66	3.8
Mexico	60	62	3.8
Netherlands	73	76	3.9
Norway	76	79	3.6
Portugal	67	73	7.9
Spain	62	71	13.0
Sweden	74	80	6.7
Switzerland	79	82	3.8
Turkey	46	51	10.6
United Kingdom	72	76	5.8
United States	71	75	6.2
European Union (27 countries)	62	68	8.8
OECD (34 countries)	65	70	7.0
Brazil	67	73	8.9
Russian Federation	66	71	7.8

Source: OECD Labor Force Statistics database.

by particular characteristics of each labor market. All the figures given in table 9.1 correspond to the standardized definition of unemployment, but national specificities are important sources of heterogeneity. For example, generous unemployment benefits may impel individuals to look for a job, or claim to be doing so, to gain access to unemployment benefits.

Scrutiny of table 9.1 and figure 9.1 indicates however that countries with relatively high unemployment rates also have relatively low rates of employment. In particular,

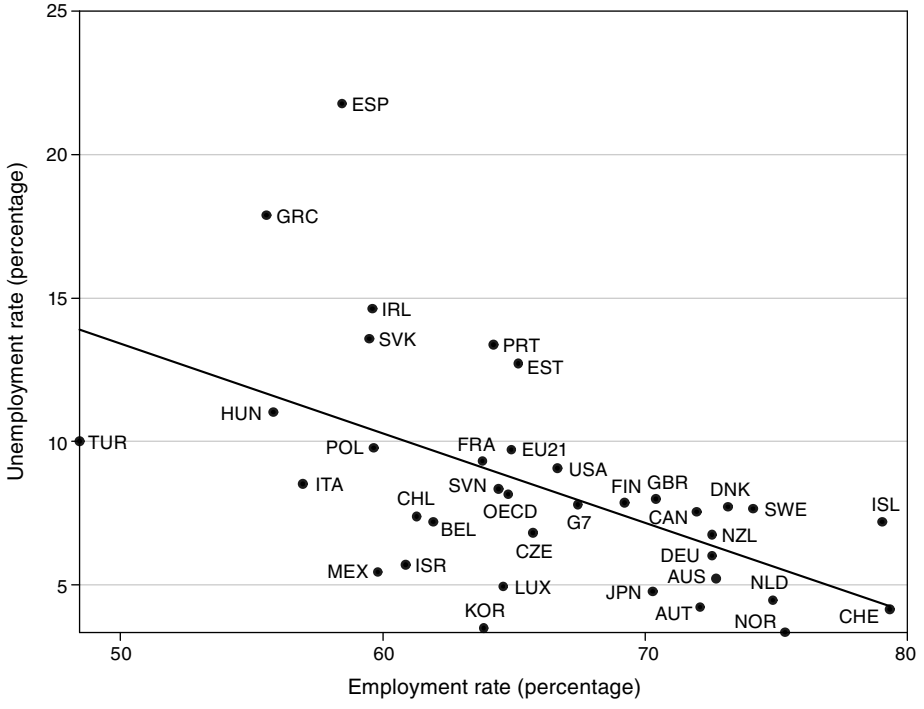


Figure 9.1
The relationship between the unemployment rate and the employment rate in the OECD countries in 2011.

Source: OECD Labor Force Statistics database.

figure 9.1 shows that there exists a decreasing relation between the unemployment rate and the rate of employment and that the dispersion of the cluster of points around the regression line is relatively weak. The unemployment rate is thus a relevant indicator of the abundance of jobs in a country. The third column of table 9.1 also shows that participation rates (the participation rate is the ratio of the labor force to the working-age population) are highly dispersed, since they vary from 51% in Turkey to 82% in Switzerland. Moreover, countries that face a high unemployment rate generally have a relatively weak rate of participation. This observation is illustrated in figure 9.2, which reveals a decreasing relation between participation rates and unemployment rates. Thus, high unemployment does not result from an excessively high participation rate.

This rapid overview of unemployment, employment, and labor market participation as experienced in different OECD countries suggests that certain countries face a relatively high unemployment rate because of insufficient job creation, not abnormally high participation rates. Examination of *changes over time* since the beginning of the 1960s in employment, unemployment, and the labor force in the United States, Japan, and three continental European countries—Germany, France, and Italy—that have experienced high unemployment will throw further light on the origins of underemployment.

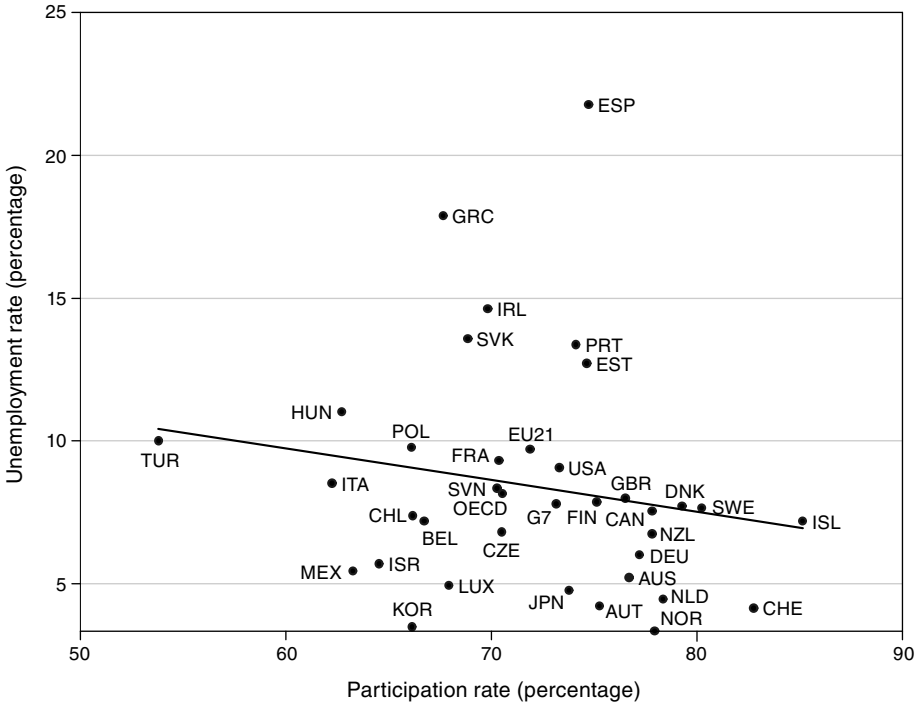


FIGURE 9.2
The relationship between the unemployment rate and the participation rate in the OECD countries in 2011.

Source: OECD Labor Force Statistics database.

1.1.2 CHANGES IN EMPLOYMENT, UNEMPLOYMENT, AND LABOR FORCE

Figure 9.3 shows that the unemployment rate has evolved very differently in Japan, the continental European countries—Germany, France, and Italy—and the United States. Between 1960 and 1994 Japan was characterized by great stability in this indicator, so much so that the two oil shocks of 1974 and 1979 seem not to have had much impact. But between 1994 and 2001 the unemployment rate rose steadily in this country. Conversely, the American unemployment rate fluctuates significantly. The unemployment rate in the continental European countries stayed relatively low until the 1970s but rose steadily until the late 1990s; from then to 2008, it diminished, but it increased steeply again after 2008, during the Great Recession.

The Relations Between Unemployment, Employment, and the Labor Force

We can assess change in the unemployment rate with the help of the following accounting equality:

$$N_t \tau_t = L_t + U_t$$

In this relation N_t, L_t, U_t, and τ_t designate respectively the population of working age, the level of employment, the number of unemployed, and the participation rate at period t.

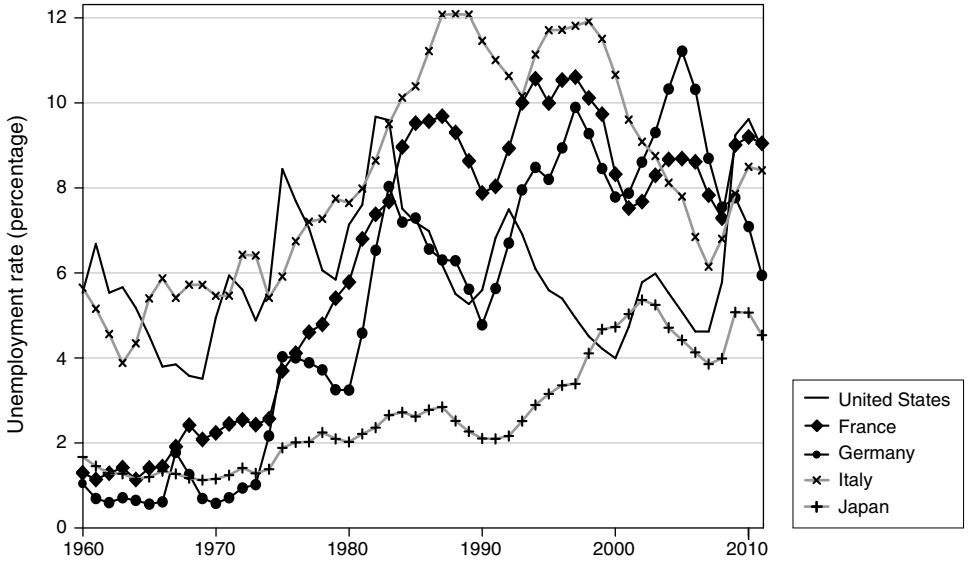


Figure 9.3
Unemployment rate in the United States, Japan, and continental Europe (Germany, France, Italy), 1960–2011 (persons aged 15 to 64).

Note: For Germany, estimate based on the annual growth rate for West Germany before 1991.

Source: OECD Economic Outlook database.

The unemployment rate being defined by $u_t \equiv U_t/(L_t + U_t)$, we have:

$$N_t \tau_t = \frac{L_t}{1 - u_t}$$

Using this equation in logarithms at dates t and $t-1$, and using the approximation $\lim_{x \to 1} \ln x = x - 1$, we get ($\Delta$ is the difference operator, $\Delta N_t = N_t - N_{t-1}$):

$$\frac{\Delta N_t}{N_{t-1}} + \frac{\Delta \tau_t}{\tau_{t-1}} = \frac{\Delta L_t}{L_{t-1}} + \frac{\Delta u_t}{1 - u_{t-1}}$$

With the assumption that u is a small number, which is the case in reality, this relation allows us to express the variations in the unemployment rate as a function of the growth rates of the working-age population, employment, and participation:

$$\Delta u_t \simeq \frac{\Delta N_t}{N_{t-1}} + \frac{\Delta \tau_t}{\tau_{t-1}} - \frac{\Delta L_t}{L_{t-1}}$$

This decomposition shows that variations in the unemployment rate come from variations in the employment rate, from changes in the size of the working-age population, and from changes in the participation rate. The relationship between the unemployment rate and employment is thus not a simple one. It is entirely possible for the unemployment rate to fall without employment rising, if, for example, the labor force

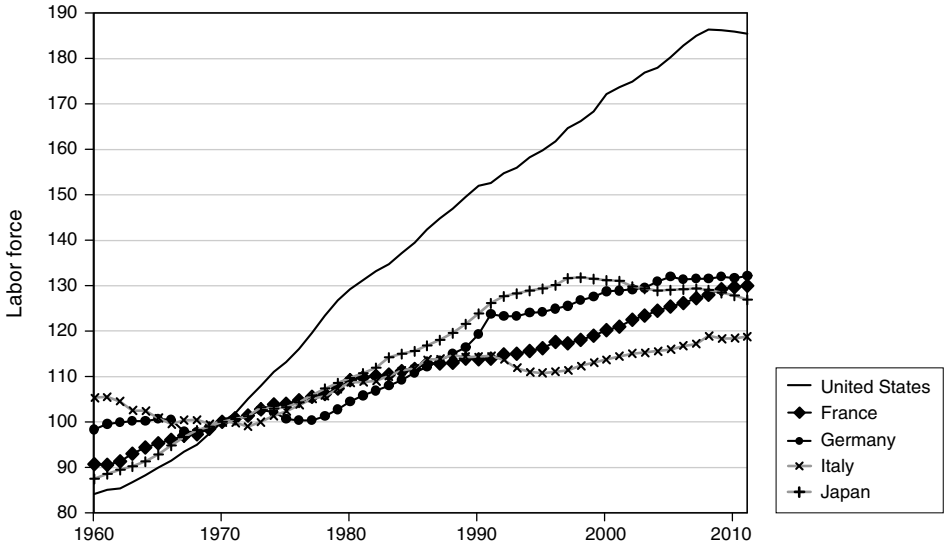


FIGURE 9.4
Changes in the labor force in the United States, Japan, and continental Europe (Germany, France, Italy), 1960–2011, base 100 in 1970.

Note: For Germany, estimate based on the annual growth rate for West Germany before 1991.

Source: OECD Economic Outlook database.

shrinks. Figure 9.4 shows that the growth of unemployment in Europe is not the upshot of this scenario; for Germany, France, and Italy, where the unemployment rates rose steeply until 1997, experienced relatively slow rates of expansion of the labor force compared to the United States, or even Japan (the sudden jump in the labor force in Germany in 1990 comes from German reunification).

The Chronic Weakness of Job Creation in Continental Europe

We observe in figure 9.4 that, without exception, the expansion of the labor force is weaker in the continental European countries and Japan than it is in the United States. The relatively strong expansion of the labor force in the United States is the result of a rise in the rates of participation and a more sustained growth in the size of the working-age population. It is interesting to note—see figure 9.5—that the labor force participation *rate* has risen considerably in the United States and Japan, but also Germany, whereas it declined in Italy and remained broadly stable in France. Thus the good performances of the United States and Japan when it comes to unemployment are not due to less growth in their labor forces. We observe further that labor force rates are higher in the countries where unemployment has not risen or has risen less over the period as a whole.

All these elements suggest that the United States, and to a lesser extent Japan, clearly have a greater capacity to create jobs than do France and Italy and than Germany did until recently. This conclusion emerges sharply in figure 9.6, which shows how the number of jobs evolved compared to levels in 1970: the increase in employment in the

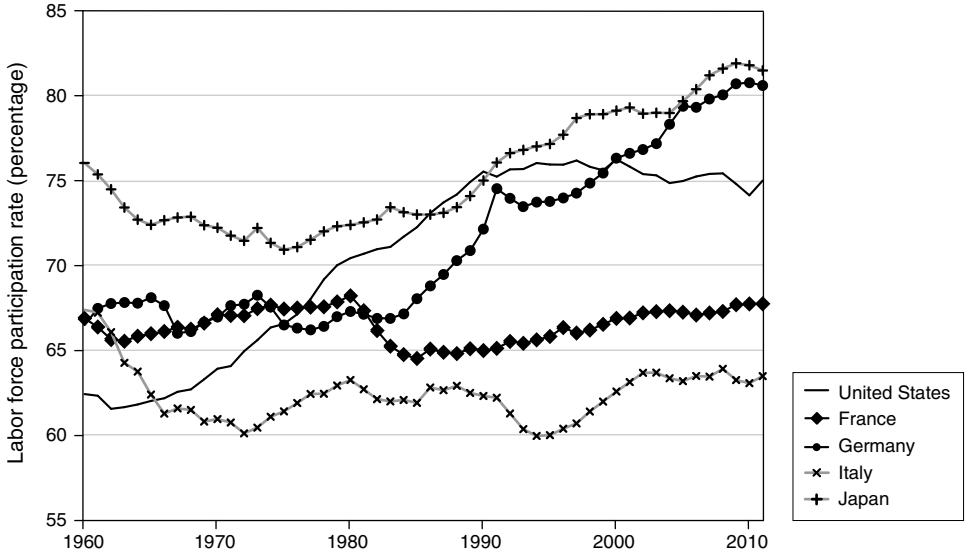


Figure 9.5
Labor force participation rate in the United States, Japan, and continental Europe (Germany, France, Italy), 1960–2011 (persons aged 15 to 64).

Note: For Germany, estimate based on the annual growth rate for West Germany before 1991.

Source: OECD Economic Outlook database.

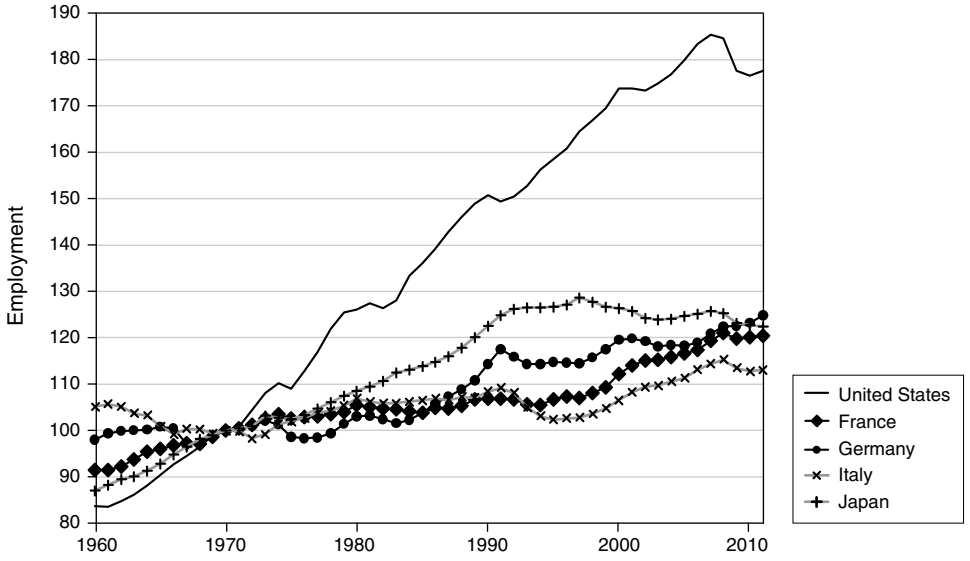


Figure 9.6
Changes in employment in the United States, Japan, and continental Europe (Germany, France, Italy), 1960–2011. Base 100 in 1970.

Note: For Germany, estimate based on the annual growth rate for West Germany before 1991.

Source: OECD Economic Outlook database.

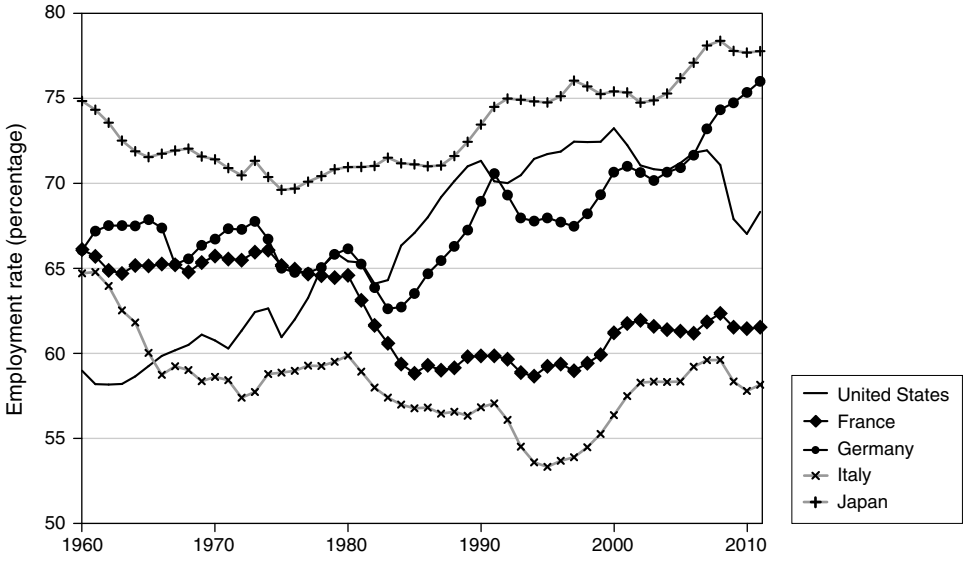


FIGURE 9.7
Employment rates in the United States, Japan, and continental Europe (Germany, France, Italy), 1960–2011 (persons aged 15 to 64).

Note: For Germany, estimate based on the annual growth rate for West Germany before 1991.

Source: OECD Economic Outlook database.

United States since 1970 has been greater than in Japan and in the European countries on which we have focused.

Additionally, figure 9.7 tells us that the European countries' poor performance in job creation leads to low employment *rates*. Since 1980 in Germany, France, and Italy, the employment rate (which, readers will remember, equals the ratio of the number of jobs to the size of the working-age population) has continuously been lower than that of the United States and Japan—though there has been a change since the late 2000s for Germany, where the employment rate rose during the Great Recession. Moreover, the employment rate has risen constantly in the United States and Japan since 1975. Thus, at the beginning of the third millennium the difference in employment rates between the continental European countries on one hand, and the United States and Japan on the other, is considerable and clearly larger than the difference in unemployment rates.

In sum, the picture is particularly negative for European countries such as Germany (until recently), France, Italy, Belgium, and Spain. It reveals a structural incapacity to create enough jobs for over 30 years. During the 1960s, this lack was offset by a significant fall in the overall participation rate. But since 1970 the latter variable has remained more or less stable, and the weakness of job creation has been fully reflected in unemployment.

1.1.3 LONG-TERM UNEMPLOYMENT

A very high proportion of long-term unemployed persons—those who have been looking for a job for more than a year—clearly sets many countries of continental Europe

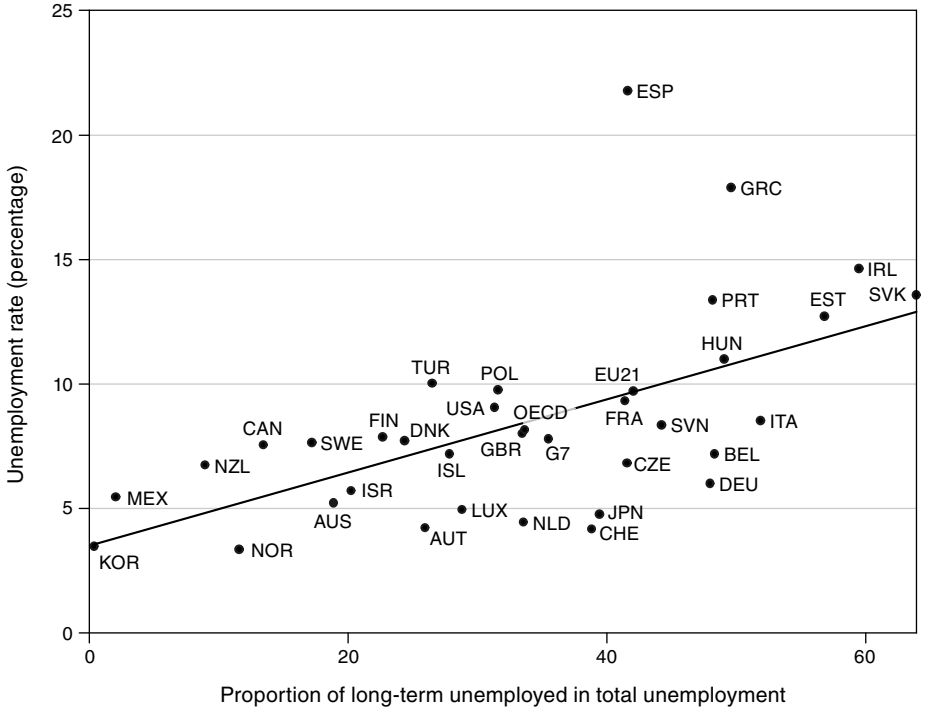


Figure 9.8
The relationship between the unemployment rate and the proportion of long-term unemployed in the OECD countries in 2011.

Source: OECD Labor Force Statistics database.

apart from certain other industrialized countries. Figure 9.8 shows that the long-term unemployed represent about 50% or more of overall unemployment in Belgium, Estonia, France, Germany, Greece, Ireland, Italy, Portugal, and Slovakia. The corresponding share in the United States and Japan is approximately 15 points lower, and even about 30 points lower in the Nordic European countries. Long-term unemployment is a phenomenon proper to certain countries of continental Europe. It is capable of having dire effects on the "employability" of suppliers of labor and constitutes an important source of degradation in the overall functioning of the labor market. The overall level of unemployment is intimately tied to the level of long-term unemployment. As we see in figure 9.8, the countries where the unemployment rate is high are also those with a strong percentage of long-term unemployed.

1.2 Jobs Flows

Two kinds of data allow us to understand the dynamics of the labor market better. The first pertains to the processes of job creation and destruction, and the second to worker flows. Net variations in the volume of employment over a given period are equal, by definition, to the difference between job creations and job destructions over that period. They are also equal to the difference between workers' entries into and exits out of

employment. In other words, variations in employment may be defined on the basis of the two following identities:

$$\text{Net employment change} = \underbrace{\text{Creations} - \text{Destructions}}_{\text{Job flows}} = \underbrace{\text{Hirings} - \text{Separations}}_{\text{Worker flows}}$$

Examination of the data reveals that the labor market is characterized by intense reallocation of jobs and workers (see Davis et al., 2006, for an excellent presentation).

1.2.1 Job Creation and Destruction

Table 9.2 provides information on the magnitude of job creation and destruction in a number of OECD and emerging economies. In this table, job creation represents the sum of job gains measured at the plant or firm level (according to the studies) over one year due to the opening of new production units and the expansion of jobs within existing workplaces. Job destruction represents the sum of job losses resulting from the closing of production units and contractions in the number of jobs in units that stay open over the same period. The job reallocation is equal to the sum of job creation and

Table 9.2
Job creation and destruction flows. Annual average rate as a percentage of total employment, all sectors of the economy.

Country (period)	Job creation	Job destruction	Job reallocation	Net employment	Excess job reallocation
Argentina (95–02)	12.7	10.7	23.4	2.0	21.4
Brazil*(96–01)	16.1	12.9	29.0	3.2	25.8
Chile*(79–99)	11.6	11.3	22.8	.3	22.5
Colombia*(82–98)	10.5	10.0	20.5	.5	20.0
Estonia (95–01)	13.3	12.0	25.3	1.3	24.0
Finland (88–98)	13.8	14.0	27.8	–.2	27.6
France (99–00)	12.0	8.3	20.3	3.7	16.6
Germany (77–99)	8.4	7.1	15.5	1.3	14.2
Hungary (92–01)	13.3	11.2	24.5	2.1	22.4
Italy (86–94)	12.3	10.2	22.5	2.1	20.4
Latvia (96–02)	15.7	10.8	26.5	4.9	21.6
Mexico (85–01)	16.9	12.0	28.9	4.9	24.0
Portugal (83–98)	12.5	10.7	23.3	1.8	21.4
Slovenia (92–01)	9.0	8.1	17.1	.9	16.2
United Kingdom*(80–98)	11.5	12.6	24.2	–1.1	23.1
United States (88–97)	12.5	10.0	22.5	2.5	20.0

Note: *Manufacturing only for Brazil, Chile, Colombia, and the United Kingdom. West Germany for Germany. Mostly private sector for Germany, Portugal, and the United States.

Source: Data from Haltiwanger et al. (2010); see their table A.1 for exact sources, except for France where data is from Picart (2008, table 2).

job destruction, whereas the net employment growth is equal to the difference between these two quantities. Excess job reallocation corresponds to the difference between job reallocation and the absolute value of net employment growth. Rates are expressed as a percentage of the average employment over the year.

In the first place, there are wide variations in job creation and destruction flows, with job reallocation ranging from 15.5% in Germany to 29% in Brazil. Correcting these flows for variation in industry composition across economies (since some sectors such as services feature more turnover than others) does not fully explain these differences (OECD, 2010). They might also relate to the share of temporary employment and the degree of stringency of employment protection legislation (EPL). Another dimension which may vary across countries and explain disparities is the distribution by size of firms, with small firms typically experiencing larger job reallocation than large firms as a percentage of their employment. In particular, countries with low excess reallocation have lower shares of temporary employment, while low EPL tends to be associated with higher reallocation rates (OECD, 2010; Haltiwanger et al., 2010). In the second place, it is evident that for all countries net employment growth is always much smaller than job creation or destruction. In the United States, for example, 10% of jobs are destroyed every year, while the proportion of jobs created with respect to the stock of existing jobs is equal to 12.5%: the net employment growth rate is thus of the order of 2.5% per year. In the third place, we observe that job reallocation belongs to a different order of magnitude than net employment growth, being about 15 times higher on average in table 9.2 (25% compared with 1.7%). This means that the excess job reallocation, equal to the difference between job reallocation and the absolute value of net employment growth, is considerable. In the United States, it would have sufficed to reallocate 2.5% of jobs in order to transform production units, but it needed a reallocation of 22.5%—an excess job reallocation of 20%—in order for these reallocations actually to take place.

It should be noted that the job creation and destruction presented in table 9.2 do not include job reallocations that take place within individual firms or plants. For example, a firm that gets rid of a worker's job in order to create a managerial job is recorded as having job creation and destruction equal to zero. Studies that have attempted to assess job reallocations within workplaces suggest that this factor is not negligible. Hamermesh et al. (1996) use a survey which indicates whether hires correspond to newly created jobs in the Netherlands. They find that reorganizations within firms explain 11% of overall job reallocations. Using data on the structure of job creation and destruction in relation to skill within firms in France, Lagarde et al. (1995) estimate that job reallocations within firms are much greater than that, representing almost half of all job reallocations.

1.2.2 The Extent of Within-Sector Reallocation

Contrary to what is sometimes stated as obvious fact, job movements most frequently take place within the *same* sector, not between *different* sectors. It is possible to assess the extent of within-sector reallocation by comparing two indicators (see Davis and Haltiwanger, 1992). If S designates the number of sectors, we look at the net employment growth in a given sector s (V_n^s) and the net employment growth in the economy as a

whole (V_n). An initial indicator assesses the extent of job reallocations due to between-sector movements. It is defined by:

$$R_E = \sum_{s=1}^{S} |V_n^s| - |V_n|$$

Let T_s be the job reallocation in sector s; the second indicator corresponds to the sum of excess job reallocations within each sector. It is defined by:

$$R_I = \sum_{s=1}^{S} (T_s - |V_n^s|)$$

The fraction of job reallocations due to between-sector shifts is then measured by the ratio $R_E/(R_I + R_E)$. Table 9.3 shows that job movements are to a large extent within sectors.

It turns out that between-sector reallocations are never more than a small component of overall job reallocations, even when sectors are broken down finely. Since the beginning of the 1980s, the process of job creation and destruction has thus been essentially within sectors.

1.2.3 Job Creation and Destruction over Time

Figure 9.9 presents quarterly rates of job creation and destruction in the United States from 1990 to 2010. These rates are high over the whole period. Contrary to the notion sometimes put forward that jobs are ever more unstable, these rates show no upswing over this period. Rather we observe a slight declining trend. Davis et al. (2006) show that this declining trend is also present in the manufacturing sector, where the available data cover a longer period than the timeframe shown in figure 9.9. In this sector the rates of job creation and destruction, on the order of 7% per quarter in the 1950s, lie at around 5% for the 1980s and 1990s.

Figure 9.9 also shows that job creation is strongly procyclical: it falls during recessions. Conversely, job destruction is countercyclical: it rises during recessions. In the United States, cycles are marked by weak variations in the number of jobs created and strong variations in the number of jobs destroyed. This entails that the rate of job

Table 9.3
Fraction of job reallocation accounted for by employment shifts between sectors.

Country	Period	Number of sectors	$R_E/(R_I + R_E)$
Germany	83–90	24	0.03
United States	72–88	980	0.14
France	84–88	15	0.06
France	84–91	600	0.17
Italy	86–91	28	0.02
Sweden	85–91	28	0.03

Source: Davis and Haltiwanger (1999, table 5).

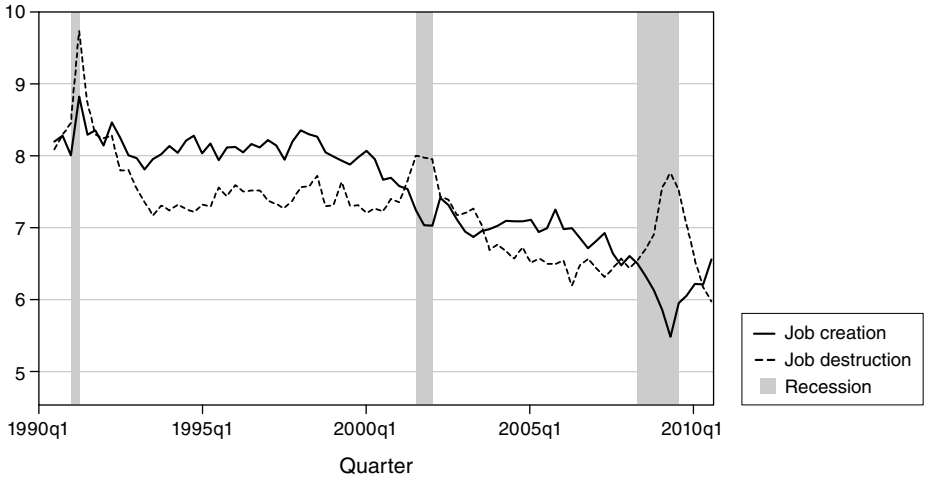


Figure 9.9
Quarterly job flows in the United States. Private sector, 1990q2–2010q2. Job creation rate and job destruction rate in percentage of employment.

Source: Davis et al. (2012) database.

reallocation is countercyclical: there is more job reallocation in phases of recession. This result is not observed in all OECD countries. Job destruction is generally countercyclical and job creation procyclical, but job destruction does not always vary to a significantly greater degree than job creation (see OECD, 1996, chapter 5).

1.3 Worker Flows

Workers' transitions between the various states they may occupy—employment, unemployment, or nonparticipation—provide a good overview of worker flows. Figure 9.10 presents monthly movements of manpower in the United States for the period 1996–2003. It shows that around 15 million persons changed states or changed jobs from one month to the next over this period, which corresponds to around 8% of the working-age population. Manpower movements thus involve a significant part of the population: every month 2.6% of wage earners change jobs, 0.8% enter into unemployment, and 2.7% pass into nonparticipation.

We will now describe more precisely inflows and outflows of employment and unemployment, as well as worker flows over business cycles.

1.3.1 Employment Inflows and Outflows

Worker flows are different from job flows, for in addition to entries and exits linked to the creation and destruction of jobs, they also include rotations in the same job: a number of workers may follow one another into and out of the same job. With data on French firms 1987–1990, Abowd et al. (1999) estimate that over the course of a year, the creation of one job corresponds to the hiring of three persons and the separation of two. As a general rule, worker reallocations clearly bulk larger than job reallocations.

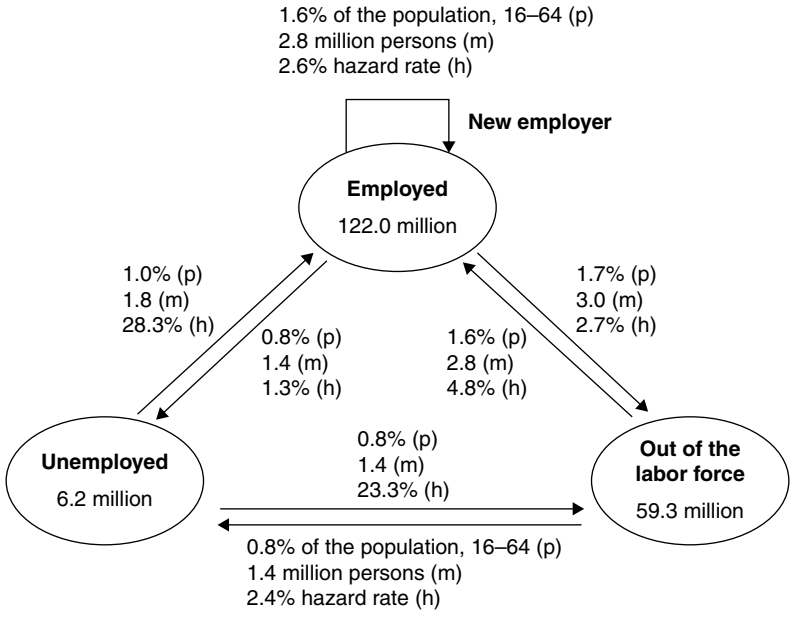


FIGURE 9.10
Average monthly worker flows in the United States. Current Population Survey, 1996–2003. The hazard rate is the rate at which persons change status each month.

Source: Davis et al. (2006, figure 1).

They are assessed by observing, for a given period, the flow of entries and exits from unemployment on one hand, and the flow of entries and exits from employment on the other. An entry into employment corresponds to a hire, and an exit from employment to a separation. Table 9.4 portrays the flow of entries and exits from employment for 26 OECD countries during the year 2011.

Table 9.4 highlights the magnitude of entries and exits from employment with respect to the stock of jobs. Worker flows are seen to be systematically greater in size than job flows. Thus, the exit rate from employment in table 9.4 is, for most countries, almost twice as large as the rate of job destruction given in the second column of table 9.2. Likewise, the rate of entry into employment is about twice as high as the rate of job creation set out in the first column of table 9.2. We observe too that worker mobility differs from country to country. The rates of entry and exit from employment are relatively high in Australia, Finland, Korea, and Mexico, whereas they are between two and three times lower in Greece, Italy, and the Czech and Slovak republics. These countries are thus characterized by low worker rotation.

Job-to-Job Mobility

An exit from employment leads to unemployment, nonparticipation, or a new hire for the person concerned. Someone may enter into employment from unemployment, nonparticipation, or another job. Figure 9.11 shows that a little less than half of

Table 9.4
Annual employment inflows and outflows. In percentages, for the year 2011.

Country	Entry rate (hirings)	Exit rate (separations)
Australia	23	24
Austria	16	16
Belgium	14	12
Canada	20	18
Czech Republic	12	11
Denmark	22	22
Finland	22	21
France	14	14
Germany	16	14
Greece	9	16
Hungary	14	13
Ireland	13	14
Italy	11	10
Korea	36	33
Mexico	26	25
Netherlands	17	4
Norway	16	15
Poland	14	13
Portugal	15	15
Slovak Republic	10	8
Spain	17	17
Sweden	21	18
Switzerland	18	16
Turkey	35	27
United Kingdom	15	14
United States	19	22
European Union (15 countries)	15	14
OECD (30 countries)	18	18

Note: 2010 for Australia, Canada, and the United States. OECD average recalculated as the weighted average of countries shown in this table, plus Estonia, Iceland, Luxembourg, and Slovenia.
Legend: The entry rate is calculated as the ratio of persons employed for less than one year to the average stock of employment over the year and the exit rate as the difference between the employment growth rate and the entry rate.

Source: OECD Labor Force Statistics database.

entry-and-exit flows from employment involve persons who are not in employment. The remainder of these flows come from direct worker mobility between two jobs with no interval of unemployment. Hence direct job-to-job mobility represents a substantial portion of all manpower movement in the OECD countries as a whole.

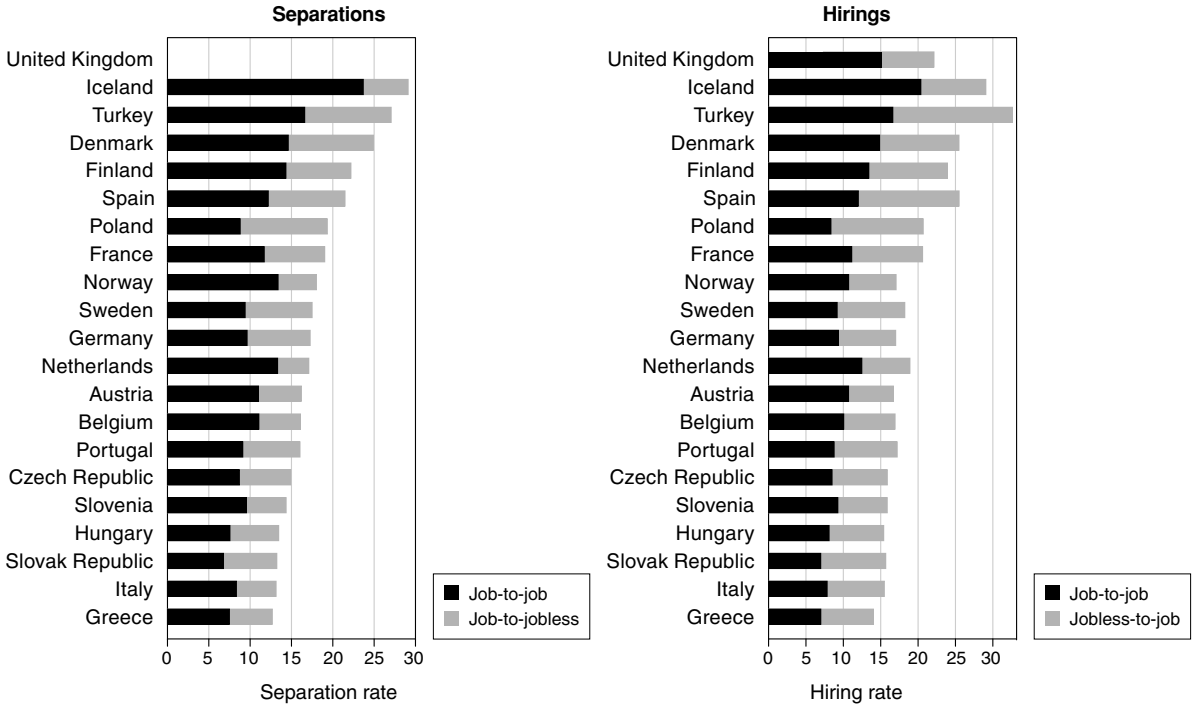


FIGURE 9.11

Job-to-job, jobless-to-job, and job-to-jobless flows in the European countries, 2000–2007.

Note: Country average rates expressed in percentages and adjusted for industry composition. Years around 2000–2007.

Source: OECD (2010, figure 3.2, p. 175).

On Displacements

Exits from employment comprise quits, the ending of short-term contracts, retirements, firings for cause, and job loss through no fault of the employee. By definition, displaced workers belong to the last category: they are defined as persons who lost or left jobs because their plant or company closed or moved, there was insufficient work for them to do, or their position or shift was abolished. It is interesting to compare figures for overall worker movements with those for displacements alone.

Figure 9.12 reproduces the values of the displacement rate for several industrialized countries. Job displacements are defined as job separations from firms that, from one year to the next, experienced an absolute reduction in employment of five employees or more, a relative reduction in employment of 30% or more (mass dismissal), or a termination of business. The displacement rate is equal to the annual number of displacements divided by the average number of persons employed during the course of the same year. Displacement rates lie between 2% and 5% in the long run (averaged over 2000 to 2008). They increased over the Great Recession (2009–2010) but not sufficiently to change this order of magnitude. Displacement rates are thus quite clearly lower than exit rates from employment. For example, table 9.4 indicates that the exit rates for Germany and the United States came to 14% and 22% respectively.

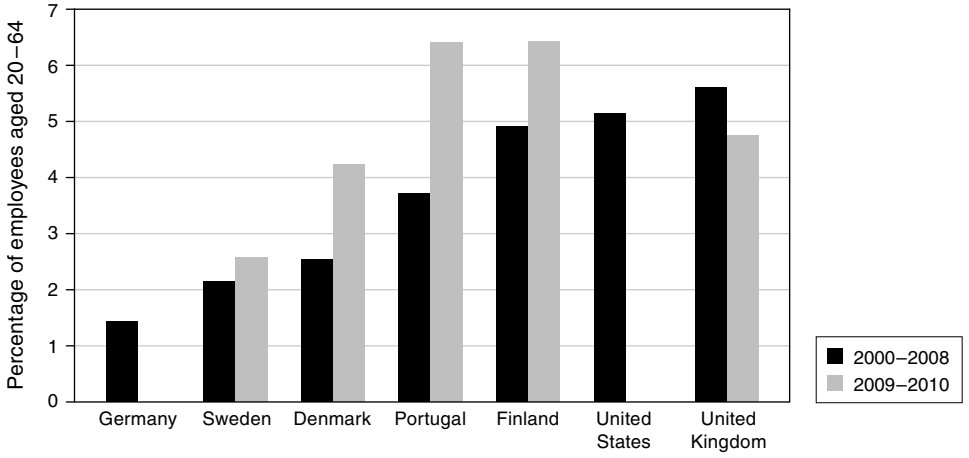


Figure 9.12
Displacement rates in 2000–2008 and 2009–2010.

Note: Percentage of employees aged 20–64 who are displaced from one year to the next, based on firms' declarations.

Source: OECD (2013).

1.3.2 Unemployment Inflows and Outflows

Figure 9.13 displays the rates of entry and exit from unemployment for 14 industrialized countries. The rates of exit from unemployment are much higher than the rates of entry into unemployment, since as a rule the average duration of unemployment is much shorter than the duration of spells in employment. The strong heterogeneity of these rates is striking: the anglophone and Nordic countries post monthly rates of exit from unemployment of more than 20%, while the corresponding rates in the countries of continental Europe are well below 10%. Correspondingly, the monthly rates of entry into unemployment for the anglophone and Nordic countries often surpass 1.5%, whereas those for continental Europe lie in the range between 0.5% and 1%. Japan occupies an intermediate position. These observations suggest that continental European labor markets are sclerotic, to the extent that they display much lower rates of reallocation of labor, as documented by, among others, Elsby et al. (2013). Figure 9.13 also shows that the United States stands apart from other nations. With an average monthly unemployment outflow rate of nearly 60% and an average inflow rate of 3.5%, it exhibits transition rates markedly superior to the transition rates of other countries.

1.3.3 Worker Reallocation Over Time

Separations, layoffs, and quits fluctuate over the course of the business cycle. Figure 9.14 shows that in the United States, layoffs are countercyclical, as firms have a tendency to lay more employees off during recessions, whereas wage earners, who have fewer job opportunities during recessions, reduce their voluntary quits. We note that variations in quits are of the same order of magnitude as variations in layoffs. During the two recessions of the 2000s, the decrease in quits was even greater than the increase in layoffs, with the result that separations as a global category (including quits, layoffs,

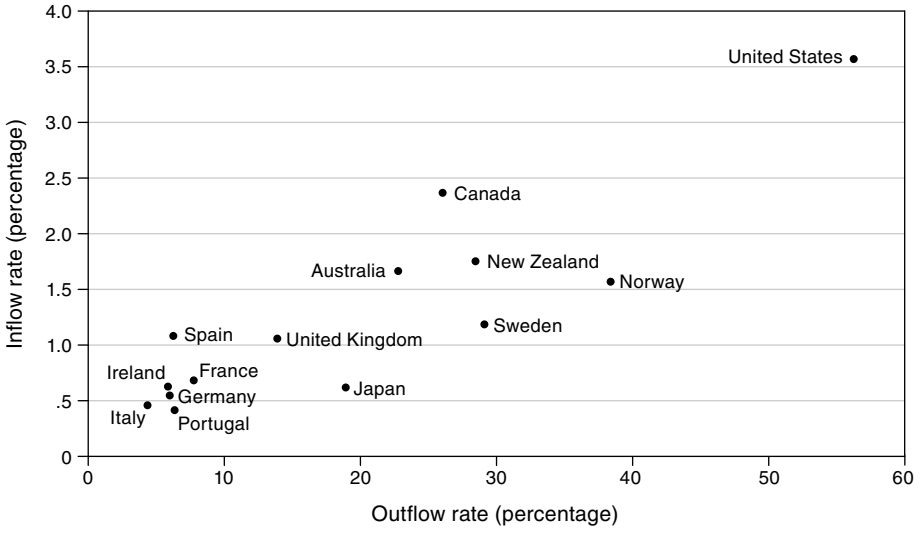


FIGURE 9.13

Unemployment inflow and outflow monthly rates in the OECD countries. The inflow rate is the ratio between monthly entries into unemployment and the total number of employed persons during the month in question; the outflow rate is the ratio between monthly exits from unemployment and the total number of unemployed persons during the month in question. The starting year for the available series varies between 1968 (for the United States) and 1986 (for New Zealand and Portugal). For all countries, the data end in 2009.

Source: Elsby et al. (2013).

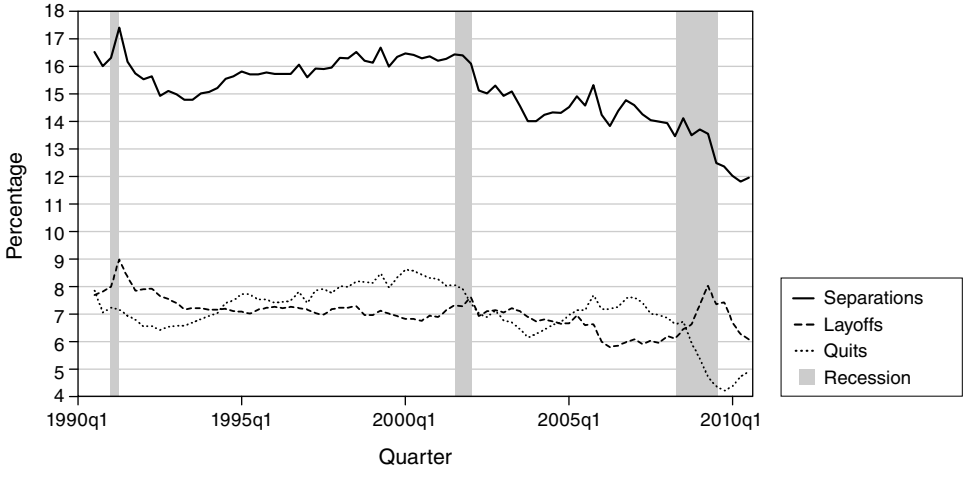


FIGURE 9.14

Quits, layoffs, and separations quarterly rates in the United States. Private sector, 1990q2–2010q2.

Source: Davis et al. (2012) database.

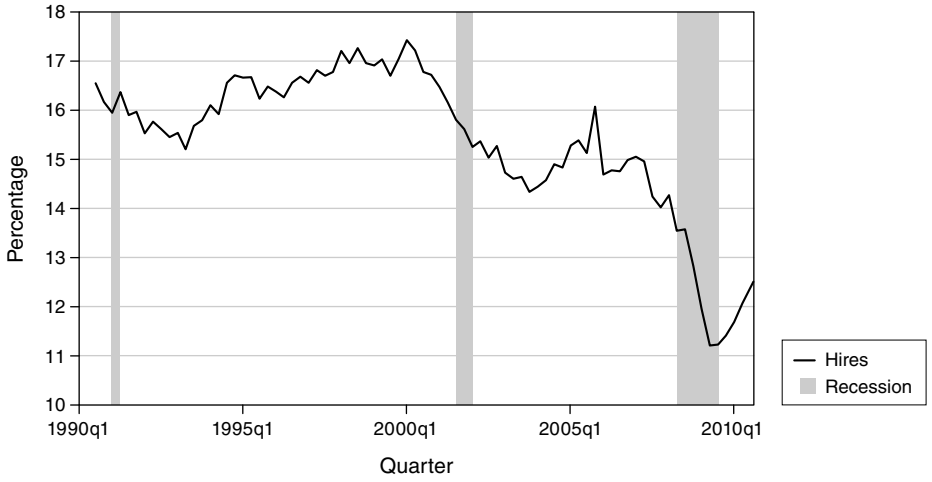


Figure 9.15
Quarterly hiring rates in the United States. Private sector, 1990q1–2010q2.

Source: Davis et al. (2012) database.

and such other exits as retirement, death, and intrafirm tranfers) declined during these two recessions.

The fluctuation in hirings over the course of the cycle in the United States are represented in figure 9.15. Just like job creations, hirings are generally procyclical: they decline during recessions.

Unemployment Dynamics

Variation in worker flows over the course of time determines the dynamics of unemployment and employment. The unemployment rate may rise on both accounts: an increase in entries into unemployment and a decrease in exits from unemployment. Elsby et al. (2013) have assessed the respective contributions of entries and exits from unemployment to overall variation in unemployment rates. They find that variation in the outflow rate explains 85% of overall variation in the unemployment rate for the anglophone countries (and consequently variation in the inflow rate explains just 15% of overall variation). Shimer (2012) arrives at a closely similar result for the United States, utilizing data that cover the period 1948–2010. He finds that the job-finding probability accounted for three quarters of the fluctuation in the unemployment rate, and the employment exit probability for one quarter. For continental European and Nordic countries, Elsby et al. (2013) find a more even split: they estimate that 55% of overall variation in the unemployment rate is accounted for by variation in the outflow rate (and consequently 45% by variation in the inflow rate). Petrongolo and Pissarides (2008) obtain similar results for the European countries they study, the United Kingdom, France, and Spain. The underlying reasons for these differences between the United States and the European countries are still not clearly understood.

Elsby et al. (2013) observe too that during recessions, when the unemployment rate goes up, there is initially an increase in workers flowing into unemployment, rather

than a decline in the number of workers flowing out of it. Correspondingly, the outflows from unemployment increase as the economy recovers. In other words, in all countries, a rise in inflows into unemployment leads the rise in the unemployment rate, whereas a rise in outflows lags the drop in the unemployment rate. (See Rogerson and Shimer, 2011, for a comprehensive survey of the dynamics of entry and exit flows from unemployment and employment, and the linkage between these flows and variations in the unemployment rate.)

1.3.4 The Beveridge Curve

The sheer mass of job flows and worker flows reveals that the labor market is permanently reorganizing itself. At every moment, a large number of jobs are being created and others are being destroyed; at every moment a large number of workers are losing their jobs and others are being hired. These reallocations of jobs and workers give rise to frictional unemployment, or in concrete terms, the simultaneous existence of vacant jobs and individuals seeking work.

The English economist William Beveridge proposed in 1944 to use the relationship between vacant jobs and the level of unemployment to assess the extent of worker reallocation. Problems of reallocation ought to be greater, the higher the number of jobs vacant for a given number of unemployed. The "Beveridge curve" illustrates this linkage between the unemployment rate u and the vacancy rate v (the ratio of the number of vacant jobs to the labor force). It is shown in figure 9.16. When economic activity slows, firms open up few vacant jobs, and there are many unemployed persons searching for a vacancy. During the recovery phase, the point representing equilibrium in the economic system shifts along the Beveridge curve, as more job vacancies are opened up and the number of job seekers falls.

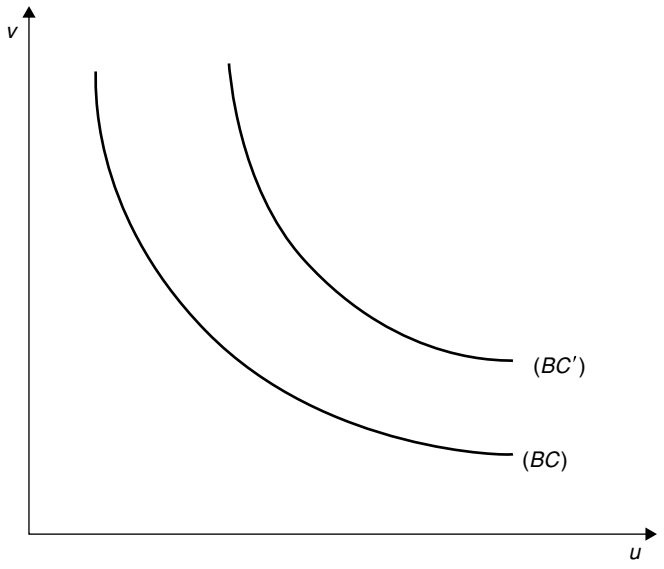


Figure 9.16
The Beveridge curve: The relation between the unemployment rate (u) and the vacancy rate (v).

The very existence of a Beveridge curve signifies that there is a simultaneous presence of unemployed persons and vacant jobs. This simultaneity originates from mobility costs associated with location and skill and from imperfect information. One of the purposes of labor markets is to allow the best possible matchup between the skills required by firms and the skills existing in the labor force. The search activity requires time and resources, but it is indispensable, given that the information necessary to both sides constitutes a rare resource.

The greater or lesser efficiency of the adjustment process is shown by the position of the Beveridge curve with respect to the origin of the axes in figure 9.16. The closer this curve lies to the origin of the axes, the more efficient the process of reallocating manpower is, for in these circumstances every vacant job will quickly be filled by an unemployed person. For example, in figure 9.16, curve (BC) reflects a more efficient process of allocating manpower resources than does curve (BC'). In a labor market described by (BC), for the *same* number of vacant jobs, there will be fewer unemployed persons than there will in the labor market described by (BC').

Differences in the adjustment process as shown by the Beveridge curve have been illustrated by the experiences of the United States and Germany.

The Beveridge Curve in the United States

In the United States, the Bureau of Labor Statistics (BLS) publishes an updated Beveridge curve every month using unemployment rates from the Current Population Survey (CPS) and job opening rates from the Job Openings and Labor Turnover Survey (JOLTS). Figure 9.17 reproduces the Beveridge curve for the United States over the period January 2001 to December 2012.

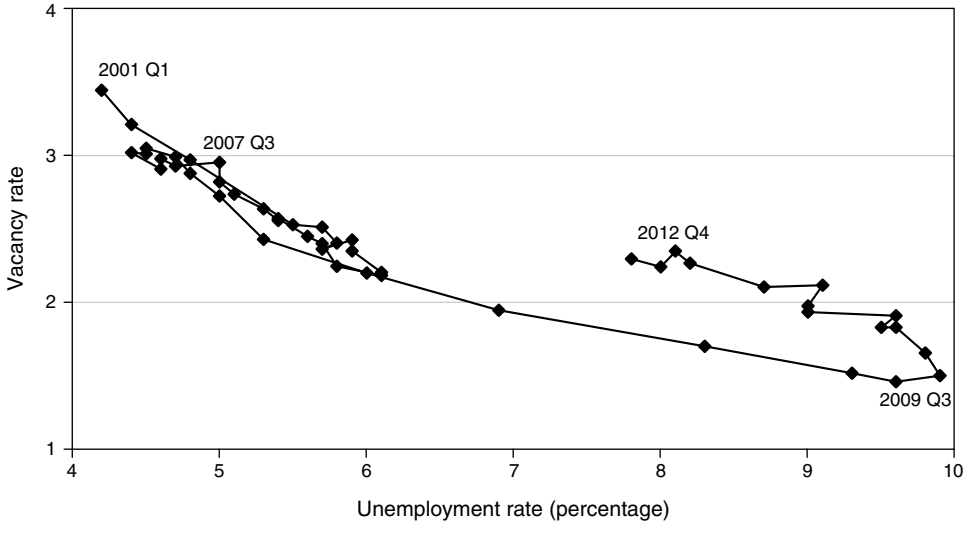


FIGURE 9.17

The Beveridge curve in the United States, 2001–2012. The vacancy rate is defined as the number of job openings divided by the sum of employment and job openings.

Source: Bureau of Labor Statistics data on openings.

We see that during the recession between March 2001 and November 2001, and even more during the Great Recession (December 2007 to June 2009), the unemployment rate rises and the vacancy rate falls. The Great Recession is detectable in a significant shift along the Beveridge curve: between the third quarters of 2007 and 2009, the job opening rate goes from 3% to 1.5%, while the unemployment rate goes from 4.5% to 9.6%. Moreover, we observe that the path followed by the American economy during the recovery phase (from the third quarter of 2009 to the fourth quarter of 2012) is not the mirror image of the path it followed during the Great Recession. During the recovery phase, more vacant jobs than before were required to reach the same unemployment rate. It is thus possible that the Beveridge curve has shifted toward the right since the third quarter of 2009, which would correspond to a shift of the curve (BC) towards the curve (BC') in figure 9.16.

Shifts in the Beveridge curve up and to the right are typically interpreted as structural shifts reflecting a reduced efficiency of the matching of workers to jobs. It is possible that the Great Recession may have entailed reallocation shocks that have permanently reduced the dimensions of certain sectors like banking or construction. It may thus have increased sectoral and occupational mismatch. When the housing bubble burst, one consequence was a fall in housing prices, and this may have increased mismatch by reducing the geographic mobility of unemployed persons who are unable to sell their houses and move because the value of their houses has fallen below that of their mortgage. The Great Recession may also have accelerated ongoing structural changes towards an upskilling of the economy. To the extent that there is a tendency to try to replace low-skilled layoffs by more skilled workers in the recovery, this will increase skill mismatch. This could also explain why the employment situation of low-skilled workers has continued to deteriorate into the recovery (the possible reasons for a shift of the Beveridge curve toward the right for the United States and other countries are analyzed in greater detail in OECD, 2012, chapter 1).

It is also conceivable that the Beveridge curve has not shifted toward the right but rather that the recovery phase is situated on a counterclockwise transitory path, as was the case with previous recessions and recoveries, and that ultimately this transitory path will restore the equilibrium of the economy on the initial Beveridge curve. A transitory dynamic of this kind results from the fact that there are always delays between posting a vacancy and actual hiring. The temporary increase in the duration of unemployment benefit claims during the Great Recession, which has not come to an end as of 2013, may have helped to prolong these delays. Davis et al. (2012) observe, too, that from the start of the recession until 2011 there was a falloff in recruiting intensity (advertising, hiring standards, compensation packages) by employers: this would also contribute to the transitory counterclockwise dynamic around the Beveridge curve. Lazear and Spletzer (2012) estimate that it is too soon to decide between a structural interpretation of the shift of the Beveridge curve and a cyclical interpretation. Such a conclusion will only be possible when unemployment rates return to their pre-recessionary levels, which is not yet the case as these lines are written.

The Beveridge Curve in Germany

The unemployment rate fell sharply in Germany between 2005 and 2012: whereas it stood at 11.4% in the first quarter of 2005, it reached 5.7% in the fourth quarter of 2012. Germany over this period presents a contrasting image to the United States

(and the euro zone as a whole), where the unemployment rate stood higher at the end of 2012 than it did in 2005. Figure 9.18 indicates that over this period the Beveridge curve for Germany shifted toward the left. The fall in the German unemployment rate would thus appear to be linked in part to an improvement in the matching process between job seekers and vacant jobs.

According to Burda and Hunt (2011) the reasons for this German "miracle" are very likely to be sought in the application of the Hartz reforms (so called because they were put in place by chancellor Gerhard Schröder following a report on the functioning of the German labor market written by Peter Hartz, at the time a member of the board of Volkswagen). The Hartz reforms came gradually into effect starting on 1 January 2003. They eased the regulations governing layoffs and stiffened the ones governing access to unemployment benefit. Since 2005 all benefit claimants are obliged to accept jobs offered them, even ones that do not correspond to their qualifications or that are located at a distance from their place of residence. The Hartz reforms also favored resort on a large scale to low-paid and short-term jobs (called mini-jobs and midi-jobs).

Burda and Hunt (2011) also draw attention to the mechanisms of short-time work and "working time accounts." Short-time work is a program that began in 1924 where a firm can get government subsidies to replace two thirds of workers' wages if it cuts the hours of its workers instead of laying them off. It has generally helped smooth employment over cycles in Germany compared to the United States. Working time accounts have spread more recently. An employer can have employees work overtime without an overtime bonus and instead bank the overtime hours in the worker's account. The worker then redeems the banked hours in the form of time off during a slack period. Such mechanisms do exist in certain other OECD countries in different forms but were

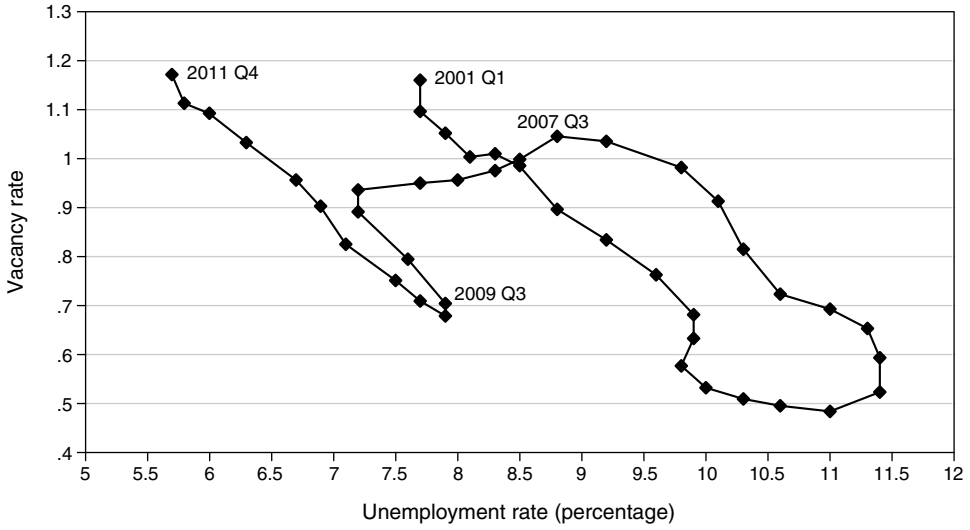


Figure 9.18

The Beveridge curve in Germany, 2001–2011. The vacancy rate is defined as the number of job openings divided by the sum of employment and job openings.

Source: OECD Main Economic Indicators database and national sources.

widely used in Germany during the Great Recession. Burda and Hunt argue that in the recession, it made sense for employers to first draw down the large quantities of hours that had accumulated in the workers' accounts rather than to lay them off, since laid-off workers would have had to be compensated for their banked hours at the overtime premium. As the working time accounts were drawn down, employers increasingly resorted to short-time work, but the delay meant that less short-time work was used than would have been expected based on past recession patterns. If the recession had persisted, the working time accounts would have dwindled to nothing and employers would likely have begun to lay workers off. But the recovery intervened before the accounts were emptied. The combination of short-time work and working time accounts meant that fewer workers were laid off in the 2008–2009 recession than in earlier recessions.

In sum, it is reasonable to suppose that these new labor market rules in Germany shifted the Beveridge curve towards the origin, which means that mismatch probably diminished.

Figure 9.19 shows that other countries appear to have experienced shifts of the Beveridge curve as well. This is notably the case for Sweden (toward the right) and the Netherlands (toward the left). Figure 9.19 also reveals that in some countries no such phenomenon is evident, as in Australia and Spain, although in the latter country the unemployment rate has spiked sharply.

This presentation of the functioning of the labor market reveals intense activity as jobs and workers are reallocated. This is why models that explicitly integrate labor market flows have gradually come to the fore. They are known in the literature as *matching models*. The main question these models have to answer is: what is the relation between unemployment and this reallocation activity? But before examining what they have to tell us, we will do well to review the principal lessons to be learned from the traditional approach to the labor market, based on the competitive model. This review follows.

2 THE COMPETITIVE MODEL WITH LABOR ADJUSTMENT COSTS

An initial approach to understanding the consequences of labor market friction is to introduce adjustment costs of employment into the model of perfect competition. The competitive model, discussed already in chapter 3, is a benchmark representation of the labor market, allowing us to analyze the influence of the turnover of jobs and workers. Here we extend this representation by taking into account the adjustment costs linked to turnover.

2.1 Job Reallocation and Labor Market Equilibrium

In the competitive model, labor supply and demand result from decisions made by agents who have no power over the setting of prices. Hence wages equalize labor supply and demand. Let us assume that the labor force is composed of a large number N of

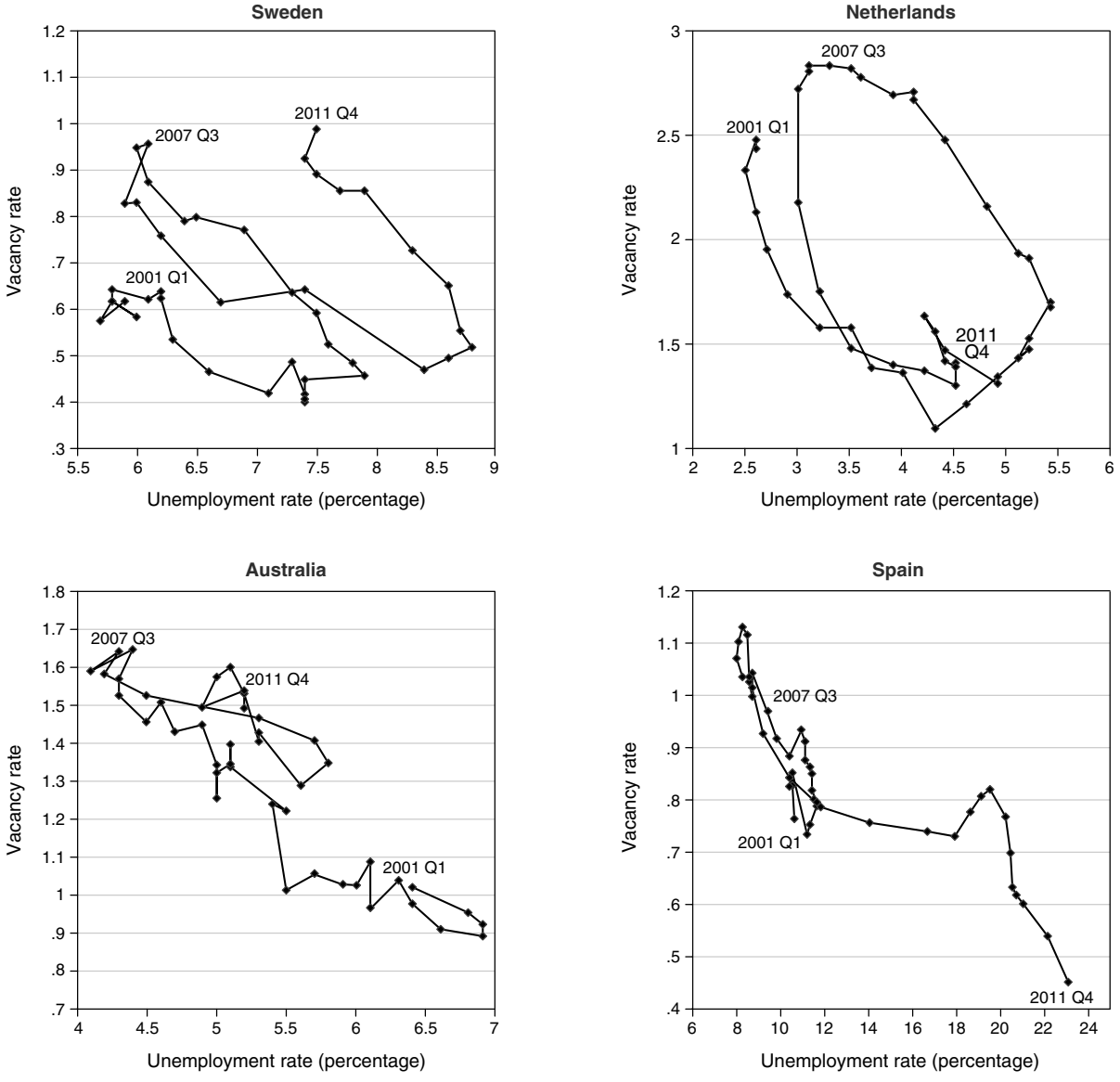


Figure 9.19

The Beveridge curve in Australia, the Netherlands, Spain, and Sweden, 2001–2011. The vacancy rate is defined as the number of job openings divided by the sum of employment and job openings.

Source: OECD Main Economic Indicators database and national sources.

individuals having different reservation wages z, the distribution of which is given by the cumulative distribution function $H(.)$. Readers will recall that in labor supply theory the reservation wage represents the remuneration threshold at which an individual will accept to work (see chapter 1). It can also be interpreted as the domestic production

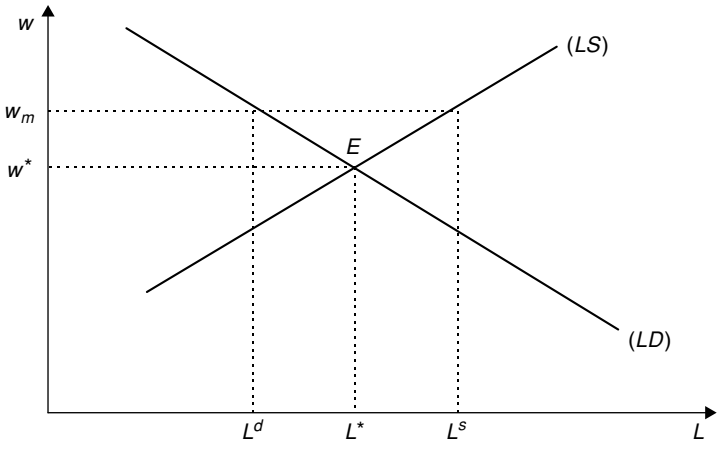


FIGURE 9.20
The competitive equilibrium.

achievable by this person outside the labor market. If we assume that every individual offers a unit of labor when the current wage w is superior to his reservation wage z, then labor supply is equal to $NH(w)$. It is an increasing function of wages, the graph of which is identified by the symbol (LS) in figure 9.20.

In chapter 2, we saw that labor demand could be deduced from profit maximization in the presence of employment adjustment costs. Let us assume, in order to simplify, that the production function of a representative firm has constant returns to scale and that each worker is capable of producing an exogenous quantity y of goods. Let L be the level of employment, and let us suppose that an exogenous proportion q of jobs is destroyed at every instant. As in chapter 2, we represent adjustment costs by a function $C(\Lambda)$ where Λ designates net variations in the level of employment. Function $C(.)$ is assumed to be increasing and convex; consequently $C' > 0$ and $C'' > 0$. In a stationary state, the stock of jobs L is constant and the firm thus hires qL workers per unit of time. Instantaneous profit is then written:

$$\Pi = Ly - [wL + C(qL)]$$

Instantaneous profit maximization[1] with respect to employment entails:

$$y = qC'(qL) + w \tag{9.1}$$

This equality shows that at the firm's optimum, the marginal productivity y of labor is equal to the marginal adjustment cost $qC' + w$ of a job. Equation (9.1) defines labor demand. Adjustment cost $C(.)$ being a convex function, labor demand is decreasing with respect to wages. Its graph is identified by the symbol (LD) in figure 9.20. It should be

[1] We leave out problems related to discounting by implicitly assuming, in order to simplify, that the interest rate is null. We return to these problems in section 4 of this chapter.

noted that a rise in the rate q of job destruction increases marginal adjustment costs $C'(qL)$ and thus increases the total marginal cost of a job. In these circumstances, the firm reduces its demand for labor. In figure 9.20, an increase in q leads to a downward shift of curve (LD). An exogenous rise in adjustment costs $C(.)$ has the same effect. Conversely, an increase in marginal productivity y shifts curve (LD) upward.

The competitive equilibrium lies at the intersection of curves (LS) and (LD). As labor supply is simply equal to $NH(w)$, wages w^* and equilibrium employment L^* are defined by the following system of equations:

$$y = qC'[qNH(w^*)] + w^*, \quad L^* = NH(w^*) \tag{9.2}$$

The hypotheses made about functions $H(.)$ and $C(.)$ entail that there is a unique competitive equilibrium. As an increase in q leads to a downward shift of curve (LD), figure 9.20 also indicates that an increase in the rate q of job destruction leads to a fall in employment and the equilibrium wage. An improvement in individual productivity y has the opposite effect.

It is worth noting that although certain individuals are not employed, there is no unemployment in this model, since every person who wants to work at the current wage can do so. Individuals who are not employed simply prefer to remain outside the labor market and do not look for a job. In sum, the competitive model makes it possible to understand certain determinants of employment. It shows that the process of job destruction is capable of having a negative impact on employment if adjustments in this variable are costly. However, it does not help us in understanding unemployment.

2.2 The Efficiency of the Competitive Equilibrium

As a general rule, a competitive market arrives at an efficient allocation of resources. Within the framework of the model just presented, this result is easily established by considering the problem of a benevolent social planner seeking to maximize collective welfare. For simplicity, we assume, on one hand, that individuals are risk neutral—the indirect utility function is linear—and on the other, that the planner has no preference for the present. In these conditions, his objective is to maximize the sum of instantaneous production realized inside and outside the market, minus labor turnover costs, since these represent a loss for the collectivity.

If we assume that the productivity z of an individual outside the market is again a random variable with cumulative distribution function $H(.)$, the question of the optimal allocation of resources boils down to the search for a threshold $\bar{z}$ of productivity—and thus a proportion $H(\bar{z})$ of individuals that must be employed in the labor market—that makes it possible to maximize net aggregate production. The planner's problem is written as follows:

$$\max_z yNH(z) - C[qNH(z)] + N\int_z^{+\infty} x\mathrm{d}H(x)$$

In this expression, the term in which the integral appears represents total production outside the market, whereas the product $yNH(z)$ designates the production of goods achieved by the market. In the market, the costs due to employment adjustments amount

to $C[qNH(z)]$. The first-order condition entails that the threshold $\bar{z}$ is the solution of equation:

$$y = qC'[qNH(\bar{z})] + \bar{z}$$

This equality defines an optimal value for the productivity threshold identical to the equilibrium wage w^* given by equation (9.2). The competitive equilibrium is thus indeed a social optimum. The planner actually decides to allocate workers to the technology used in the market as long as the marginal productivity, net of turnover costs, of one more individual is greater than what that worker is able to achieve outside the market. This result shows that at the competitive equilibrium, the level of employment is socially optimal, even if some individuals are not employed. It should also be noted that the process of job destruction exerts a negative effect on the stock of jobs in the presence of labor turnover costs but that this process entails *no inefficiency* in the allocation of resources.

2.3 THE LIMITATIONS OF THE COMPETITIVE MODEL

The competitive model displays significant limitations that make it ill-adapted to the study of problems linked to unemployment and the determinants of employment.

1. Most empirical studies show that productivity shocks have much more effect on employment than on wages (see Hall, 1999; Rogerson and Shimer, 2011). Now, the competitive model summed up in figure 9.20 arrives at predictions that contradict this. With a labor supply close to the vertical (which agrees with the small wage elasticity of labor supply found by empirical studies; see chapter 1), a productivity shock affecting labor demand leads to strong variation in wage and weak variation in employment. Many strategies to elaborate competitive models predicting small variations in the wage when the economy undergoes productivity shocks have been tried. The dynamic model of labor supply presented in chapter 1 is one of these strategies. However, these attempts have not yet led to a convincing rehabilitation of the competitive model as a representation of the labor market.

2. The hypothesis of perfect competition does not allow us to explain inefficiencies arising from the functioning of the labor market. The allocation of resources is optimal in this model, which entails particularly the absence of unemployment. As we have seen, the existence of the Beveridge curve illustrates the simultaneous presence of unemployed persons and vacant jobs. This stems from the imperfect information and the mobility costs prevailing in the labor market. Within this framework, unemployed workers adopt job search strategies, and firms adopt recruitment strategies, which may give rise to externalities that are themselves sources of inefficiency in the allocation of resources.

3. The hypothesis of perfect competition also postulates a mode of wage formation that ignores the institutional characteristics of labor markets. In chapters 6 and 7 we emphasized that wage bargaining and manpower management policies have a preponderant influence on levels of remuneration. Here again, the strategic dimensions of behavior can have consequences very different from those we find in the

competitive model, in which wages are determined by an abstract process that is assumed to equalize supply and demand.

Thus, in the presence of imperfect information, and when wages do not clear markets, it is highly likely that the labor market will operate inefficiently. That makes it important to have at our disposal an analytical tool that does not postulate the absence of inefficiency a priori, a tool enabling us to identify, understand, and if necessary define remedies for these inefficiencies. To furnish a representation of the labor market possessing these qualities has been the aim of a number of studies. Of these, the matching model proposed and developed by Pissarides (2000) is, at the present time, the analytic framework most often used (see also Mortensen and Pissarides, 1999).

3 THE MATCHING MODEL

We now develop a simple model of the labor market in which transaction costs explain the simultaneous existence of vacant jobs and unemployed persons. Wage formation is here described by a bargaining process between employers and workers; in other words, the hypothesis of competitive wages is dropped. The model is structured around the concept of "matching function," which sums up, at the aggregate level, the outcomes of numerous encounters between persons in search of a job and firms with positions vacant.

At every instant, the number of hires depends on the interface between vacant jobs and workers looking for a job. For given levels of supply and demand, and when workers are perfectly suited to the jobs offered and there is no imperfection in the available information, the number of hires is equal to the minimum of job seekers and job vacancies, and the labor market functions efficiently. But in reality jobs and workers are heterogeneous, and information never circulates perfectly. Hence some workers risk not finding work at the same time that some firms have positions vacant. The existence of these transaction costs in the labor market is usually represented by a matching function that determines the number of hires on the basis of the quantity of labor being supplied and demanded. This matching function and the equilibrium conditions of flows in the labor market make it possible to give an analytical foundation for the Beveridge curve.

3.1 The Matching Function and the Beveridge Curve

In practice, job search procedures are characterized by a large number of "frictions." The most important of these have to do with the mismatch between certain vacant jobs and the skills of workers, as well as ignorance of the whereabouts and/or the actual characteristics of the jobs available. Faced with these frictions, employers and job seekers adopt search strategies that include reading newspapers, browsing the Web, applying to government employment offices, using personal networks, sending letters of application, and so on. All these actions take time and often have high costs. But at every instant they produce a certain number of "successes," which can be measured by the number of hires at the date in question. The *matching function* goes straight to an *aggregate* level (for example, a country, region, or industry) and does not take into account

the diversity of individual actions. It summarizes the entire search process in a single relation giving the flow M of hires achieved over a given interval of time as a function of the stock of vacant jobs V and persons in search of work D. The matching function is analogous to other aggregate functions used by macroeconomists, like the aggregate production function. For it to be a useful instrument, we have to be able to give it extremely precise properties that rest, if possible, on microeconomic foundations, and above all, we need to verify that the empirical estimates of such a function are coherent with these properties.

3.1.1 Microeconomic Foundations

A simple, but not truly realistic, way of obtaining an aggregate matching function consists of assimilating vacant jobs to "urns," and job applications to "balls" tossed at the urns by job seekers (Hall, 1979b; Pissarides, 1979; Blanchard and Diamond, 1994). A match occurs when a ball goes into an urn. The inefficiency of the job search process is reflected in the greater or lesser precision with which the balls are tossed in the direction of the urns. We omit the time index for simplicity, and D and V will again denote respectively the number of job seekers and the number of vacant jobs at a given date. Let us assume that job seekers know the locations of all vacant jobs and that a particular job seeker, whom we call Mr. i, simultaneously sends e_i applications out randomly among the V jobs vacant. Parameter $e_i \leq V$ is an indicator of the effort which Mr. i puts into his job search. When more than one application is received for the same vacant job, a random draw determines who will get it, and the other applications go into the wastepaper basket. Let us further suppose that there is no coordination among the job seekers. That being so, it is possible that one vacant job will receive a heap of applications, while another will not receive any. More precisely, the probability that a given vacant job will receive the application from Mr. i is equal to e_i/V. Conversely, the probability that this job will not receive an application from Mr. i amounts to $1 - (e_i/V)$. It results that the probability of a vacant job receiving no applications takes the value $\prod_{i=1}^{i=D} [1 - (e_i/V)]$. In consequence, the probability of a vacant job receiving at least one application is equal to $1 - \prod_{i=1}^{i=D} [1 - (e_i/V)]$. As we have assumed that, for each vacant job, the firms draw the successful applicant at random from among the applications received, the number of hires M is given by relation:

$$M = V \left[1 - \prod_{i=1}^{i=D} \left(1 - \frac{e_i}{V} \right) \right]$$

If V is large with respect to e_i (which is a reasonable hypothesis), it is possible to approximate $1 - (e_i/V)$ by $\exp[-(e_i/V)]$. Let $\bar{e}$ be the average of the e_i; the matching function is finally written:

$$M = M(V, \bar{e}D) = V \left\{ 1 - \exp \left[- \left(\frac{\bar{e}D}{V} \right) \right] \right\}$$

It can be verified that this function is increasing in V and D and that it is homogeneous of degree 1 with respect to its two arguments. The value $\bar{e}$ of the average search intensity also appears among the arguments of the matching function. That justifies the

inclusion, in the estimates of the matching function, of all the variables that may affect job search effort, like the characteristics of the unemployment insurance system, the demographic profile of job seekers, indicators of the ease of geographical mobility, and so on. Note further that, the total number of applications being equal to $\bar{e}D$, the probability of Mr. i finding a job is written $e_i M(V, \bar{e}D)/\bar{e}D$. He thus has a better chance, the greater his level of relative effort $e_i/\bar{e}$.

Simple "urns and balls" models thus give us the foundations of the aggregate matching function. These models have their uses for analyzing the search strategies of firms and workers. Calvo-Armengol and Zénou (2005) and Cahuc and Fontaine (2009) have used urns and balls models to analyze the influence of the networked personal relationships, within which wage earners and unemployed persons can transmit information, on the efficiency of job search.

Other models have been adopted in order to take into account strategic, nonrandom elements which can play a role in the job search, on the part of both workers and firms.

Ranking models, like that of Blanchard and Diamond (1994), start from the hypothesis that firms have preferences among the applications they receive. They will, for example, prefer skilled employees to unskilled ones, or short-term unemployed persons to long-term ones. That being the case, the matching function depends, directly or indirectly, on the preferences of employers and the characteristics of job seekers. So, if firms give priority to the short-term unemployed, it can be shown that the average probability of finding a job diminishes with the incidence of long-term unemployment. This result has been confirmed by the work of Mumford and Smith (1999) for Australia and that of Burgess (1993) for the United Kingdom. Petrongolo and Pissarides (2001) do point out, though, that a result of this type does not necessarily reinforce the hypothesis that applicants are ranked. It might also be caused by reduced search effort on the part of the long-term unemployed.

Stock-flow matching models begin with the idea that the existence of stocks of vacant jobs and unemployed persons reflects, to some degree at least, an inadequate fit between the characteristics of vacant jobs and those of job seekers which is already perfectly well *known* and does not need to be discovered. From that it follows that the job search process, on the part of both firms and workers, will privilege new inflows of applications over stocks already examined. Coles and Smith (1998) construct a model of this type, which they estimate using British data for 1987–1995. The empirical results partially corroborate their hypotheses. They find that only new flows of vacant jobs significantly increase the hazard rates of the long-term unemployed, while the hazard rates for the short-term unemployed are positively affected both by stocks of vacant jobs and by new flows.

3.1.2 The Properties of the Matching Function

With no loss of generality, we will simply denote the aggregate matching function by $M(V, D)$. In a model in continuous time such as the one we will use throughout the rest of this book, $M(V, D)$ represents the *instantaneous* flow of hires at a given date. In other words, if V_t and D_t designate respectively the stock of vacant jobs and the stock of persons looking for work at date t, the number of hires over interval $[t, t + \mathrm{d}t]$ is equal to $M(V_t, D_t)\mathrm{d}t$. In order to simplify the notation, we generally omit the time

index. Function $M(V,D)$ will be assumed to be strictly increasing with respect to each of its arguments and such that $M(V,0) = M(0,D) = 0$. These hypotheses signify, on one hand, that hires increase when the number of job applicants, or the number of vacant jobs, increases, and on the other, that no hire can occur without at least one vacant job and one job applicant. A frequently used formulation of the matching function adds two supplementary hypotheses (Pissarides, 2000). First, only unemployed persons are assumed to be job applicants. If U designates the number of unemployed persons, then we will have $U = D$. This hypothesis amounts to setting aside the job search activities of wage earners who are already employed (see Mortensen, 1994, and Pissarides, 2000, who present models that include this possibility). Finally, we assume that the matching function has constant returns to scale. The probability of filling a vacant job per unit of time is then expressed as follows:

$$\frac{M(V,U)}{V} = M(1,U/V) \equiv m(\theta);\ \theta \equiv V/U \tag{9.3}$$

Parameter θ, which equals the ratio of the number of vacant jobs to the number of unemployed persons, is an indicator of the "tightness" prevailing in the labor market. Differentiating the expression (9.3) with respect to U, we get:

$$m'(\theta) = -\frac{U^2}{V^2} M'_U(1,U/V) < 0$$

Hence vacant jobs are filled at a rate that diminishes with the labor market tightness. The reason for this is as follows: for a given number U of unemployed persons, each firm has greater difficulty in filling its vacant positions when the total number of vacant jobs rises. For an unemployed person, the exit rate from unemployment—also called the *hazard rate*; see chapter 5—also depends on the labor market tightness; it is defined by:

$$\frac{M(V,U)}{U} = \frac{V}{U}\frac{M(V,U)}{V} = \theta m(\theta) \tag{9.4}$$

Differentiating this relation with respect to V, we find:

$$[\theta m(\theta)]' \equiv m(\theta) + \theta m'(\theta) = M_V(V,U) > 0$$

In consequence, the exit rate from unemployment is an increasing function of the labor market tightness. That means that for a given number of unemployed persons, each of them has a greater chance of finding a job when the number of vacant jobs increases. It can also be verified that the absolute value of the elasticity of function $m(\theta)$, $\eta(\theta) = -\theta m'(\theta)/m(\theta)$, is inferior to unity.

Scrutiny of the exit rates from unemployment and employment shows that there are *trading externalities*. The increase in the number of vacant jobs diminishes the rate at which vacant jobs are filled and increases the exit rate from unemployment. So it is in the interest of unemployed persons for firms to create jobs, but in the interest of each firm for the number of vacancies to be as low as possible, so as to have the benefit of numerous applications for the jobs it needs to fill. It is also in the interest of each

unemployed individual for other job seekers to withdraw from the labor market, so as to reduce the competition. *Between-group* externalities are positive, therefore, but *within-group* externalities are negative, corresponding to congestion effects.

3.1.3 Some Empirical Elements

The matching function can be estimated on the basis of time series data. It is generally assumed that it takes a Cobb-Douglas form with constant returns such that the number of matches at date t, denoted M_t, is equal to:

$$M_t = A_t V_t^{1-\eta} U_t^{\eta}$$

Parameter A_t represents the efficiency of the matching process at period t. The dependent variable is represented by the number of hires during period t, and the explanatory variables are the stocks of unemployed persons (U_t) and vacant jobs (V_t). This equation may be written by dividing each side by U_t and by taking the logarithm:

$$f_t = (1 - \eta) \ln \theta_t + a_t \tag{9.5}$$

where f_t is the logarithm of the exit rate from unemployment into employment, $\theta_t = V_t/U_t$ represents labor market tightness, and a_t designates the logarithm of the efficiency parameter. This efficiency parameter may be written in the form $a_t = a + \varepsilon_t$ so as to decompose it into a constant a, corresponding to its average value over the ensemble of periods covered by the estimation, and a residual term, ε_t, corresponding to variations around the average of the efficiency of the matching process proper to each period.

Most often the estimation uses simple OLS regression procedures (Petrongolo and Pissarides, 2001). With a few notable exceptions, like Blanchard and Diamond (1990) on data from the manufacturing sector in the United States and Yashiv (2000) on Israeli data, most empirical studies based on macroeconomic data accept the hypothesis of constant returns. If the flows of hires are hires coming from unemployment only, the elasticity of the matching function with respect to the stock U of unemployed persons lies in the range [0.5, 0.7]. For instance, using a simple OLS regression based on monthly data covering the period 2001–2009 for the United States, Rogerson and Shimer (2011) find an elasticity equal to 0.58. But if the dependent variable comprises all hires (taking in persons who move from one job to another and hires of nonparticipants), this elasticity lies in the range [0.3, 0.4]. Barnichon and Figura (2011) have used this approach to estimate the path of the efficiency of the matching process in the United States over the period 1968–2010. They find that this efficiency degraded significantly during the Great Recession of 2008–2009. They also find that the composition of the unemployment pool explains an important part of the efficiency of the matching process. This result is compatible with the analysis of the microeconomic foundations of the aggregate matching function, which suggests that all the elements that might have an influence on job search activity ought to be included among the explanatory variables. Empirical studies do indeed add variables of this type to the list of exogenous factors. It turns out that the incidence of long-term unemployment, the geographical dispersion of vacant jobs and unemployed persons, and the demographic structure of the labor force all exert significant influence on the matching process.

Borowczyk-Martins et al. (2013) have noted however that estimations based on simple OLS regressions are probably biased, since the decision to post a job vacancy is not independent of the efficiency of the matching process. Firms may have an incentive to create more vacant jobs when the efficiency of the matching process improves, for it means they can hire more rapidly. The equilibrium value of the variable θ_t is thus not independent of the value of parameter a_t and the hypothesis that the covariance between labor market tightness θ_t and the residual term ε_t in equation (9.5) is null, necessary to obtain an unbiased OLS estimator of parameter η, is highly likely to remain unsatisfied.

3.1.4 Equilibrium of Flows and the Beveridge Curve

Labor market tightness and the rate of job destruction, along with the matching technology, condition the dynamics of flows of jobs and workers. To show this, we designate the stock of unemployed persons by U, employment by L, and the size of the labor force at a given date by N. At every instant, the labor force grows by quantity $\dot{N}$. Assuming that all the new entrants into the labor force begin by looking for a job, the number of unemployed persons is increased by the total of these new entrants, to whom must be added the qL workers who have just lost their jobs. Unemployment thus increases by $\dot{N} + qL$. Conversely, at every instant there are $\theta m(\theta)U$ unemployed persons who find a job. The variation $\dot{U}$ in the stock of unemployed persons is then written:

$$\dot{U} = \dot{N} + qL - \theta m(\theta)U \tag{9.6}$$

Let $n = \dot{N}/N$ be the rate of growth of the labor force, and $u = U/N$ the rate of unemployment. As we have $N = L + U$ and also $\dot{U} = \dot{u}N + u\dot{N}$, the law of motion of the rate of unemployment is found by dividing the two sides of relation (9.6) by N. The result is:

$$\dot{u} = q + n - [q + n + \theta m(\theta)]\, u \tag{9.7}$$

The stationary value of the unemployment rate, the only thing that interests us here, corresponds to $\dot{u} = 0$. It is thus given by:

$$u = \frac{q + n}{q + n + \theta m(\theta)} \tag{9.8}$$

If we define the vacancy rate by $v = V/N$, the labor market tightness θ is also equal to the ratio v/u. Equation (9.8) then describes a relationship between the unemployment rate u and the vacancy rate v. This linkage expresses the equilibrium of worker flows between employment and unemployment, given the properties of the matching function. In the plane (v, u), this relationship yields the Beveridge curve. It is possible to show, using the hypotheses made about the matching function, that the Beveridge curve is decreasing and convex, as shown by curve (BC) in figure 9.16. Moreover, the position of the Beveridge curve reflects the efficiency of the matching technology, for this curve lies farther out from the origin, the more inefficient this technology is.

We will, in what follows, develop a model of labor market equilibrium based on the matching process just described and will confine ourselves to the stationary state

(the dynamics are presented in section 6). We begin by studying the behaviors that firms and workers will adopt when faced with the matching process.

3.2 The Behavior of Firms and Workers

There are only two goods in the economy: a good produced by the firms and consumed by all individuals; and labor, assumed to be homogeneous, which is the sole factor of production. The good produced by the firms is the numeraire. In the standard version of the matching model, each firm has one job that can be either vacant or filled; when this job is filled, it enables the production of an exogenous quantity y of the good per unit of time. Then, we revert to the traditional representation of the firm using a production function and bring in capital as another input. This more general model does not produce markedly different conclusions, but it does supply the foundations of the simplified model we use here, and it allows us to specify the impact of variations in the cost of capital on investment and employment. We begin by defining the expected profit from a job in order to determine the labor demand of firms.

3.2.1 Firms

At every instant, a job can be either filled or vacant. When it is filled, it yields an expected profit Π_e which is different from the profit expected Π_v when the job falls vacant.

The Profit Expected from a Filled Job

In each small interval of time $\mathrm{d}t$, a filled job is liable to fall vacant with an exogenous probability $q\mathrm{d}t$. This probability covers all exits from employment, whether their cause is layoffs, the destruction of jobs, or whatever. It must be remembered, though, that letting an employee go and destroying a job are by nature endogenous decisions, made on the basis of an analysis of the present and future prospects of the firm. So to choose an exogenous probability q to describe these phenomena is not a satisfactory solution. Chapters 10 and 13 will show how it is possible to make this probability endogenous (see also Mortensen and Pissarides, 1994, and Pissarides, 2000). A large number of results (but not all) still stand with the hypothesis of an exogenous probability of exiting from employment.

We also assume that the real interest rate r is exogenous. Implicitly, then, we place ourselves in the framework of a small open economy with perfect mobility of financial assets. The existence of a financial market entails that a dollar invested at date t brings in $1 + r\mathrm{d}t$ dollars in $t + \mathrm{d}t$, or, in other words, that the discounted value of a dollar at date t which will be available at date $t + \mathrm{d}t$ is $1/(1 + r\mathrm{d}t)$. So the term $1/(1 + r\mathrm{d}t)$ represents the discount factor for each small interval of time $\mathrm{d}t$. In the stationary state, if we denote by w the real wage received at every instant by an employee, the profit expected from a filled job takes this form:

$$\Pi_e = \frac{1}{1 + r\mathrm{d}t}\left[(y - w)\mathrm{d}t + q\mathrm{d}t\Pi_v + (1 - q\mathrm{d}t)\Pi_e\right] \tag{9.9}$$

This relation indicates that the expected profit from a job is equal to the discounted sum of the flow of instantaneous profit $(y - w)\mathrm{d}t$ in the interval of time $\mathrm{d}t$

and of the discounted expected future profits. With a probability $q\mathrm{d}t$ these future profits coincide with the expected profit Π_v from a vacant job, and with the complementary probability $(1 - q\mathrm{d}t)$ they coincide with the expected profit Π_e from a filled job. It is particularly interesting to note that relation (9.9) can be rewritten in simpler form:

$$r\Pi_e = y - w + q(\Pi_v - \Pi_e) \tag{9.10}$$

It is worth noting that this equation portrays the equality of the returns of different assets in a perfect financial market. In the present case, an asset worth Π_e invested in the financial market brings in $r\Pi_e$ at every instant. This same asset invested in the labor market offers an instantaneous profit $(y - w)$, to which is added the average gain $q(\Pi_v - \Pi_e)$ associated with the job possibly changing state. For a filled job, this gain is in fact a loss resulting from the employee's leaving. Several times before—see chapter 5 in particular—we have encountered formulas analogous to relation (9.10). Mathematical appendix D at the end of the book supplies a rigorous proof of these formulas, showing that they do indeed correspond to the stationary state of a model in which a particular event (here, the destruction of jobs) follows a Poisson process.

The Profit Expected from a Vacant Job

The costs of a vacant job per unit of time are denoted h. These costs represent the expenses incurred in holding the position open and looking for an employee with the right skills to fill it (advertising, agency fees, the services of a consultant, etc.). Since vacant jobs are filled at rate $m(\theta)$, the profit expected from a vacant job is written:

$$\Pi_v = \frac{1}{1 + r\mathrm{d}t}\{-h\mathrm{d}t + m(\theta)\mathrm{d}t\Pi_e + [1 - m(\theta)\mathrm{d}t]\,\Pi_v\}$$

Or again, rearranging the terms of this relation:

$$r\Pi_v = -h + m(\theta)(\Pi_e - \Pi_v) \tag{9.11}$$

This relation equates the instantaneous return $r\Pi_v$ of the "unfilled job" asset in the financial market to its return in the labor market. Its return in the labor market comprises the instantaneous cost $-h$ and the average gain $m(\theta)(\Pi_e - \Pi_v)$ associated with a change of state (in this case, the passage from the vacant state to the filled state).

Labor Demand

As long as the profit expected from a vacant job remains strictly positive, new entrepreneurs enter the market to create jobs. This inflow ends when the profit expected from a vacant job goes to zero. We thus have the *free entry* condition; it is written simply $\Pi_v = 0$. When this condition is satisfied, relation (9.11) then entails $\Pi_e = h/m(\theta)$. On the other hand, equation (9.10) defining the profit expected from a filled job also gives $\Pi_e = (y - w)/(r + q)$. Equalizing these two values of Π_e we arrive at the following equation:

$$\frac{h}{m(\theta)} = \frac{y - w}{r + q} \tag{9.12}$$

The left side of this equation represents the average cost of a vacant job. At every instant a vacant job brings an expense equal to h and is filled at rate $m(\theta)$. We know[2] that, on average, this vacant job remains unfilled for an interval of time $1/m(\theta)$. So the average cost of a vacant job is indeed equal to quantity $h/m(\theta)$. Recalling that the right side of relation (9.12) is equated to the profit expected from a filled job, the interpretation of this relation becomes very simple: at free entry equilibrium, the average cost of a vacant job must be equal to the profit expected from a filled job.

Since the rate $m(\theta)$ at which vacant jobs are filled decreases with labor market tightness θ, equation (9.12) defines a decreasing relation between the wage and the labor market tightness. This negative relation is analogous to *labor demand* in the neoclassical theory of the firm (see chapter 2). It reveals the fact that a rise in the wage w degrades the profit outlook of a filled job. Since at free entry equilibrium the expected profit of a filled job equals the average cost of a vacant job, entrepreneurs react to a decrease in the expected profit of filled jobs by creating fewer vacant jobs, which lowers the expected duration and then the expected cost of vacant jobs.

Since we have shown that the unemployment rate can be deduced from labor market tightness using the Beveridge curve (9.8), it is possible to define the equilibrium values of the unemployment rate u and of labor market tightness θ using the system of equations (9.8) and (9.12) when wages are exogenous. Readers are invited to perform this exercise for themselves.

In matching models, wages are usually bargained over between each employer and each employee. This is a very natural approach, for as relation (9.12) shows, the fact that there is a cost attached to creating jobs induces a strictly positive profit for employers with filled jobs. A strictly positive profit from filled jobs is indeed required if employers are to have an interest in posting vacant slots. In these circumstances, part of the profit will flow to the employees if they have bargaining power. To grasp the way a labor market with transaction costs functions, it is therefore important to represent the process of sharing the gains produced by filled jobs, and analyze its influence. For that, it is necessary first to specify the way in which workers derive benefit from being employees and from being unemployed.

3.2.2 Workers

The labor force is composed of N individuals, whose lifespan is infinite. Any worker can be either employed, with an expected utility V_e, or unemployed, with an expected utility $V_u \leq V_e$. When a worker is employed, she produces a quantity y and gets a real wage w per unit of time. She also risks losing her job at the rate of q. Assuming that workers are risk neutral (which amounts to assuming that the indirect instantaneous utility function is linear), the expected utility of an employee at stationary equilibrium is found by repeating the procedure used to calculate the value of a job, so that:

$$rV_e = w + q(V_u - V_e) \tag{9.13}$$

[2] If a variable can change state at rate p, it will, on average, remain in the state it is in at the present moment for an interval of time equal to $1/p$ (see mathematical appendix D at the end of this book, which is dedicated to the properties of Poisson processes).

An unemployed worker is always in search of a job. At each instant, this search procures her a net gain denoted z. We saw in chapter 5, in studying the theory of job search, that this net gain comprises benefits linked to being unemployed (unemployment insurance, social welfare transfers, and whatever utility comes from not having to leave home to go to work) minus the various costs attached to searching for a job (transportation, postage, perhaps extra training, etc.). Since the exit rate from unemployment is $\theta m(\theta)$, the expected utility of an unemployed person satisfies:

$$rV_u = z + \theta m(\theta)(V_e - V_u) \tag{9.14}$$

3.3 Wage Bargaining

When a worker and a vacant job come together, the employer and the potential employee bargain over the wage. Theory suggests that this bargaining yields a wage which increases with labor market tightness. Empirical studies confirm the existence of a relation of this type.

3.3.1 Surplus Sharing

Under suitable assumptions, the wage bargaining outcome is a simple surplus sharing rule, a rule for the sharing of the surplus yielded by a filled job between employer and employee. Moreover, it turns out that very simple noncooperative games make it possible to explain this sharing rule.

Surplus and the Nash Criterion

In dealing with the problem of bargaining, it is often helpful to work with the *surplus* S that derives from the match between an employee and an employer. This surplus is defined by the sum of the *rents* that a filled job paying negotiated wage w procures. Rent represents the difference between what individuals obtain through the contractual relationship and what the best opportunity outside the contract would bring them (see chapters 6 and 7). In the present context, for the employee the rent amounts to $(V_e - V_u)$, while for the employer it is equal to $(\Pi_e - \Pi_v)$. The surplus is thus defined by:

$$S = V_e - V_u + \Pi_e - \Pi_v \tag{9.15}$$

Bargaining gives each participant a share of the surplus proportional to his relative power. Let $\gamma \in [0, 1]$ be the relative power of the worker; the result of the negotiation is written:

$$V_e - V_u = \gamma S \quad \text{and} \quad \Pi_e - \Pi_v = (1 - \gamma)S \tag{9.16}$$

There are several ways to explain such a division of the surplus. In chapter 7, we learned that the outcome of bargaining between two players could, under certain conditions, equal the maximum of the generalized Nash criterion. In this case, the value of the wage negotiated at each date is the solution of the following problem:

$$\max_w (V_e - V_u)^\gamma (\Pi_e - \Pi_v)^{1-\gamma} \tag{9.17}$$

Using equations (9.10) and (9.13), which define respectively the expected gain of an employee and an entrepreneur, we can easily verify that the first-order condition of this problem gives the sharing rule (9.16).

A Bargaining Game

We can also explain the surplus sharing rule (9.16) with the help of a noncooperative bargaining game. Let us assume, for example, that the bargaining unfolds, at each instant, as a two-stage game with the following characteristics:

Stage 1: The two players propose a contract that stipulates a wage to be paid in the future small interval of time dt.

Stage 2: If one of the two players has refused to sign the contract proposed in stage 1, the worker makes a new, take-it-or-leave-it offer with probability γ, and the employer in turn makes an offer of the same kind, with the complementary probability $(1 - \gamma)$. If there is again no agreement, the job is destroyed.

It is not hard to show that the surplus sharing rule (9.16) emerges as the subgame perfect equilibrium in this bargaining game (see chapter 7 for a definition of this equilibrium). If it is the worker who makes the offer in stage 2, the employer obtains a gain of Π_v, and the worker takes the whole surplus, which means that her expected utility amounts to $(S + V_u)$. If, on the other hand, it is the employer who makes the second-stage offer, the worker obtains V_u, the employer takes the whole surplus, and his expected profit amounts to $(S + \Pi_v)$. So in the first stage, the worker knows that at the outcome of stage 2, her expected utility will amount to $(1 - \gamma)V_u + \gamma(S + V_u)$, which is equal to $V_u + \gamma S$. Symmetrically, the employer knows that his expected profit will be equal to $(1 - \gamma)(S + \Pi_v) + \gamma\Pi_v$, which amounts to $\Pi_v + (1 - \gamma)S$. In consequence, it makes no difference to either player whether they sign a contract at stage 1 stipulating an expected utility V_e equal to $V_u + \gamma S$ for the employee, and an expected profit Π_e equal to $\Pi_v + (1 - \gamma)S$ for the employer, or wait until stage 2 to make the offers already defined. In the first stage, then, to sign a contract conforming to sharing rule (9.16) constitutes a subgame perfect equilibrium of the bargaining game. If we assume that there is a cost attached to going to stage 2, even a small cost, the bargaining game possesses a single equilibrium, corresponding to the immediate agreement of a surplus sharing contract as described by condition (9.16).

To this point, we have set out a very simple and excessively artificial game that leads to the surplus sharing rule usually adopted in matching models. Actually, it is possible to construct a large number of bargaining games that all lead to this sharing rule. These different games yield different interpretations of parameter γ, which can, in particular, depend on the preference of the players for the present and their degree of risk aversion (see chapter 7 for fuller exposition of this type of problem, and the work of Osborne and Rubinstein, 1990). At present, we concentrate on the consequences of the surplus sharing rule.

The Negotiated Wage

In the first place, we get a simple expression of the surplus by adding up relations (9.10) and (9.13), which define respectively the expected utility and profit associated with a

filled job for which the wage negotiated amounts to w. We thus have:

$$S = \frac{y - r(V_u + \Pi_v)}{r + q} \tag{9.18}$$

Moreover, definitions (9.10) and (9.13) of the profit and utility expected from a filled job can be written as follows:

$$V_e - V_u = \frac{w - rV_u}{r + q} \quad \text{and} \quad \Pi_e - \Pi_v = \frac{y - w - r\Pi_v}{r + q} \tag{9.19}$$

Combining the two first equalities of relations (9.16) and (9.19) with the expression (9.18) of the surplus taken at free entry equilibrium, where $\Pi_v = 0$, we arrive at a formula characterizing the negotiated wage. It is written:

$$w = rV_u + \gamma(y - rV_u) \tag{9.20}$$

This expression has a very intuitive interpretation. When the employee has all the bargaining power ($\gamma = 1$), then she garners all of production y at every date. If, on the contrary, the employer possesses all the bargaining power ($\gamma = 0$), the wage w is then equal to rV_u and relation (9.19) shows that $V_e = V_u$; the employee then obtains no rent. In the intermediate cases, ($0 < \gamma < 1$), the wage negotiated is a convex combination of the value y of the production and of the reservation wage, rV_u, weighted by the respective power of the employee and the employer.

3.3.2 The Wage Curve

The wage curve synthesizes the linkages between the wage and the labor market tightness as they emerge out of the bargaining process. Estimates of numerous wage equations allow us to specify the properties of this curve.

Wage Curve and Labor Supply

It is possible to obtain a relationship between the wage w and the tightness θ of the labor market using equation (9.20), which gives us the value of the negotiated wage. To that end, it is enough to note that definition (9.14) of V_u and surplus sharing rule (9.16) entail $rV_u = z + \gamma\theta m(\theta)S$, and, taking into account form (9.18) of the value of surplus S at free entry equilibrium, we arrive at:

$$rV_u = \frac{z(r + q) + \gamma y\theta m(\theta)}{r + q + \gamma\theta m(\theta)}$$

Substituting this expression of rV_u in wage equation (9.20), we get:

$$w = z + (y - z)\Gamma(\theta) \quad \text{with} \quad \Gamma(\theta) = \frac{\gamma[r + q + \theta m(\theta)]}{r + q + \gamma\theta m(\theta)} \tag{9.21}$$

Since the exit rate $\theta m(\theta)$ from unemployment increases with labor market tightness θ, function $\Gamma(\theta)$ likewise increases with θ. This function represents the *actual* weight of the employee in the bargaining. Hence, the balance of power shifts in favor of the employee when θ increases, for in this case the probability of exiting from unemployment, and thus the value V_u of the outside opportunity climb in tandem. The employee then fears the prospect of unemployment less, which pushes the negotiated wage up. A similar line of reasoning will show us why function $\Gamma(\theta)$ is decreasing with the exit rate q from employment. Of course, this function increases with the intrinsic weight γ of the employee in the bargaining. In sum, if $y > z$, equation (9.21) defines a rising monotonic curve between the negotiated wage w and the labor market tightness θ. In the literature, it has become habitual to use the abbreviation (*WC*)—for *wage curve*—to denote the curve that precisely encapsulates the outcome of this bargaining. It is worth noting that the wage curve replaces the labor supply curve from the competitive model. For a given number of vacant jobs, it defines a decreasing relation between wages and the stock of unemployed persons, which is equivalent to a rising relation between wages and employment. Now this property also characterizes the labor supply function in certain circumstances. But this formal analogy should not conceal the profound differences that distinguish the wage curve from the labor supply curve, when workers have bargaining power greater than zero. The wage curve is the upshot of a bargaining process over wages and takes into account characteristics of the labor market like the job destruction rate q and the form $m(.)$ of the matching function. All these parameters are absent in the standard labor supply function, which is the outcome of a limit case in which workers have no bargaining power. In that situation, the gains of unemployed persons z are interpreted as the reservation wage (see chapter 1 for a definition of this notion) below which workers turn down jobs offered to them. That makes the wage offered by employers independent of labor market tightness.

Empirical Elements Relating to Bargaining Power

Much empirical work has been devoted to estimating wage equations similar in form to the one given by relation (9.21)—see Blanchflower and Oswald (1995) and chapter 7. Some of these works aim to estimate the bargaining power of workers by trying to establish that they do in fact obtain a portion of the rent of firms. Abowd and Lemieux (1993) have shown that wages are higher in Canadian firms with little exposure to international competition. They estimate that workers capture 30% of the rent obtained by firms protected from competition. Van Reenen (1996) has, for his part, studied the shareout of rents created by innovation, using British data for the period 1945–1983. He obtains a result similar to that of Abowd and Lemieux, since he estimates that 29% of rent is captured by employees. Blanchflower et al. (1996) carried out the same sort of exercise, attempting to estimate the relationship between wages and profit per capita in the United States in the period 1964–1985. The elasticity of wages with respect to profit per capita amounts to 8%. Manning (2011, table 4) presents the principal estimates of parameter γ supplied by various studies. Large variations appear, most likely linked to the nature of the labor market institutions proper to each country and each sector. But in the majority of studies, the estimate of parameter γ proves to be different from zero. On the whole, these results suggest that workers do in fact capture a portion of the rent

of jobs. The representation of the mode of wage formation as a process of rent-sharing is therefore not invalidated empirically.

3.4 LABOR MARKET EQUILIBRIUM

In the matching model, three relations make it possible to characterize completely the equilibrium values of the unemployment rate, wages, and labor market tightness. They are labor demand, the wage curve, and the Beveridge curve.

3.4.1 THE DETERMINATION OF WAGES, TIGHTNESS, AND THE UNEMPLOYMENT RATE

In the competitive model, summed up by figure 9.20, the intersection of the labor supply and demand curves determines the equilibrium values of wages and employment. In the matching model, the wage curve takes the place of the supply curve. Hence, in plane (θ, w), the equilibrium values θ^* and w^* of the labor market tightness and the wage correspond to the coordinates of the intersection of the wage curve with labor demand respectively defined by relations (9.21) and (9.12). In figure 9.21, we identified the labor demand curve and the wage curve by the abbreviations (LD) and (WC) respectively.

For some of what follows, it will be useful to have a relation that completely defines the equilibrium value of labor market tightness. We obtain this relation by eliminating the wage w between equations (9.12) and (9.21). Taking into account the definition of function $\Gamma(\theta)$—see (9.21) again—we finally get:

$$\frac{(1-\gamma)(y-z)}{r+q+\gamma\theta m(\theta)} = \frac{h}{m(\theta)} \tag{9.22}$$

Most often the impact of exogenous parameters on labor market equilibrium can easily be deduced by looking at the shifts of the (WC) and (LD) curves that they cause.

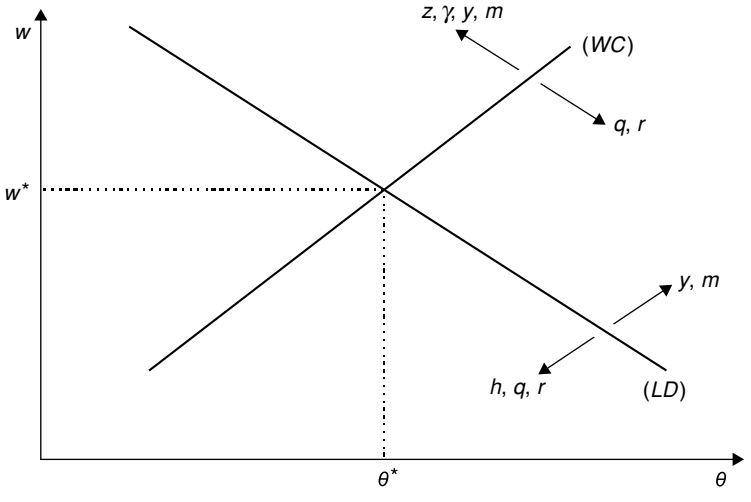


FIGURE 9.21
The negotiated wage and labor market tightness.

But certain ambiguities sometimes persist, and it is then useful to refer to relation (9.22). It is interesting to note that the left side of this relation represents the value of the profit expected from a filled job when the value of the negotiated wage is taken into account; it is a decreasing function of θ. Readers are reminded that the right side represents the average cost of a vacant job; it is an increasing function of θ.

We can easily deduce the equilibrium unemployment rate from that of labor market tightness, taking into account entries into and exits from unemployment. Figure 9.22 represents labor market equilibrium in the plane (v, u). Knowing the equilibrium value θ^* of labor market tightness, the equilibrium value u^* of the unemployment rate is equal to the abscissa of the intersection of the Beveridge curve (BC) and the line that starts from the origin with slope θ^*. This line shows the supply of jobs that maximizes profits when wages and employment are in equilibrium.

3.4.2 Comparative Statics

The comparative statics properties of labor market equilibrium can be deduced by examining figures 9.21 and 9.22, and using equation (9.22), which defines the equilibrium value of the labor market tightness, in case of ambiguity. Table 9.5 assembles the results obtained. We limit ourselves here to presenting succinctly the impact of each parameter in order to illustrate the functioning of the model. The empirical dimension will later be addressed in detail.

The Growth of the Labor Force

The size N of the labor force has no influence on the equilibrium of the model. On the other hand, a rise in the *growth rate* n of the labor force shifts the Beveridge curve upward without changing the (WC) and (LD) curves. The wage remains constant, but unemployment mounts. This result is an offshoot of the hypothesis that all new entrants

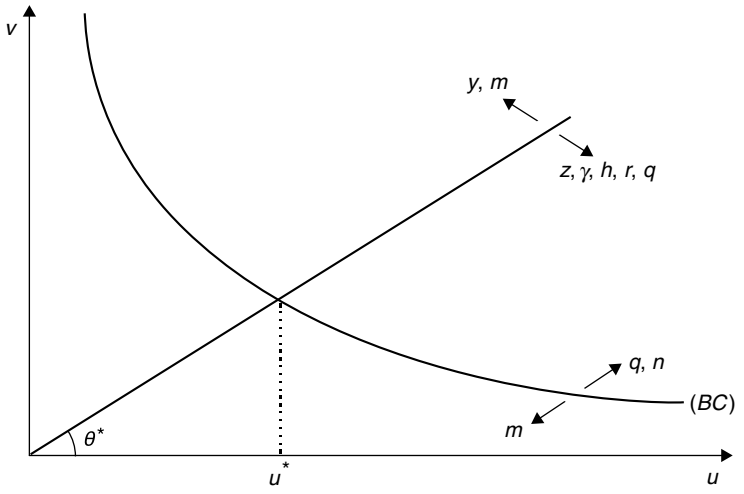


Figure 9.22
Vacant jobs and unemployment.

TABLE 9.5
Comparative statics of stationary equilibrium.

	z	γ	h	m	y	q	r	n
w	+	+	−	+	+	−	−	0
θ	−	−	−	+	+	−	−	0
u	+	+	+	−	−	+	+	+

into the labor market are unemployed. For the same number of vacant jobs, each person in search of work sees his probability of being hired diminish if the number of new entrants is increased, which is equivalent to a deterioration of the matching process.

Bargaining Power
Parameter γ measuring the bargaining power of the employee appears only in expression (9.21) of the wage curve. For a given value of θ, an increase in the employee's power pushes the negotiated wage upward. Since labor demand is unchanged, figure 9.21 shows that the rise in γ involves a shift upward of the wage curve, which in the end provokes a rise in the negotiated wage. This wage rise lowers the profit expected from a filled job, which at free entry equilibrium ought to be equal to the average cost of a vacant job. There will thus be a fall in the number of vacant jobs, which is equivalent to a diminution of θ. The Beveridge curve being independent of γ, unemployment is, in sum, going to increase.

Unemployment Insurance Benefits
The effect of an increase in unemployment insurance benefits z is exactly the same as that of an increase in the bargaining power γ of the employee. By improving the expected utility of an unemployed person, it increases wage pressure. In figure 9.21, we see that the wage curve shifts upward, which pushes the wage up. In total, unemployment increases. Yet, as we saw in chapter 5, section 2.2.1, unemployment benefits are also attended by an *eligibility effect* that runs counter to the effects at work in this simple model.

Productivity
Figure 9.21 shows that a rise in productivity y increases the negotiated wage but has an effect which is a priori ambiguous on the equilibrium value of labor market tightness θ. This ambiguity arises from two effects that have the same origin yet work in opposite directions. A rise in y mechanically increases the "size of the pie" which the worker and the entrepreneur have to divide up. Consequently, with bargaining power held constant, the two protagonists obtain more wages for the one and more profit for the other. The first movement drives firms to diminish the number of vacant jobs; the second gives them an opposing incentive to increase it. This ambiguity as regards the final outcome is illustrated by a simultaneous shift upward of the (WC) and (LD) curves in figure 9.21. Nonetheless, this ambiguity disappears if we go back to equation (9.22), characterizing the equilibrium value of θ. It then becomes evident that an increase in y has a positive effect on θ overall and reduces the unemployment rate. This result is due to the fact that the profit expected from a filled job, taking account of the negotiated value of the

wage—which corresponds to the left side of equation (9.22)—always increases with labor productivity.

It is important to note that these productivity effects depend strongly on the hypotheses that the gains of unemployed persons z and recruitment costs h do not hinge on labor productivity. Now there are good reasons to think that these two parameters are not independent of productivity in the long run: unemployment benefits are most often defined as a fraction of past wages—which is the same as linking them to labor productivity—and search costs certainly rise with the cost of labor. If z and h were perfectly indexed to wages (i.e., $z = z'w$ and $h = h'w$, where z' and h' are constants), it is easy to verify by referring to the main equations that the level of productivity would no longer have any influence on labor market equilibrium. This result signifies that the unemployment rate is likely affected by the *level* of productivity in the short to medium run but is independent of it in the very long term. As we will see in chapter 10, however, the *rate of growth* of productivity affects the unemployment rate even when the gains of unemployed persons and the costs of vacant jobs are perfectly indexed to productivity.

The Efficiency of the Matching Process

Formally, improved efficiency in the matching process comes to the same thing as multiplying the matching function $m(.)$ by a positive coefficient greater than unity. In figures 9.21 and 9.22, we have identified this operation by the letter m. Improved efficiency in the matching process increases the probability of individuals returning to work. The expected utility of an unemployed person increases, which likewise increases the actual power $\Gamma(\theta)$ of workers in wage bargaining. Upward pressure on wages follows; it is revealed in figure 9.21 by an upward shift of the wage curve. In parallel fashion, greater efficiency in the matching process increases the probability of filling vacant jobs, which lowers their average cost. For a given wage, then, firms offer more vacant jobs and θ increases. In figure 9.21, the (LD) curve shifts to the right. In total, wages rise, but the effect on θ is ambiguous, since on one hand this wage rise reduces the number of vacant jobs that are opened up, but on the other the reduction in the average cost of vacant jobs provides an opposing incentive to open up more of them. Relation (9.22), defining the equilibrium value of the labor market tightness, allows us to solve this indeterminacy. We verify that θ increases when the matching process improves. Once again, therefore, the effect on labor demand (LD) proves to be dominant. Finally, figure 9.22 indicates that the unemployment rate falls, since improved efficiency in matching shifts the Beveridge curve downward.

The Job Destruction Rate

Figures 9.21 and 9.22 describing labor market equilibrium show that a rise in the job destruction rate q is strictly equivalent to lowering the efficiency of the matching process m. This is indeed a perfectly logical result, for in this simple model the job destruction rate q and the rate at which vacant jobs are filled, identified by m, represent two facets of the same phenomenon: the reallocation of jobs and workers. The variable m reflects the "job creation" facet, while parameter q reflects, by hypothesis, the "job destruction" facet. Chapters 10 and 13 focus on making the job destruction rate endogenous. This enrichment of the basic model will shed valuable light on the consequences of job protection and technological innovation.

The Interest Rate

As shown by equation (9.22), a rise in the interest rate decreases the surplus of filled jobs. Relation (9.22) indicates that a rise in the interest rate, by depreciating the discounted value of future profits, reduces the incentive to post vacant jobs and, in consequence, increases the unemployment rate. It is important to point out that the interest rate can also affect employment by altering capital investment, and thus labor productivity. This problem will be dealt with below in section 5.

4 THE EFFICIENCY OF MARKET EQUILIBRIUM

The matching process guiding the allocation of labor resources in the market is characterized by the presence of positive between-group externalities and negative within-group congestion effects. An efficient state of the economy will combine these two types of externalities in an adequate fashion.

4.1 TRADING EXTERNALITIES

If the number of vacant jobs rises, each vacant job has a smaller probability of being matched with a worker, but each unemployed person has a higher probability of finding a job. Firms prefer to have as few vacant jobs as possible, so that they will be filled as rapidly as possible, but unemployed persons prefer the inverse: that there should be many vacant jobs, so as to increase their likelihood of being hired. Correspondingly, if the number of unemployed persons rises, each of them has fewer chances of finding a job, while firms see their chances of being able to fill their vacant positions increase. To put it in summary fashion: every unemployed person would like to be the only member of that category and would like the category of vacant jobs facing him to be as full as possible, while every employer would like to be the only one with positions vacant and to be facing a wide array of job seekers. There are *congestion effects* within each category and *positive externalities* between the categories.

An omniscient planner who wished to maximize efficiency would internalize these externalities and would arrive at a social optimum in which the congestion effects and the positive externalities would be "blended" in the manner that best met her choice criterion. Now, wage negotiations taking place *after* the matchup between a vacant job and an unemployed person has occurred will not internalize these externalities, and the decentralized equilibrium of the labor market is not required a priori to correspond to a social optimum. Still, given that the partners to wage bargaining evidently have opposing interests, it is possible that in certain circumstances the optimal blend of positive externalities and congestion effects may occur at labor market equilibrium. More precisely, the decentralized equilibrium of the matching model studied to this point is generally inefficient, except for a particular situation that verifies the Hosios-Pissarides condition (see Hosios, 1990, and Pissarides, 2000, chapter 7, for an exhaustive analysis of the effects of the job search process on global efficiency), which we present below.

However, this result falls within the framework of a model where wages are bargained over and where workers have no way to distinguish among the different jobs offered. We will see that under the hypothesis of directed search, where workers can

orient their job search as a function of the wages offered by firms, the competition in which firms engage to attract workers restores the efficiency of decentralized equilibrium. Thus, unlike the model of random search with wage bargaining on which we have focused thus far, the model of directed search with wage posting entails that decentralized equilibrium is socially efficient.

4.2 The Social Optimum

We begin by defining the social optimum when agents have no preference for the present (the interest rate r goes to zero). That allows us to characterize efficient allocation simply, setting aside the problem of dynamic optimization. The general case is addressed subsequently.

4.2.1 A Useful Particular Case

Assuming that individuals are risk neutral, the planner's criterion corresponds to the discounted value of production per capita, since the marginal utility of a unit of output is independent of the level of income and so is identical for employers, employees, and the unemployed. Reverting to the notations already used, total instantaneous production, denoted Ω, is defined in the following manner:

$$\Omega = yL + zU - hV$$

Note that in this definition of aggregate production, search costs hV linked to the existence of vacant jobs are counted negatively, as they correspond to a loss. Note further that, strictly speaking, the gain z of an unemployed person does not include any transfers like unemployment benefits. In this formulation, z represents an indicator of the return to leisure or to domestic production. Finally, aggregate production evidently takes positive account of production yL of employees. Dividing by the size N of the labor force, and recalling that, by definition $v = \theta u$, we arrive at the expression of ouput per capita:

$$\omega = y(1-u) + zu - h\theta u \tag{9.23}$$

With a constant labor force ($n = 0$), it is possible to characterize the properties of the social optimum very simply when the interest rate r goes to zero. In this case, the planner attempts to maximize output per capita, given the equilibrium of flows in the labor market described by equation (9.8) of the Beveridge curve. The planner's problem is then written:

$$\max_{\{\theta,u\}} \omega = y(1-u) + zu - h\theta u \quad \text{s.c.} \quad u = \frac{q}{q + \theta m(\theta)}$$

Substituting the value of u given by the Beveridge curve equation in ω, the planner's problem takes the form:

$$\max_{\theta} y + \frac{q(z - h\theta - y)}{q + \theta m(\theta)}$$

The first-order condition of this problem yields an equation implicitly defining the optimal value of labor market tightness:

$$\frac{[1-\eta(\theta)](y-z)}{q+\theta m(\theta)\eta(\theta)} = \frac{h}{m(\theta)}, \quad \eta(\theta) = -\frac{\theta m'(\theta)}{m(\theta)} \tag{9.24}$$

This equation highlights the elasticity $\eta(\theta)$ of the matching function with respect to the unemployment rate—readers will easily verify that $\eta(\theta) = UM_U(V,U)/M(V,U)$—although this quantity played no role in decentralized equilibrium. It acquires great importance here, for it is the sensitivity of the matching function that defines the blend of congestion effects and positive externalities in the matching process. When $r = 0$, comparison of relation (9.24) with equation (9.22) giving the value of tightness at decentralized labor market equilibrium, shows that this equilibrium coincides with the social optimum if and only if $\gamma = \eta(\theta)$. This condition, known as the "Hosios condition," indicates that only a value of employee bargaining power equal to the elasticity of the matching function with respect to the unemployment rate gives the right blend of congestion effects and positive externalities. As a general rule, there is no reason for this equality to be satisfied, so market equilibrium is inefficient when wages are negotiated in a decentralized fashion. The following, more strictly technical, subsection shows that the Hosios condition remains true with a strictly positive interest rate.

4.2.2 The General Case

When the interest rate r is superior to zero, welfare analysis no longer comes down to the maximization of the criterion ω in the *stationary* state of the economy, for the social planner must now take into account losses deriving from the inertia present in the evolution of certain variables—here, the evolution of the unemployment rate described by equation (9.7). Again assuming that the labor force remains constant ($n = 0$), the planner's problem takes the following form:[3]

$$\max_{\theta} \int_0^{+\infty} \omega e^{-rt} \mathrm{d}t \tag{9.25}$$

subject to constraint:

$$\dot{u} = q(1-u) - \theta m(\theta) u$$

Let μ be the multiplier associated with this constraint; the Hamiltonian of the planner's problem is written:

$$H = [y(1-u) + zu - h\theta u]\, e^{-rt} + \mu\,[q(1-u) - \theta m(\theta) u]$$

The first-order conditions are given by equations:

$$\frac{\partial H}{\partial \theta} = 0 \quad \text{and} \quad \frac{\partial H}{\partial u} = -\dot{\mu} \tag{9.26}$$

[3] This is a problem of dynamic optimization, which is studied in mathematical appendix B at the end of the book.

Differentiating the Hamiltonian with respect to θ, the first of conditions (9.26) entail, after rearranging terms:

$$he^{-rt} = -\mu m(\theta)[1 - \eta(\theta)] \tag{9.27}$$

And the transversality condition is written:

$$\lim_{t\to\infty} \mu \cdot u = 0$$

If we now derive the Hamiltonian with respect to u, the second of the first-order conditions (9.26) yields:

$$(z - y - h\theta)e^{-rt} - \mu[q + \theta m(\theta)] = -\dot{\mu} \tag{9.28}$$

From this point on we consider only the stationary equilibrium ($\dot{\theta} = 0$), and derivation of relation (9.27) with respect to t entails $\dot{\mu} = -r\mu$. Substituting this value of $\dot{\mu}$ in (9.28) and taking into account the expression of μ extracted from the first-order condition (9.27), it is possible after several rearrangements to write the equation giving the optimal value of labor market tightness in the following form:

$$\frac{[1 - \eta(\theta)](y - z)}{r + q + \theta m(\theta)\eta(\theta)} = \frac{h}{m(\theta)} \tag{9.29}$$

Comparison of this relation with equation (9.22) giving the value of labor market tightness at decentralized labor market equilibrium shows that this equilibrium coincides with the social optimum if and only if $\gamma = \eta(\theta)$. So, with a strictly positive interest rate, we again find ourselves at the Hosios condition.

4.3 Is Labor Market Equilibrium Necessarily Inefficient?

In the matching model utilized to this point, the inefficiency of decentralized equilibrium comes from the absence of mechanisms giving agents an incentive to take the externalities linked to their decisions into account. However, in a great many situations, these mechanisms do exist, thanks to wage setting rules or wage contracts more elaborate than those encompassed by our basic model.

4.3.1 A Model with Directed Search and Wage Posting

In the basic model, wages are bargained over in such a way as to share the rent deriving from job–worker matches. But there exist other modes of wage setting. Employers often announce the remunerations attached to their vacant jobs, for example. In order to show that a mode of wage setting different from that of the basic model is capable of restoring efficiency to decentralized equilibrium, we consider a model close to that proposed by Moen (1997). He assumes that wages are no longer bargained over, but are fixed by employers at the time they open up vacant jobs.

The economy comprises a large number of labor pools, or "islands," indexed by i. The mobility of workers between labor pools is perfect, and a vacant job can be created in any labor pool whatsoever. Unemployed workers are assumed to have perfect information on the situation in each labor pool. Their search activity can be directed toward their preferred employment pool.

At every instant, the number of hires in each labor pool is determined by a matching function identical to the one considered hitherto. In consequence, if there are U_i unemployed persons and V_i vacant jobs, the exit rate from unemployment and the rate at which vacancies are filled in labor pool i are respectively equal to $\theta_i m(\theta_i)$ and $m(\theta_i)$. In each labor pool, the employers with vacant jobs decide to post a hiring wage, denoted w_i. We assume that all employers offer the same wage in each labor pool.[4] This wage is not renegotiable and applies throughout the employer–employee relationship.

The hypothesis of directed search by workers and perfect mobility implies that the expected utility of an unemployed person is the same in all the labor pools, so it will simply be denoted V_u. Assuming further that the job destruction rate q is identical in each labor pool, the expected utility V_{ei} of a person employed in labor pool i satisfies:

$$rV_{ei} = w_i + q(V_u - V_{ei}) \tag{9.30}$$

If the instantaneous gain z of an unemployed person is the same everywhere, the expected utility V_u of a person in search of work satisfies:

$$rV_u = z + \theta_i m(\theta_i)(V_{ei} - V_u) \quad \forall i \tag{9.31}$$

Eliminating V_{ei} between these last two equations, we get, for given V_u, a decreasing relation between w_i and θ_i taking the form:

$$\theta_i m(\theta_i) = (r + q)\frac{(rV_u - z)}{w_i - rV_u} \tag{9.32}$$

This last equation reveals the implications of the competition among entrepreneurs to attract workers into their respective labor pools. Each entrepreneur must offer the same expected utility V_u to those in search of work, but this objective may be attained in several ways. An entrepreneur may open up few jobs, which entails a low exit rate from unemployment $\theta_i m(\theta_i)$, balanced against a high wage. Or he might open up many jobs, which entails a high exit rate from unemployment, balanced against a low wage. Mobility of the unemployed among the different labor pools thus entails that each entrepreneur must trade off between opening up a large number of jobs and offering high wages to attract enough workers.

4.3.2 THE EFFICIENCY OF DECENTRALIZED EQUILIBRIUM

For a given number U_i of unemployed persons in pool i, the optimal strategy for the entrepreneurs present in this pool consists of offering a wage w_i so as to maximize the expected gain from vacant jobs, subject to constraint (9.32). Now the expected gain Π_{vi} from an unfilled job, and the expected profit Π_{ei} from a filled one in pool i are defined by:

$$r\Pi_{vi} = -h + m(\theta_i)(\Pi_{ei} - \Pi_{vi}) \tag{9.33}$$

$$r\Pi_{ei} = y - w_i + q(\Pi_{vi} - \Pi_{ei}) \tag{9.34}$$

[4] Moen (1997) considers a more general case, where the entrepreneurs in the same labor pool can offer different wages, but, at equilibrium, offer the same wage.

Eliminating Π_{ei} between these last two equations, we get the expression of the profit expected from a vacant job as a function of the wage w_i and the labor market tightness θ_i:

$$\Pi_{vi} = \frac{-h(r+q) + m(\theta_i)(y - w_i)}{r + q + m(\theta_i)} \tag{9.35}$$

We can consider that relation (9.32) defines θ_i as a function of w_i; setting to zero the derivative of Π_{vi} with respect to w_i then gives us the first-order condition of the entrepreneurs' problem in labor pool i. It comes to:

$$\left[(y - w_i)m'(\theta_i)\frac{\partial \theta_i}{\partial w_i} - m(\theta_i)\right][r + q + m(\theta_i)] - m'(\theta_i)\frac{\partial \theta_i}{\partial w_i}[m(\theta_i)(y - w_i) - h(r+q)] = 0 \tag{9.36}$$

with, following (9.32):

$$\frac{\partial \theta_i}{\partial w_i} = \frac{-\theta_i}{[1 - \eta(\theta_i)](w_i - rV_u)}; \quad \eta(\theta_i) \equiv -\frac{\theta_i m'(\theta_i)}{m(\theta_i)} \tag{9.37}$$

The free entry condition entails that the entrepreneurs open up jobs as long as the opportunities for profit linked to the opening up of a vacant job are positive. This comes to a stop when $\Pi_{vi} = 0$. The definition (9.35) of the profit expected from a vacant job entails, then, that at equilibrium the last term between brackets in the first-order condition (9.36) is null. Substituting the value of $\partial\theta_i/\partial w_i$ specified by (9.37) in (9.36) and rearranging terms, we arrive at:

$$w_i = rV_u + \eta(\theta_i)(y - rV_u)$$

Comparison of this equation with equality (9.20), which characterizes the negotiated wage in the basic model, shows that the mode of wage setting we have just set out arrives systematically at the Hosios condition $\gamma = \eta(\theta_i)$ and so ensures the efficiency of decentralized equilibrium. This example suggests that competition among firms to attract workers is capable of restoring the efficiency of market equilibrium. It is worth noting, however, that this result arises from the hypothesis that labor contracts are not renegotiated, since they specify a fixed wage. Actually, if $\gamma \neq \eta(\theta_i)$, either party has an interest in proposing a new round of wage bargaining once the hires have been made. So the equilibrium efficiency of the market rests on the hypothesis that employers can make very firm commitments—and this is not necessarily satisfied.

4.3.3 When Union Power Leads to Efficient Allocation

It is interesting to note that other ways of organizing the labor market also make it possible to arrive at an efficient allocation. In particular, a union setting wages for the economy as a whole chooses an efficient allocation if its objective is to maximize the expected utility of the unemployed. This is easily seen if we note that the expected

utility V_u of an unemployed person, eliminating V_e from equations (9.13) and (9.14), is written:

$$rV_u = \frac{z(r+q) + w\theta m(\theta)}{r + q + \theta m(\theta)}$$

Maximization of V_u subject to the labor demand constraint (9.12) gives the solution (9.29) corresponding to the social optimum.

Efficiency and the Incompleteness of Markets

In the presence of externalities, the inefficiency of decentralized equilibrium is caused by the fact that the economy does not comprise enough markets capable of giving individuals the incentive to take all the consequences of their decisions into account. But in that situation, there are incentives to create supplementary markets, and thus the possibility of offering mutually advantageous contracts. In the matching model, as in every configuration, it is necessary to specify the origin of the incompleteness of markets. From this standpoint, Greenwald and Stiglitz (1988) and Mortensen and Pissarides (1999) have proposed models in which intermediaries intervene in the labor market, offering contracts to both unemployed persons and employers with vacancies, in which the wages that will apply to future hires are specified as a function of the amount of time that passes before the hires take place. In that setting, competition among the intermediaries leads to a social optimum.

These examples show that it is possible to imagine institutions compatible with the efficiency of decentralized equilibrium. But it is far from certain that the *actual* functioning of the different labor markets comes close to these theoretical constructs.

5 INVESTMENT AND EMPLOYMENT

In the preceding sections, the problem of choosing capital, and the consequences of this for employment, were set aside completely. This is a major limitation of the model, inasmuch as labor productivity, which influences employment, is itself conditioned by capital. We will see that it is possible to represent investment decisions quite simply in the matching model (Pissarides, 2000; Cahuc and Wasmer, 2001). Taking capital into account leads naturally to the use of a "large-firm" model, in which the firm no longer comprises just one job, as in the basic model studied before, but many wage earners, the number of which is chosen by the firm. When the firm is able to employ many workers, it may have an interest in manipulating their number to influence wages. This mechanism, highlighted by Stole and Zwiebel (1996), modifies the behavior of labor demand and investment in firms. It only comes about when capital does not adjust instantaneously.

In what follows, we begin by analyzing investment decisions in the simplest setting, where firms can adjust their capital instantaneously. The setting with a delay in the adjustment of capital is examined subsequently.

5.1 The Investment Decision

We study the determinants of investment within the traditional framework, in which the technology of firms is represented by a production function with two substitutable factors, labor and capital. We henceforth assume that the production sector of the economy is composed of a large number of identical firms bearing the index i. At every instant, firm i utilizes quantities K_i of capital and L_i of labor to produce a quantity $F(K_i, L_i)$ of the numeraire good. This last expression represents the production function of firm i; it is taken to be strictly increasing with respect to each of its arguments, strictly concave, and with constant returns to scale. The behavior of workers is identical to the one described in the basic model; in particular, all individuals are assumed to be risk neutral. It should be kept in mind that all the variables in the model depend on time, but in what follows we omit the time index for the sake of simplicity.

5.1.1 The Problem of the Firm

In every firm, at every instant, decisions unfold in the following order:

1. The firm decides on its hires. The employers therefore preserve the "right to manage," the principal consequences of which were discussed in chapter 7.

2. The employer negotiates over wages with each worker, one to one, so there is no collective bargaining between the employer and a union representing the interests of the employees. Capital is chosen simultaneously with the wage bargaining. This hypothesis signifies that the employer cannot commit himself to a stock of capital in order to manipulate the wage being negotiated, which depends on productivity. Concretely, we assume that there exists a capital market in which the firm can buy and sell without delay (see Cahuc and Wasmer, 2001).

Hence, at every date, firm i opens up V_i vacant jobs, each of which is filled at rate $m(\theta)$. The number of hires per unit of time is then equal to $m(\theta)V_i$. It should be noted that the rate $m(\theta)$ is given for the firm because labor market tightness is a macroeconomic variable (formally, θ is not indexed by i). Let I_i be the instantaneous investment of firm i, and δ the rate of depreciation of capital. If w_i designates the prevailing wage in firm i, then the problem of this firm is written:

$$\max_{V_i, I_i} \Pi_i = \int_0^{+\infty} \left[F\left(K_i, L_i\right) - w_i L_i - hV_i - I_i\right] e^{-rt} \mathrm{d}t \tag{9.38}$$

subject to:

$$\dot{L}_i = m(\theta)V_i - qL_i \tag{9.39}$$

$$\dot{K}_i = I_i - \delta K_i \tag{9.40}$$

In these expressions, h, q, and r are exogenous parameters representing, as in the basic model, the cost of a vacant job, the exit rate from employment, and the real

interest rate. Constraint (9.40) expresses the law of motion of capital, and constraint (9.39) signifies that in firm i the variation of employment $\dot{L}_i$ is equal to hires $m(\theta)V_i$ minus quits qL_i.

5.1.2 The Optimal Solutions

Problem (9.38), the maximization of the firm's intertemporal profit, is a dynamic optimization problem, in which the state variables are employment L_i and capital K_i. The solution of this type of problem is explained in detail in mathematical appendix B at the end of the book. Let μ and λ be the multipliers associated respectively with constraints (9.39) and (9.40); the Hamiltonian of this problem is written:

$$H = [F(K_i, L_i) - w_i L_i - hV_i - I_i]\, e^{-rt} + \mu\,[m(\theta)V_i - qL_i] + \lambda(I_i - \delta K_i) \tag{9.41}$$

The first-order conditions read:

$$\frac{\partial H}{\partial I_i} = 0 \quad \text{and} \quad \frac{\partial H}{\partial K_i} = -\dot{\lambda} \tag{9.42}$$

$$\frac{\partial H}{\partial V_i} = 0 \quad \text{and} \quad \frac{\partial H}{\partial L_i} = -\dot{\mu} \tag{9.43}$$

To these equations must be added the transversality conditions:

$$\lim_{t\to\infty} \mu L_i = 0 \quad \text{and} \quad \lim_{t\to\infty} \lambda K_i = 0 \tag{9.44}$$

Equalities (9.42) entail $e^{-rt} = \lambda$ and $\lambda\delta - F_K(K_i, L_i)e^{-rt} = \dot{\lambda}$. The first equation entails that $\dot{\lambda} = -r\lambda$. Substituting this expression of $\dot{\lambda}$ into the second equation, we arrive at:

$$F_K(K_i, L_i) = r + \delta \tag{9.45}$$

Relation (9.45) expresses the usual equality between the marginal productivity of capital and its user cost $(r + \delta)$. Conditions (9.43) in turn entail $he^{-rt} = \mu m(\theta)$ and $q\mu - [F_L(K_i, L_i) - w_i]e^{-rt} = \dot{\mu}$. At stationary equilibrium, where $\dot{\theta} = 0$, after several simple calculations we get:

$$F_L(K_i, L_i) = w_i + \frac{h(r+q)}{m(\theta)} \tag{9.46}$$

Relation (9.46) conveys that the marginal productivity of labor must be equal to the wage plus the employment adjustment costs at the optimum. Relations (9.45) and (9.46) show that capital and employment depend on parameters like wages and job destruction rates that are, in principle, specific to each firm, but also on macroeconomic variables like the labor market tightness and the interest rate.

5.2 Wage Bargaining

If we follow the decision sequence set out at the beginning of this section, in stage 2 each employee bargains over his wage individually with the employer. Accordingly, bargaining concerns the *marginal* surplus created by each job—by definition, the expected supplementary gains produced by this job. The value of a marginal job is easily defined in a stationary situation. The marginal job brings in a flow of gains $F_L(K_i, L_i) - w_i$; as well, it is destroyed with a probability q per unit of time. Since every job destroyed brings in zero profit, the value π_i of a marginal job in firm i is written as follows:

$$\pi_i = \left(\frac{1}{1 + r\mathrm{d}t}\right)\{[F_L(K_i, L_i) - w_i]\,\mathrm{d}t + [1 - q\mathrm{d}t]\,\pi_i\} \Leftrightarrow \pi_i = \frac{F_L(K_i, L_i) - w_i}{r + q}$$

This definition of the value of the marginal job is identical to that giving the value of a filled job in the basic model—see (9.9)—on condition of having $\Pi_v = 0$ and identifying individual production y with marginal productivity $F_L(K_i, L_i)$.

From this point of view, it is important to note that the hypothesis of constant returns to scale entails that the marginal productivity of labor does not depend on employment when capital reaches its optimal level. Let us set $k_i = K_i/L_i$ and $f(k_i) = F(K_i, L_i)/L_i$; differentiating this last equation with respect to K_i and L_i we find the marginal productivities of capital and labor, that is, $F_K(K_i, L_i) = f'(k_i)$ and $F_L(K_i, L_i) = f(k_i) - k_i f'(k_i)$. Equality (9.45) between the marginal productivity of capital and its user cost shows that the capital–labor ratio k_i is the same in all firms; we simply denote it by k. In this case, the negotiated wage is also the same in all firms; we denote it by w. More precisely, the first-order conditions (9.45) and (9.46) entail:

$$f'(k) = r + \delta \tag{9.47}$$

$$f(k) - kf'(k) = w + \frac{h(r + q)}{m(\theta)} \tag{9.48}$$

As the capital–labor ratio k is completely defined by the user cost of capital $(r + \delta)$, the marginal productivity of labor $f(k) - kf'(k)$ is also completely determined by knowledge of r and δ. This result allows us to justify the hypothesis of constant individual production y in the basic model, since in reality it represents the marginal productivity of labor, which, with the hypotheses of constant returns of the production function and an exogenous interest rate, does not depend on employment. It should be noted that this marginal productivity is a decreasing function of the interest rate. With this new definition of y, equation (9.48) is identical to relation (9.12) defining labor demand in the basic model.

A further task is to verify the transversality conditions (9.44). When computing the first-order conditions, we saw that multipliers λ and μ were proportional to e^{-rt} at stationary equilibrium. Since $K_i = kL_i$ and since, in the stationary state, L_i grows at rate n in all firms, we observe that the transversality conditions are satisfied if and only if $r > n$.

Finally, the Beveridge curve derives directly from condition (9.39) describing the evolution of employment in the representative firm. Since, by definition, $L = N + U$ with $\dot{N}/N = n$, we come back exactly to equation (9.8) characterizing the Beveridge curve. In sum, this analysis of the matching model with large firms both justifies and

clarifies the use of the simplified model in section 3. In particular, it enables us to study the impact of variations in the interest rate on unemployment in greater depth.

5.3 The Adjustment Lag of Capital

The large-firm model studied to this point assumes an instantaneous adjustment of capital. In reality, capital is a variable that generally adjusts only after a certain delay: it takes time to install new equipment or construct new buildings. For this reason, capital is frequently considered a predetermined variable that cannot adjust instantaneously. Under this hypothesis, the marginal productivity of labor is no longer constant, but decreasing. Hence the value of a marginal job, $F_L(K_i, L_i) - w_i$, is a function of the level of employment, and the wage bargained over by the marginal worker also becomes a function of employment.

Stole and Zwiebel (1996) have elaborated an intrafirm bargaining model to analyze this issue. They consider a partial equilibrium with a single isolated firm. They assume that contract incompleteness prevents either party from committing to future wages and employment decisions. They show that intrafirm bargaining yields no rent for employees and gives rise to overemployment compared to a competitive labor market. The basic idea of Stole and Zwiebel is that a firm with a diminishing marginal productivity of labor can decrease the bargained wage, which increases with the marginal productivity of labor, by raising employment. In Stole and Zwiebel's setting, firms overemploy strategically, up to the point where workers get their reservation wage. Therefore, in their partial equilibrium analysis, intrafirm bargaining implies overemployment.

Cahuc et al. (2008) have studied this issue in a matching model in order to assess this result on equilibrium unemployment. They show that the mechanism highlighted by Stole and Zwiebel does not necessarily give rise to overemployment at market equilibrium. Actually, quantitative exercises suggest that the opposite holds true: the strategic interactions involved in intrafirm bargaining help to increase unemployment because the distortions induced by intrafirm bargaining reduce profits and then job creation. Moreover, increases in the bargaining power of workers are more detrimental to employment when firms use both employment and capital rather than employment alone as strategic instruments to manipulate wages. From this point of view, it would seem that the overemployment phenomenon foregrounded by Stole and Zwiebel does not play an important role at the macroeconomic level.

6 Unemployment Fluctuations

To this point, we have limited ourselves to the study of stationary equilibrium. The study of out-of-stationary-state dynamics allows us to analyze the cyclical variations of unemployment and vacancies.

6.1 The Dynamics of the Vacancies and Unemployment

Analysis of the dynamics of the basic model requires that we reconsider the equations defining the expected utilities and profits. Hence, when the economy moves away

from its stationary state, relations (9.13) and (9.14) defining the expected utility of an employee and an unemployed person, respectively, are now written:[5]

$$rV_e = w + q(V_u - V_e) + \dot{V}_e \tag{9.49}$$

$$rV_u = z + \theta m(\theta)(V_e - V_u) + \dot{V}_u \tag{9.50}$$

The terms $\dot{V}_e$ and $\dot{V}_u$, which represent the time derivatives of V_e and V_u, are interpreted as expected capital gains from changes in the valuation of the assets V_e and V_u. As there is no source of regular growth in the basic model, these terms are null at stationary equilibrium. Correspondingly, profits expected from a filled job and a vacant one, defined by equations (9.10) and (9.11), now take the form:

$$r\Pi_e = y - w + q(\Pi_v - \Pi_e) + \dot{\Pi}_e \tag{9.51}$$

$$r\Pi_v = -h + m(\theta)(\Pi_e - \Pi_v) + \dot{\Pi}_v \tag{9.52}$$

The matching of an unemployed person to a vacant job occasions a surplus S, the time derivative of which is denoted by $\dot{S}$. By definition, we will thus have:

$$S = V_e - V_u + \Pi_e - \Pi_v \quad \text{and} \quad \dot{S} = \dot{V}_e - \dot{V}_u + \dot{\Pi}_e - \dot{\Pi}_v \tag{9.53}$$

Just as in the basic model, we assume that the free entry condition $\Pi_v = 0$ is satisfied at every date, so it likewise comes to $\dot{\Pi}_v = 0$. With the help of definitions (9.53), adding up equations (9.49) and (9.51), which characterize respectively an employee's expected utility and the profit expected from a filled job, entails:

$$(r + q)S = \dot{S} + y + \dot{V}_u - rV_u \tag{9.54}$$

This differential equation describes the time path of the surplus. The surplus is independent of the wage. Accordingly, just as in the basic model, the wage bargaining outcome is similar to a surplus sharing rule conditioned by the respective powers of the participants. So we will again have:

$$V_e - V_u = \gamma S \quad \text{and} \quad \Pi_e - \Pi_v = (1 - \gamma)S \tag{9.55}$$

The free entry condition ($\Pi_v = \dot{\Pi}_v = 0$) and definition (9.52) of the profit expected from a vacant job yield the usual equality between expected profit and average cost $\Pi_e = h/m(\theta)$. The second of the sharing rules (9.55) then entails:

$$S = \frac{h}{(1 - \gamma)m(\theta)} \Longrightarrow \dot{S} = -\frac{hm'(\theta)}{(1 - \gamma)m^2(\theta)}\dot{\theta} \tag{9.56}$$

[5] See mathematical appendix D at the end of this book.

This equation, relation (9.50) characterizing the expected utility of an unemployed person, and the first of the sharing rules (9.55) again entail:

$$rV_u - \dot{V}_u = z + \theta m(\theta)\gamma S = z + \frac{\gamma\theta h}{1-\gamma} \tag{9.57}$$

Bringing the values of S, $\dot{S}$, and $(rV_u - \dot{V}_u)$ given by relations (9.56) and (9.57) into differential equation (9.54) describing the time path of the surplus, and rearranging terms, we arrive at:

$$\frac{hm'(\theta)}{(1-\gamma)m^2(\theta)}\dot{\theta} + \frac{h[r+q+\gamma\theta m(\theta)]}{(1-\gamma)m(\theta)} - y + z = 0 \tag{9.58}$$

This differential equation completely characterizes the path of labor market tightness. In the stationary state ($\dot{\theta} = 0$), this equation is of course identical to relation (9.22) giving the stationary value θ^* of labor market tightness. Equation (9.58) is a first-order, nonlinear differential equation of the form $\varphi(\dot{\theta}, \theta) = 0$. The convergence of θ in the neighborhood of stationary equilibrium can nevertheless be studied very easily by linearizing function φ around point $(\dot{\theta} = 0, \theta = \theta^*)$. After several calculations,[6] we arrive at the following linear differential equation:

$$\dot{\theta} + a\theta = a\theta^* \quad \text{with } a = \gamma\frac{m^2(\theta^*)}{m'(\theta^*)} - (r+q) < 0$$

The general solution of this equation is of the form $\theta = Be^{-at} + \theta^*$, where B is a constant. Parameter a being negative, the unique stable path of θ corresponds to $B = 0$. We then have, at every instant, $\theta = \theta^*$. This result signifies that variable θ immediately "jumps" to the stationary value. It arises from the fact that opening up a vacant job is a "forward-looking" decision, one that takes into account only expectations of *future* profit and contains no inertia factor. The number of vacant jobs can thus adapt immediately to any change in the environment. More generally, all decisions of agents are directed toward the future, so it is easy to verify that the wage negotiated is also a variable that jumps instantaneously to its stationary value.

When labor market tightness has reached its stationary value θ^*, the differential equation (9.7) describing the evolution of the unemployment rate takes the following form:

$$\dot{u} + [q + n + \theta^* m(\theta^*)]u = q + n$$

[6] Let us rewrite equation (9.58) as:

$$m'(\theta)h\dot{\theta} + h[r+q+\gamma\theta m(\theta)]m(\theta) - (1-\gamma)m^2(\theta)(y-z) = 0$$

Differentiation with respect to $\dot{\theta}$ and θ in the neighborhood of $(\dot{\theta} = 0, \theta = \theta^*)$ yields:

$$hm'(\theta^*)\mathrm{d}\dot{\theta} + \left\{h\left[[r+q+2\gamma\theta^* m(\theta^*)]m'(\theta^*) + \gamma m^2(\theta^*)\right] - 2m'(\theta^*)m(\theta^*)(1-\gamma)(y-z)\right\}\mathrm{d}\theta = 0$$

Using the steady-state equilibrium condition (9.22), we get:

$$m'(\theta)\mathrm{d}\dot{\theta} + \left[\gamma m^2(\theta) - m'(\theta)(r+q)\right]\mathrm{d}\theta = 0$$

where $\mathrm{d}\dot{\theta} = \dot{\theta}$ and $\mathrm{d}\theta = \theta - \theta^*$.

This is a first-order linear differential equation in which the coefficient of u is positive. The unemployment rate thus exhibits a monotonic convergence to its stationary value given by relation (9.8). Note that the unemployment rate is thus not a purely forward-looking variable. The average duration of a job search being a positive quantity, there exists at every instant a stock of unemployed persons who represent an element of inertia for the dynamics of the economy. Following a shock, the unemployment rate only gradually reaches its new stationary value.

6.2 The Unemployment Volatility Puzzle

Is the matching model capable of reproducing, to a reasonable approximation, certain important stylized facts about the dynamics of unemployment and vacant jobs? In an oft-cited paper, Shimer (2005) answers this question in the negative. He argues that the matching model cannot reproduce unemployment dynamics well because any productivity shock is immediately absorbed into the wage with little effect on unemployment. Hence, in comparison to observed fluctuations, the matching model would generate too much volatility for the real wage and not enough for the unemployment rate.

We will start by presenting Shimer's critique. Then we will see how it is possible to modify the matching model so as to improve its empirical predictions.

6.2.1 The Puzzle

Shimer (2005) carries out a calibration of the matching model described in section 3 with the help of data on the American economy between 1951 and 2003. He then simulates this model on the assumption that productivity y and the separation rate q follow Markov processes. Shimer observes that, according to empirical data, the tightness ratio is extremely procyclical with a standard deviation of 0.38 around its trend, whereas the standard deviation of the simulated model amounts only to 0.035. It is thus more than 10 times weaker. Likewise the job finding rate is 12 times more volatile according to empirical data than in the simulation of the model. On the other hand, the simulation of the model generates excessive wage volatility. The elasticity of the wage with respect to productivity is close to unity in the simulated model, whereas it would be on the order of 0.45 according to the estimates of Hagedorn and Manovskii (2008), carried out on U.S. data covering almost the same period. For Shimer, it is the mode of wage determination by Nash bargaining that fails to generate enough wage rigidity (defined as a wage rate that changes less than in proportion to the average product of labor over the cycle).

A Calibration of the Matching Model

To grasp Shimer's arguments, it is helpful to use the calibration carried out by Pissarides (2009), which is applied to monthly data with the assumption that $y = 1$. For the interest rate and the rate of job destruction, Pissarides adopts the estimates of Shimer, that is, $r = 0.004$ and $q = 0.036$. He uses a matching function of the Cobb-Douglas type $m_0 u^{\eta} v^{1-\eta}$ with $\eta = \gamma = 0.5$. It is also necessary to assign values to the shift parameter m_0 of the matching function and to parameter h measuring the cost of posting a vacant job. As these two parameters have no obvious counterparts in reality, they cannot be obtained by gathering empirical data, so they must be estimated indirectly. To assign a value to m_0, Pissarides starts by calculating the sample mean for

θ by making use of the Job Openings and Labor Turnover Survey (JOLTS) and the Help-Wanted Index (HWI). He finds that the sample mean for θ comes to 0.72. With the Cobb-Douglas form used for the matching function, the job finding rate is given by $\theta m(\theta) = m_0\theta^{0.5}$. Shimer had calculated that the monthly job finding rate, $\theta m(\theta)$, had an average value of 0.594; from that we can derive m_0 by the equality $m_0 = \theta m(\theta)/\theta^{0.5}$, and we find $m_0 \simeq 0.7$. In order to assign a value to parameter h, we must return to equation (9.22) defining the equilibrium value of the tightness indicator. In this equation, we have $y = 1, r = 0.004$, $q = 0.036$, $y = 1$, $\gamma = 0.5$, $\theta m(\theta) = 0.594$, $\theta = 0.72$, and thus $m(\theta) = 0.825$, which entails:

$$h = 1.224(1 - z) \tag{9.59}$$

Divergent estimates have been made of the net gain of unemployed persons. Pissarides privileges the estimate of Hall and Milgrom (2008), which arrived at $z = 0.71$ (in productivity units), but he also presents results using the variant $z = 0.4$, which was the figure adopted by Shimer (2005). Relation (9.59) indicates that for $z = 0.71$, we get $h \simeq 0.356$, and for $z = 0.4$ we get $h \simeq 0.734$.

It is then possible to calculate the value of the wage with the help of equation (9.12) defining labor demand. Taking into account the numerical values of the parameters, we have:

$$w = y - \frac{h(r+q)}{m(\theta)} = y - 0.059(1 - z)$$

We find $w \simeq 0.98$ for $z = 0.71$ and $w \simeq 0.96$ for $z = 0.4$. This means that wage earners obtain, according to this calibration, a wage lying very close to their marginal productivity.

The Elasticity of the Tightness Ratio and the Wage with Respect to Productivity

With the help of the empirical data assembled by Shimer, Pissarides calculates that the value of the elasticity of the tightness ratio with respect to the productivity of labor comes to 7.56. Does the basic model just calibrated furnish a comparable value for this elasticity? To answer this question, we use equation (9.22) completely defining the equilibrium value of the tightness ratio. With the help of (9.59), this equation becomes:

$$\frac{1.224(1 - z)}{m(\theta)} = \frac{(1 - \gamma)(y - z)}{r + q + \gamma\theta m(\theta)}$$

Or again:

$$\frac{r+q}{m(\theta)} + \gamma\theta = \frac{(1 - \gamma)(y - z)}{1.224(1 - z)}$$

Given the Cobb-Douglas form of the matching function, derivation with respect to y yields:

$$\frac{\mathrm{d}\theta}{\mathrm{d}y} = \frac{1}{\frac{r+q}{2m_0}\theta^{-0.5} + \gamma} \frac{1 - \gamma}{1.224(1 - z)}$$

Using the values of the parameters and the variables for $y = 1$, we find after several calculations:

$$\frac{d\theta}{dy} = \frac{0.76}{1-z}$$

And so:

$$\eta_y^\theta = \frac{y}{\theta}\frac{d\theta}{dy} = \frac{1}{0.72}\frac{0.76}{1-z} = \frac{1.06}{1-z} \tag{9.60}$$

With the numerical values of the net gain of unemployed persons selected by Pissarides, we find $\eta_y^\theta = 3.65$ for $z = 0.71$ and $\eta_y^\theta = 1.77$ for $z = 0.4$. The values of the elasticities yielded by the calibration of the model are thus very distant from the empirically derived value of 7.56. In the matching model, the tightness ratio reacts much less to variations in the productivity of labor than in reality. One may also note that relation (9.60) indicates that we would need to have $z = 0.86$ to reproduce the empirical value of 7.56 for η_y^θ. Such a level of gain for the unemployed is not at all likely.

Conversely, the wage reacts excessively to variations in labor productivity. We can observe this by calculating the elasticity of the wage with respect to this productivity, using the labor demand equation (9.12). Deriving this last equation with respect to y, we arrive, with $\gamma = 0.5$, at:

$$\eta_y^w = \frac{y}{w} - \eta_y^\theta \frac{(y-w)}{2w}$$

Given the values of the parameters, of w and of η_y^θ, we find $\eta_y^w = 0.988$ for $z = 0.71$ and $\eta_y^w = 1.005$ for $z = 0.4$. Wage elasticity with respect to labor productivity is thus practically equal to 1, whereas its empirical value lies at around 0.45 according to the results, mentioned above, of Hagedorn and Manovskii (2008).

6.2.2 Solutions to the "Unemployment Volatility Puzzle"

A range of solutions have been put forward to the "unemployment volatility puzzle." The first assigns a high value to nonmarket activity, the second attempts to make wages more rigid, the third draws a distinction between the wage rigidity of employees in place and the wage rigidity of new entrants, while a fourth solution is based on the heterogeneity of jobs.

The High Value of Nonmarket Activity

Hagedorn and Manovskii (2008) have proposed a model in which the instantaneous income of the unemployed is high and in which they have little bargaining power. The high value of the instantaneous income of the unemployed is justified by the high value of nonmarket activity. In this setting, the matching model correctly reproduces the elasticity of the exit rate from unemployment with respect to productivity.

Wage Rigidity

Many researchers have tried to explain unemployment volatility by wage rigidity. An example is Hall (2005), who assumes in an ad hoc manner that wages do not respond to aggregate productivity shocks. It is then the level of employment that absorbs these shocks. Hall and Milgrom (2008) modify the threat point in Nash bargaining by assuming that it is possible to prolong the bargaining. The threat is then no longer that the work relationship will be broken off but that time and production will be wasted during the bargaining. The empirical data assembled by Hall and Milgrom show that the value of the time and production wasted in bargaining is less cyclical than the dissolution value of the job, which corresponds to the discounted expected utility of unemployed workers. In consequence, in the model of Hall and Milgrom the wage reacts less to variations in productivity, which entails greater volatility in the unemployment rate. Kennan (2010) reaches a similar conclusion with a framework that provides another explanation for wage rigidity. He assumes that productivity is subject to publicly observed aggregate shocks and to idiosyncratic shocks that are seen only by the employer. Kennan shows that small fluctuations in productivity that are privately observed by employers can give rise to a kind of wage stickiness in equilibrium, and the informational rents associated with this stickiness are sufficient to generate relatively large unemployment fluctuations.

Pissarides (2009) has contested this hypothesis of wage rigidity, stressing that, as a factual matter, hiring wages are flexible. He proposes an alternative solution to the unemployment volatility puzzle resting on the presence of fixed hiring costs rather than a hypothesis of wage rigidity.

Mover, Stayer, Starting Wage, Continuing Wage, and Fixed Hiring Costs

Pissarides notes that the wage stickiness highlighted by empirical research is a property of the average wage. It does not apply to all wages. When we concentrate on the wages of employees who are starting new jobs (movers), wage stickiness looks a lot different than it does when we examine the wages of those who have remained in the same job for a while (stayers). The interpretation of the empirical research advanced by Pissarides (2009) and his own estimates indicate that in the United States, the wage elasticity of movers is at least equal to 1, whereas it would lie between 0.3 and 0.5 for the stayers. The calculations of Robin (2011) confirm this observation. They indicate that wage elasticities differ according to deciles. For men for example, wage elasticity amounts to 0.92 for the bottom decile—where the majority of starting wages lie—and to around 0.40 in the middle of the distribution.

In order to adhere to these stylized facts, the model of Pissarides (2009) distinguishes between hiring wages and continuing wages. He assumes that the hiring of a wage earner gives rise to a fixed cost H (which might be an outlay for training, recruitment, increased floor space, etc.). Assuming that the hiring wage is bargained over before the fixed cost kicks in, the global surplus at the moment of hiring—denoted S_0—is written $S_0 = \Pi_0 - \Pi_v + V_0 - V_u$, where Π_0 and V_0 designate respectively the profit expected from a starting filled job and the expected utility for a worker in a starting job. Conversely, the wage for a continuing job is bargained over after the fixed cost has kicked in, and the global surplus—denoted S—is written $S = \Pi_e - (\Pi_v - H) + V_e - V_u$ where Π_e and V_e designate respectively the profit expected from a continuing job and the expected utility for a worker in a continuing job.

The principal departure from the canonical model comes from the expression of the expected profit for a vacant job, which is now written:

$$r\Pi_v = -h + m(\theta)(\Pi_0 - H - \Pi_v)$$

Designating the hiring wage by w_0, the profit expected from a starting filled job reads:

$$r\Pi_0 = y - w_0 + q(\Pi_v - \Pi_0)$$

With the help of the free entry condition $\Pi_v = 0$, these last two equations yield the labor demand, which is written:

$$H + \frac{h}{m(\theta)} = \frac{y - w_0}{r + q} \tag{9.61}$$

Assuming that the hiring wage results from a shareout of the surplus S_0, the same line of reasoning as in the canonical model brings us to:

$$w_0 = z + (y - z)\Gamma(\theta) \quad \text{with } \Gamma(\theta) = \frac{\gamma[r + q + \theta m(\theta)]}{r + q + \gamma\theta m(\theta)} \tag{9.62}$$

Eliminating the wage w_0 between (9.61) and (9.62), we arrive at the equation defining the equilibrium value of the tightness ratio:

$$\frac{(1 - \gamma)(y - z)}{r + q + \gamma\theta m(\theta)} = \frac{h}{m(\theta)} + H \tag{9.63}$$

Pissarides then calibrates the model with fixed hiring cost by taking the case where $z = 0.71$ with the same parameter values as those used to calibrate the base model which served to bring out the unemployment volatility puzzle ($r = 0.004,\ q = 0.036,\ y = 1,\ \gamma = 0.5,\ \theta m(\theta) = 0.594$, and $m(\theta) = 0.825$). Equation (9.63) then yields the following relation between h and H:

$$h = 0.825(0.43 - H) \tag{9.64}$$

To calculate the elasticity of the tightness ratio with respect to productivity, it is convenient to write equation (9.63) as follows:

$$(1 - \gamma)(y - z) = h\left[\frac{r + q}{m(\theta)} + \gamma\theta\right] + H\left[r + q + \gamma\theta m(\theta)\right]$$

Since $\theta m(\theta) = m_0\theta^{0.5}$ and $m(\theta) = m_0\theta^{-0.5}$, deriving with respect to y yields:

$$1 - \gamma = \frac{\mathrm{d}\theta}{\mathrm{d}y}\left[\gamma h + \frac{\theta^{-0.5}}{2}\left(h\frac{r + q}{m_0} + \gamma m_0 H\right)\right]$$

Taking the parameter values into account, we find, using (9.64):

$$\frac{\mathrm{d}\theta}{\mathrm{d}y} = \frac{1}{0.38 - 0.47H} \tag{9.65}$$

Since $\eta_y^\theta = \frac{y}{\theta}\frac{\mathrm{d}\theta}{\mathrm{d}y} = 1.39\frac{\mathrm{d}\theta}{\mathrm{d}y}$, the result is:

$$\eta_y^\theta = \frac{1}{0.27 - 0.34H}$$

It is apparent that η_y^θ is an increasing function of H. In order to adhere to the empirical value $\eta_y^\theta = 7.56$, we would need to set $H = 0.41$ and thus $h = 0.825(0.43 - 0.41) = 0.016$. The average cost of posting a vacant job, given by $h/m(\theta)$, is then equal to 0.019. So the hiring costs would have to be markedly greater than the cost of posting a vacant job in order for the model to reproduce the stylized facts. Pissarides takes the view that such a situation is verisimilar. We saw in chapter 2 that, as a general rule, studies carried out on American data come to the conclusion that hiring costs are much greater than separation costs (Hamermesh, 1993) and that research carried out by recruiting agencies and human resources departments indicates that the replacement cost of a wage earner who quits a firm ranges from 25% of the annual wage of a low-skilled worker to more than 100% for highly skilled employees (Nase, 2009).

The model also yields an elasticity suitable to a hiring wage. To show this, we may write the labor demand in this fashion:

$$w_0 = y - (r + q)\left[H + \frac{h}{m(\theta)}\right] \tag{9.66}$$

Now, (9.63) shows that $H + \frac{h}{m(\theta)}$ takes the same value as in the model without fixed costs. The estimate of w_0 is thus equal to 0.98. With $m(\theta) = m_0\theta^{-0.5}$, deriving (9.66) with respect to y brings us to:

$$\frac{\mathrm{d}w_0}{\mathrm{d}y} = y - \frac{h(r+q)}{2m_0\theta^{0.5}}\frac{\mathrm{d}\theta}{\mathrm{d}y}$$

Taking into account the numerical values of the parameters, and since we have calculated that $\frac{\mathrm{d}\theta}{\mathrm{d}y} = 5.44$, we find $\eta_y^{w_0} = 1.02$. The elasticity of the hiring wage thus lies close to unity.

To determine the rule yielding the continuing wage, denoted w, it suffices to revert to the reasoning of the base model as set forth in section 3. We find:

$$w = rV_u + \gamma(y - rV_u) = \gamma y + (1 - \gamma)\left[rV_u - H(r + q)\right]$$

Since definition (9.14) of V_u and surplus sharing rule (9.16) entail $rV_u = z + \gamma\theta m(\theta)S$, we then have:

$$rV_u = z + \theta m(\theta)\frac{\gamma}{1 - \gamma}\frac{y - w}{r + q}$$

Eliminating rV_u between these last two equations, we find the value of the wage for continuing jobs:

$$w = \gamma y + (1 - \gamma)\frac{r + q}{r + q + \gamma\theta m(\theta)}\left[z - H(r + q)\right] \tag{9.67}$$

Reverting to the numerical values used hitherto, we can calculate the equilibrium values of this wage. It comes to:

$$w = 0.542 - 0.00237H$$

In particular, for the value $H = 0.41$, which allows us to adhere to the empirical value for the elasticity of the tightness ratio with respect to productivity, we find $w = 0.541$.

The elasticity of the wage w with respect to productivity is obtained by deriving relation (9.67) with respect to y. After several calculations, and taking into account the numerical values of the parameters and relation (9.65), we find:

$$\eta_y^w = \frac{y}{w}\frac{\mathrm{d}w}{\mathrm{d}y} = \frac{y}{w}\left(0.5 - 0.036\frac{z - 0.04H}{0.38 - 0.47H}\right)$$

For the value $H = 0.41$, which allows us to adhere to the empirical value of elasticity of the tightness ratio, we then find $\eta_y^w = 0.68$. The wage elasticity of continuing jobs is indeed less than that of starting jobs, even though it is still clearly greater than the empirical elasticity.

Job Heterogeneity and On-the-Job Search

Previous explanations of the unemployment volatility puzzle rely on a matching model that leaves out the heterogeneity of jobs and movements from job to job. Menzio and Shi (2011) and Robin (2011) have looked at the consequences of aggregate productivity shocks in models with heterogeneous jobs and on-the-job search. In these models, aggregate productivity shocks affect movement from job to job and have a major impact on the destruction and creation of low-productivity jobs, with no need to advance any hypothesis about wage rigidity.

Elsby and Michaels (2013) have studied labor market dynamics with a model including large firms. Due to the diminishing marginal product of labor, the model simultaneously generates a large average surplus and a small marginal surplus to employment relationships. The small marginal surplus to employment relationships allows their model to match the volatility of the job finding rate over the cycle, whereas the large value of the average surplus allows their model to match the rate of entries into unemployment. This is progress with respect to the strategy that assumes a small job surplus to generate enough cyclicality in job creation in the standard matching model, because assuming small job surplus with constant marginal returns to labor yields excessive employment-to-unemployment transitions. The calibrated version of the model of Elsby and Michaels (2013) provides a coherent account of the distributions of firm size and employment growth, the amplitude and propagation of the cyclical dynamics of

worker flows, the Beveridge curve relation between unemployment and vacancies, and the dynamics of the distribution of firm size over the business cycle.

On the whole, research on the unemployment volatility puzzle tells us that the matching model is capable of reproducing the relation between productivity and unemployment. There is not, however, a consensus about which hypotheses are most pertinent when it comes to reproducing this relation.

7 SUMMARY AND CONCLUSION

- Over 2000–2011, unemployment touched all OECD countries but in very different proportions. Some countries, like the United States, Japan, Norway, the Netherlands, Switzerland, and the United Kingdom, have an average unemployment rate at or below 6%. But other countries, like France, Greece, and Spain, display an unemployment rate at or above 9%. For the European Union as a whole, the average unemployment rate is in the neighborhood of 9%.
- In most industrialized countries, job creation and destruction are large-scale phenomena. The combined total of these two flows amounts to between 15% and 30% of overall employment every year. For all countries, net employment growth is always much smaller than job creation or destruction. Movements in employment most often take place within the same sector. There is no tendency for the between-sector reallocation of jobs to increase.
- Worker flows are systematically greater in size than job flows. The exit rate from employment is, for most countries, almost twice as large as the rate of job destruction. The rate of entry into employment is about two times greater than the rate of job creation. Worker mobility differs from country to country. The rates of entry into and exit from employment are relatively high in Australia, Finland, Korea, and Mexico, while they are between two and three times lower in Greece, Italy, and the Czech and Slovak republics. A little less than half of the flow of entries and exits from employment concerns persons who are not in employment. The rest comes from direct mobility of workers between two jobs, with no interval of unemployment.
- In the presence of transaction costs, reallocation of jobs and workers can lead to the simultaneous existence of unfilled jobs and unemployed persons. The process through which unemployed persons and vacant jobs are brought together is usually represented by a matching function, indicating the number of hires as a function of the number of vacant jobs and unemployed persons. This function is characterized by positive between-group externalities (the unemployed have an interest in job creation by firms) and congestion effects (each job seeker has an interest in the number of job seekers being as low as possible). The matching process and the equilibrium of worker flows entail a Beveridge curve that links the unemployment rate to the vacancy rate.
- Transaction costs in the labor market lie at the source of exchange externalities which entail that decentralized equilibrium is generally inefficient when wages

are bargained over between employers and workers. There do, nevertheless, exist modes of wage determination such as, for example, competition among entrepreneurs who post wages to attract workers, that make it possible to restore the efficiency of decentralized equilibrium. Overall, the inefficiency of decentralized equilibrium is an open question.

- Shimer (2005) argued that the matching model cannot reproduce unemployment dynamics well because any productivity shock is immediately absorbed into the wage with little effect on unemployment. A range of solutions have been proposed to resolve this "unemployment volatility puzzle." They include assigning a high value to nonmarket activity, distinguishing between the formation of the wages of employees in place and the wage formation of new entrants, and introducing heterogeneity in jobs. With the help of these techniques, the matching model succeeds in reproducing the unemployment dynamics suitably. But at the time of writing, there is no consensus to adopt any one solution.

8 RELATED TOPICS IN THE BOOK

- Chapter 2, section 3: Dynamic labor demand
- Chapter 3, section 1: The competitive model
- Chapter 5: Job search
- Chapter 7, section 2: Bargaining theory
- Chapter 8, section 2.1: Taste discrimination
- Chapter 10, section 1: Technological progress and unemployment
- Chapter 10, section 2.2: A model with skills and tasks
- Chapter 11, section 1.3: Firms' selection and trade
- Chapter 12, section 1.2: The effect of taxes on the labor market
- Chapter 12, section 2.2.2: Minimum wage in labor market with frictions
- Chapter 13, section 2: Employment protection
- Chapter 14, section 2.3: Employment subsidies and the creation of public-sector jobs
- Chapter 14, section 2.4: The equilibrium effects of targeted measures

9 FURTHER READINGS

Davis, S., Faberman, J., & Haltiwanger, J. (2006). The flow approach to labor markets: New data sources and micro-macro links. *Journal of Economic Perspectives*, *20*(3), 3–26.

Elsby, M., Hobijn, B., & Şahin, A. (2013). Unemployment dynamics in the OECD. *Review of Economics and Statistics*, *95*(2), 530–548.

Mortensen, D., & Pissarides, C. (1999). Job reallocation, employment fluctuations and unemployment. In M. Woodford & J. Taylor (Eds.), *Handbook of macroeconomics* (vol. 1B, chap. 18, pp. 1171–1228). Amsterdam: Elsevier Science.

Petrongolo, B., & Pissarides, C. (2001). Looking into the blackbox: A survey of the matching function. *Journal of Economic Literature*, *39*, 390–431.

Pissarides, C. (2000). *Equilibrium unemployment theory* (2nd ed.). Cambridge, MA: MIT Press.

Rogerson, R., & Shimer, R. (2011). Search in macroeconomic models of the labor market. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 4A, chap. 7). Amsterdam: Elsevier Science.

REFERENCES

Abowd, J., Corbel, P., & Kramarz, F. (1999). The entry and exit of workers and the growth of employment: An analysis of French establishments. *Review of Economics and Statistics*, *81*(2), 170–187.

Abowd, J., & Lemieux, T. (1993). The effect of product market competition on collective bargaining agreements: The case of foreign competition in Canada. *Quarterly Journal of Economics*, *108*, 983–1004.

Barnichon, R., & Figura, A. (2011). Labor market heterogeneities, matching efficiency, and the cyclical behavior of the job finding rate (Working Paper). Universitat Pompeu Fabra, Barcelona.

Beveridge, W. (1944). *Full employment in a free society*. London: Allen and Unwin.

Blanchard, O., & Diamond, P. (1990). The aggregate matching function. In P. Diamond (Ed.), *Growth, productivity and unemployment*. Cambridge, MA: MIT Press.

Blanchard, O., & Diamond, P. (1994). Ranking, unemployment duration and wages. *Review of Economic Studies*, *61*, 417–434.

Blanchflower, D., & Oswald, A. (1995). *The wage curve*. Cambridge, MA: MIT Press.

Blanchflower, D., Oswald, A., & Sanfrey, P. (1996). Wages, profits and rent sharing. *Quarterly Journal of Economics*, *111*(1), 227–251.

Borowczyk-Martins, D., Jolivet, G., & Postel-Vinay, F. (2013). Accounting for endogeneity in matching function estimation. *Review of Economic Dynamics*, *16*(3), 440–451.

Bowden, R. (1980). On the existence and secular stability of a *u*–*v* loci. *Economica*, *47*, 35–50.

Burda, M., & Hunt, J. (2011). What explains the German labor market miracle in the Great Recession? *Brookings Papers on Economic Activity*, *42*(1), 273–335.

Burgess, S. (1993). A model of competition between unemployed and employed job searchers: An application to the unemployment outflow rate in Britain. *Economic Journal*, *103*(420), 1190–1204.

Cahuc, P., & Fontaine, F. (2009). On the efficiency of job search with social networks. *Journal of Public Economic Theory*, *11*(3), 411–439.

Cahuc, P., Marques, F., & Wasmer, E. (2008). A theory of wages and labor demand with intrafirm bargaining and matching frictions. *International Economic Review*, *48*(3), 943–972.

Cahuc, P., & Wasmer, E. (2001). Does intrafirm bargaining matter in the large firm's matching model? *Macroeconomic Dynamics*, *5*, 742–747.

Calvo-Armengol, A., & Zénou, Y. (2005). Job-matching, social network and word-of-mouth communication. *Journal of Urban Economics*, *57*, 500–522.

Coles, M., & Smith, E. (1998). Marketplaces and matching. *International Economic Review*, *39*, 239–254.

Davis, S., Faberman, J., & Haltiwanger, J. (2006). The flow approach to labor markets: New data sources and micro-macro links. *Journal of Economic Perspectives*, *20*(3), 3–26.

Davis, S., Faberman, J., & Haltiwanger, J. (2012). Recruiting intensity during and after the Great Recession: National and industry evidence. *American Economic Review*, *102*(3), 584–588.

Davis, S., & Haltiwanger, J. (1992). Gross job creation, gross job destruction, and employment reallocation. *Quarterly Journal of Economics*, *107*, 819–863.

Davis, S., & Haltiwanger, J. (1999). Gross job flows. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 3, pp. 2711–2805). Amsterdam: Elsevier Science.

Elsby, M., Hobijn, B., & Şahin, A. (2013). Unemployment dynamics in the OECD. *Review of Economics and Statistics*, *95*(2), 530–548.

Elsby, M., & Michaels, R. (2013). Marginal jobs, heterogeneous firms, and unemployment flows. *American Economic Journal: Macroeconomics*, *5*(1), 1–48.

Greenwald, B., & Stiglitz, J. (1988). Pareto inefficiency of market economies: Search and efficiency wage models. *American Economic Review*, *Papers and Proceedings*, *78*, 351–355.

Hagedorn, M., & Manovskii, I. (2008). The cyclical behavior of equilibrium unemployment and vacancies revisited. *American Economic Review*, *98*(4), 1692–1706.

Hall, R. (1979a). A theory of the natural unemployment rate and the duration of employment. *Journal of Monetary Economics*, *5*, 153–169.

Hall, R. (1979b). An aspect of the economic role of unemployment. In G. Harcourt (Ed.), *Microeconomic foundations of macroeconomics*. London: Macmillan.

Hall, R. (1999). Labor-market frictions and employment fluctuations. In M. Woodford & J. Taylor (Eds.), *Handbook of macroeconomics* (vol. 1B, chap. 17, pp. 1137–1170). Amsterdam: Elsevier.

Hall, R. (2005). Employment fluctuations with equilibrium wage stickiness. *American Economic Review*, *95*(1), 50–65.

Hall, R., & Milgrom, P. (2008). The limited influence of unemployment on the wage bargain. *American Economic Review*, *98*(4), 1653–1674.

Haltiwanger, J., Scarpetta, S., & Schweiger, H. (2010). Cross country differences in job reallocation: The role of industry, firm size and regulations (EBRD Working Paper No. 116).

Hamermesh, D. (1993). *Labor demand*. Princeton, NJ: Princeton University Press.

Hamermesh, D., Hassink, W., & van Ours, J. (1996). Job turnover and labor turnover: A taxonomy of employment dynamics. *Annales d'économie et de statistique*, *34*, 1264–1292.

Hosios, D. (1990). On the efficiency of matching and related models of search and unemployment. *Review of Economic Studies*, *57*, 279–298.

Kennan, J. (2010). Private information, wage bargaining and employment fluctuations. *Review of Economic Studies*, *77*, 633–664.

Lagarde, S., Maurin, E., & Torelli, C. (1995). Flows of workers and job reallocation. Mimeo, Insee, Direction des Statistiques Démographiques et Sociales.

Lazear, E., & Spletzer, J. (2012). The United States labor market: Status quo or a new normal? (NBER Working Paper No. 18386).

Manning, A. (2011). Imperfect competition in the labor market. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 4B, chap. 11, pp. 973–1041). Amsterdam: Elsevier Science.

Menzio, G., & Shi, S. (2011). Efficient search on the job and the business cycle. *Journal of Political Economy*, *119*(3), 468–510.

Moen, E. (1997). Competitive search equilibrium. *Journal of Political Economy*, *105*, 385–411.

Mortensen, D. (1994). The cyclical behavior of job and worker flows. *Journal of Economic Dynamic and Control*, *18*, 1121–1142.

Mortensen, D., & Pissarides, C. (1994). Job creation and job destruction in the theory of unemployment. *Review of Economic Studies*, *61*, 397–415.

Mortensen, D., & Pissarides, C. (1999). Job reallocation, employment fluctuations and unemployment. In M. Woodford & J. Taylor (Eds.), *Handbook of macroeconomics* (vol. 1B, chap. 18, pp. 1171–1228). Amsterdam: Elsevier Science.

Mumford, K., & Smith, P. (1999). The hiring function reconsidered: On closing the circle. *Oxford Bulletin of Economics and Statistics*, *61*, 343–364.

Nase, D. (2009). The high cost of turnover, 25 September, *Articlesbase*, www.articlesbase.com/human-resources-articles/the-high-cost-of-turnover-1271345.html

OECD. (1995). *Employment outlook*. Paris: OECD Publishing.

OECD. (1996). *Employment outlook*. Paris: OECD Publishing.

OECD. (2010). *Employment outlook*. Paris: OECD Publishing.

OECD. (2012). *Employment outlook*. Paris: OECD Publishing.

OECD. (2013). *Employment outlook*. Paris: OECD Publishing.

Petrongolo, B., & Pissarides, C. (2001). Looking into the blackbox: A survey of the matching function. *Journal of Economic Literature*, *39*, 390–431.

Petrongolo, B., & Pissarides, C. (2008). The ins and outs of European unemployment. *American Economic Review*, *98*(2), 256–262.

Picart, C. (2008). Flux d'emploi et de main-d'œuvre en France: Un réexamen. *Economie et Statistique*, *412*(1), 27–56.

Pissarides, C. (1979). Job matching with state employment agencies and random search. *Economic Journal*, *89*, 818–833.

Pissarides, C. (2000). *Equilibrium unemployment theory* (2nd ed.). Cambridge, MA: MIT Press.

Pissarides, C. (2009). The unemployment volatility puzzle: Is wage stickiness the answer? *Econometrica*, *77*(5), 1339–1369.

Robin, J.-M. (2011). On the dynamics of unemployment and wage distributions. *Econometrica*, *79*(5), 1327–1355.

Rogerson, R., & Shimer, R. (2011). Search in macroeconomic models of the labor market. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 4A, chap. 7). Amsterdam: Elsevier Science.

Shimer, R. (2005). The cyclical behavior of equilibrium unemployment and vacancies: Evidence and theory. *American Economic Review*, *95*(1), 25–49.

Shimer, R. (2012). Reassessing the ins and outs of unemployment. *Review of Economic Dynamics*, *15*, 127–148.

Stole, L., & Zwiebel, J. (1996). Intrafirm bargaining under non-binding contracts. *Review of Economic Studies*, *63*, 375–410.

Van Reenen, J. (1996). The creation and capture of economic rents: Wages and innovations in a panel of UK companies. *Quarterly Journal of Economics*, *111*(1), 195–226.

Yashiv, E. (2000). The determinants of equilibrium unemployment. *American Economic Review*, *90*(5), 1297–1322.

CHAPTER 10

Technological Progress, Unemployment, and Inequality

In this chapter we will:

- Show how technological progress influences job creation and job destruction
- See that technological progress can drive unemployment either up or down
- Study the significant impact that technological progress has had on wage inequality and the occupational structure over the last century
- Estimate the impact of technological progress on wage inequality using the approach of Katz and Murphy (1992) (Data and programs available at www.labor-economics.org allow us to apply this approach to the United States over the period 1963–2008.)
- Analyze the phenomenon of wage and job polarization
- Look at the contrasting profiles of wage inequality and unemployment in North America and Europe
- Observe that investment in education can influence the direction of technological progress

INTRODUCTION

Does technological change induce unemployment and wage inequality? This question has provoked many disputes, which the media have blown up, with the most far-fetched answers often getting the greatest attention. The specter of machines devouring jobs is repeatedly conjured up whenever technological innovation makes it possible to replace humans with mechanical equipment for the accomplishment of certain tasks. The notion that technological progress destroys jobs, taken to the limit, gives rise to the most fantastic predictions. At the beginning of the nineteenth century, Sismondi foresaw a world "where the King sits alone on his island, endlessly turning cranks to produce, with automatons, all that England now manufactures" (Sismondi, 1991, p. 563). More recently, in a book that quickly became a worldwide bestseller and was greeted by reviewers as a prophecy, J. Rifkin predicted the "end of work" as the West moves toward an information economy practically devoid of workers (Rifkin, 1995, p. 93).

Rifkin's mode of argument is to cite examples and situations—numerous, but always one-sided—which, taken together, can give the impression that technological progress actually does destroy jobs and push up unemployment. The fact is that we need to take into account *all* reallocations of jobs and manpower. On average more than 10% of jobs are destroyed every year in the rich countries, but this phenomenon is largely offset by job creation, and we observe no systematic rise in unemployment over the long term (see chapter 9). So in order to assess the impact of technological progress on employment, we have to use a conceptual framework that combines the interactions among technological progress, job destruction, *and* job creation. Conclusions based on accumulated examples neglect the fact that technological progress triggers the process of *creative destruction* highlighted by Schumpeter (1934), the impact of which on unemployment is a priori ambiguous, since it both favors job creation and engenders job destruction. Analysis, both theoretical and empirical, of the impact of technological progress on the level of employment has to be carried out on the macroeconomic scale, not that of particular firms or sectors.

First, we analyze relationships among what happens in the labor market, technological progress, and the creation and destruction of jobs. Technological progress is an important component of growth and contributes to the endless restructuring of production units. As we will see, it has opposing effects on employment, which it favors by creating opportunities for profit but which it also destroys through restructuring. Empirical research confirms these theoretical results, suggesting that technological progress has an ambiguous effect on unemployment.

In section 2, we turn to the effects of technological progress on wage inequalities among workers with different skill levels. We will see that technological progress does not affect all workers in the same fashion. It may boost the productivity of certain workers and render the skills of others obsolete. We will also see that technological progress has helped to increase inequality between skilled and unskilled workers since the beginning of the 1890s in industrialized countries. As well, information technology has taken over jobs formerly held by workers of intermediate skill levels performing routine tasks. This has led to a shrinkage of the proportion of jobs of this type and a drop in the remuneration of the wage earners who hold them. This phenomenon of "job polarization" has occurred in the United States, but also in the European countries. With the help of this documented experience, we will see how institutions influence the impact of biased technological progress on wage inequality and employment according to skill levels. For this purpose it is instructive, as we will see, to contrast a "European" model, characterized by significant compression of wages, thanks to a minimum wage and higher minimum social standards, to an "Anglo-American" model in which the state intervenes in the labor market to a much less marked extent.

1 TECHNOLOGICAL PROGRESS AND UNEMPLOYMENT

Technological progress improves the efficiency of inputs. In the seventeenth and eighteenth centuries, the introduction of new crops and the abandonment of the practice of fallowing land led to a strong increase in agricultural production per hectare and

per worker. In the nineteenth and twentieth centuries, mastery of the powers of steam, electricity, and internal combustion made it possible greatly to increase the ratio of industrial production to the quantities of inputs used. At the end of the twentieth century, innovations in the areas of computerization and telecommunications improved productivity in the service sector. Over a span of centuries, history has been marked by technological innovations that have strongly increased the efficiency of the inputs in the rich countries.

We will see that technological progress contributes significantly to output growth, but its effect on employment and unemployment is a priori ambiguous. On one hand, by improving labor productivity, it increases profits and stimulates more job creation. But on the other, it destroys jobs the technology of which is too outdated to be profitable. Hence technological progress drives a process of job creation and destruction, the outcome of which no one knows beforehand.

We start by providing some empirical facts about productivity growth and unemployment. Then, we will see how technological progress improves labor productivity and therefore increases the profit outlook due to job creation. We use the search and matching model to analyze how this so-called capitalization effect changes the behavior of agents and influences labor market equilibrium. Finally, the last part of this section deals with the consequences of the process of *creative destruction*.

1.1 Facts About Technological Progress, Labor Productivity, and Unemployment

To grasp the impact of technological progress, we must start by defining this concept and measuring it. Then we will examine the relations among technological progress, the growth of labor productivity, and unemployment.

1.1.1 What Is Technological Progress?

Technological progress may affect the productivity of each of the factors of production differently. In order to take the different aspects of technological progress into account, we consider an economy that produces a quantity Y of aggregate output with labor L and capital K, leaving out the time index for simplicity. The aggregate production function is $Y = AF(A_K K, A_L L)$. Technological progress is represented by an increase in the coefficients A, A_K, or A_L. Let us now see how changes in productivity of labor, equal to Y/L, are related to technological progress.

Let Δ be the difference operator (for example, at date t, $\Delta K_t = K_t - K_{t-1}$), and $F_i(A_K K, A_L L)$, $i = 1, 2$, the partial derivative of function F with respect to its i^{th} argument; a first-order Taylor approximation gives:

$$\Delta Y = (\Delta A)F + [(\Delta K)A_K + (\Delta A_K)K]\,AF_1 + [(\Delta L)A_L + (\Delta A_L)L]\,AF_2 \qquad \textbf{(10.1)}$$

Let us assume, for the sake of simplicity, that markets are competitive. In competitive markets, profit maximization entails that the marginal productivity of each input, $AA_K F_1$ and $AA_L F_2$, equals the costs of these inputs. Let $\alpha = L(AA_L F_2)/Y$ be the share of labor in total income. Assuming constant returns to scale, the share of capital is then equal to $(1 - \alpha)$. Let us further denote the growth rate of a variable x by g_x. Dividing

both members of equation (10.1) by Y, and denoting by $g_y = g_Y - g_L$ the rate of growth of labor productivity, we arrive at the celebrated decomposition of Solow (1957):

$$g_y = \underbrace{g_A + (1-\alpha)g_{A_K} + \alpha g_{A_L}}_{\text{Solow residual}} + (1-\alpha)(g_K - g_L) \qquad \textbf{(10.2)}$$

According to this decomposition, labor productivity growth comes from two different sources: technological progress (which can itself take three distinct forms) and capital accumulation. Using series that describe the time path of the inputs and their respective share in GDP, formula (10.2) allows us to estimate the term $g_A + (1-\alpha)g_{A_K} + \alpha g_{A_L}$ linked to technological progress, and commonly called the Solow residual. It should be noted that the Solow residual accounts for all the qualitative factors that increase labor productivity besides capital accumulation. Accordingly, the notion of technological progress captured by the Solow residual is very extensive: not only are technological innovations included in this notion but also all improvements in management practices, in public institutions, and in any other factor likely to increase labor productivity. In all the research carried out using this approach, the significance of the technological progress term is invariably emphasized.

Table 10.1 shows that the Solow residual (denoted by r_S) accounted for most of the growth in labor productivity (measured as GDP per hour worked) in the G7 countries (Germany, Canada, United States, France, United Kingdom, Italy, Japan) during the 1990s and the 2000s. This result signifies that technological progress profoundly influences labor productivity growth in the industrialized countries.

Table 10.2 confirms this result: it displays the correlation between the Solow residual and the rate of growth of labor productivity over the period 1985–2009 for 20 OECD countries. These two variables are presented as five-year averages (1985–1989, 1990–1994...) for each country. Table 10.2, column I, brings out a strong positive linkage between labor productivity growth and the Solow residual. Table 10.2, column II, shows that this relationship still holds when the growth rate of labor productivity is regressed on the Solow residual, including country and period fixed effects in order to control for time-invariant, country-specific unobserved factors and for time-varying unobserved factors common to all countries.

TABLE 10.1
Growth rates (in percentage) of GDP per hour worked (g_y) and total factor productivity in the private sector (r_S) in the 1990s and the 2000s.

	1990–1999		2000–2009	
Country	g_y	r_S	g_y	r_S
Germany	2.23	1.62	1.22	1.02
United States	1.68	1.78	2.06	1.84
France	1.84	1.67	1.18	1.06
Japan	2.44	1.90	1.33	1.36
Italy	1.35	1.89	1.39	0.06
Canada	1.49	1.46	1.13	1.11
United Kingdom	2.40	2.01	1.76	1.60

Source: OECD Productivity Database and Annual National Accounts.

TABLE 10.2
The relationships of the Solow residual, the unemployment rate, and the growth of labor productivity in 20 OECD countries (Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany, Ireland, Italy, Japan, Korea, Netherlands, New Zealand, Portugal, Spain, Sweden, Switzerland, United Kingdom, and the United States) over the period 1985–2009.

	Labor productivity	Labor productivity	Unemployment	Unemployment
	I	II	III	IV
Solow residual	1.00 (0.03)	0.92 (0.05)	−0.18 (0.28)	−0.24 (0.30)
Country and period fixed effects	No	Yes	No	Yes
Adj R^2	0.92	0.93	−0.01	0.69

Note: Variables averaged over 5-year periods. GDP per hour worked is measured in U.S. dollars, at constant prices and constant PPPs. Standard errors in parentheses.

Source: OECD Productivity Database and Annual National Accounts.

1.1.2 TECHNOLOGICAL PROGRESS AND LABOR TURNOVER

The reorganization of the apparatus of production needed to realize productivity gains may take the form of the creation of firms, the destruction of firms, or the reallocation of jobs between firms or within the same firm. For example, Foster et al. (2006) have analyzed the consequences of the evolution of economic activity undergone in the retail sector in the 1990s in the United States. During that period, the information technology revolution has had a strong impact on the retail sector. The adoption of systems that electronically link cash registers to scanners, credit card processing machines, customer relationship management systems, and inventory management systems allowed establishments to increase labor productivity. Foster et al. (2006) find that virtually all of the labor productivity growth in the retail sector is accounted for by more productive entering establishments displacing much less productive exiting establishments. In another study carried out on the automobile repair sector in the United States between 1987 and 1992, Foster et al. (2001) estimate that the contribution of new firms to the growth of labor productivity in this sector was greater than the total growth of this variable. This result means that the "older" firms still in business contribute *negatively* to the growth of labor productivity in that sector.

More generally, the method of Griliches and Regev (1995) makes it possible to pinpoint the respective contributions of firm creation, firm destruction, the reallocation of jobs between continuing firms, and also the gains in labor productivity within firms, to the growth of the labor productivity of an entire sector of the economy. Griliches and Regev write the variation of the labor productivity of a sector of the economy as follows:

$$\Delta P_t = \sum_{\text{Continuers}} \left[\underbrace{\bar{\theta}_i \Delta p_{it}}_{\text{Within}} + \underbrace{\Delta \bar{\theta}_i \left(\bar{p}_i - \bar{P} \right)}_{\text{Between}} \right] + \sum_{\text{Entries}} \underbrace{\theta_{it} \left(p_{it} - \bar{P} \right)}_{\text{Entry}} - \sum_{\text{Exits}} \underbrace{\theta_{it-k} \left(p_{it-k} - \bar{P} \right)}_{\text{Exit}}$$

where Δ means changes over the $k-$year interval between the first year $t-k$ and the last year t; θ_{it} is the employment share of firm i in the given industry at time t; and p_{it} is the labor productivity of firm i. A bar over a variable indicates the averaging of the variable over the first year $t-k$ and the last year t. In the equation above, the first term

is the "within" component; the second is the "between" component, while the third and fourth terms are the entry and exit components, respectively.

Figure 10.1 gives the results of this decomposition of the growth in labor productivity in the manufacturing sectors in eight OECD countries for two five-year intervals, 1987–1992 and 1992–1997. This figure shows that the gains in labor productivity were essentially realized within firms. The contribution from exiting firms is generally positive, which means that the firms destroyed generally have weaker productivity than those which survive. The contribution from net entries (entries minus exits) is positive for the whole group of countries, except for Germany in the 1990s. On average this contribution is significant, accounting for 20% to 40% of the gains in labor productivity.

The results portrayed in figure 10.1 must be interpreted in light of the method used. They are especially dependent on the length of the intervals of time $(t - k)$ chosen: five years in this case. Assuming that it takes a few years for entering firms to realize significant productivity gains, choosing a shorter interval would diminish the contribution of entering firms to labor productivity gains. Still, these results show that overall the process of creation and destruction of firms, as well as reallocations of production

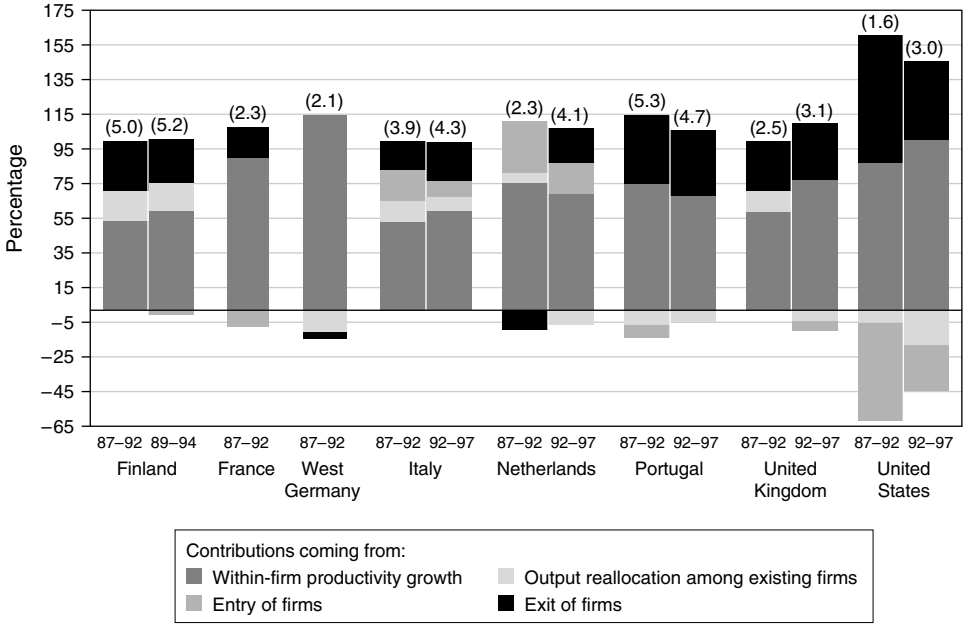


FIGURE 10.1

Decomposition of labor productivity growth in manufacturing. Percentage share of total annual productivity growth of each component.

Note: Figures in parentheses are overall productivity growth rates (annual percentage change).

Source: OECD (2003, figure 4.1, p. 134).

between firms, contribute significantly to the gains in labor productivity of the manufacturing sector over the period in question. This conclusion also holds good for the service sector and for the case where the decomposition bears on multifactor productivity growth (i.e., the Solow residual) rather than on labor productivity (OECD, 2003).

Productivity gains within firms can be achieved by improving the productivity of the workforce in place, especially through training, but also by renewing it. Thus Bauer and Brender (2004) and Givord and Maurin (2004) find, for Germany and France respectively, that firms that use information and communication technologies most intensively have higher manpower rotation. Lynch and Black (1998) for the United States and Behaghel et al. (2012) for France find that the adoption of new technologies is attended by higher outlays on training. In total, it appears that firms with the highest productivity gains adopt a more dynamic style of workforce management, relying more heavily on internal promotion, and hiring and firing more frequently. This phenomenon is well illustrated by Bloom et al. (2012), who show that U.S. multinationals have higher productivity from information and communication technologies than non-U.S. multinationals, primarily due to their tougher "people management" practices, which include more intensive use of promotions, rewards, hirings, and firings.

1.1.3 Technological Progress and Unemployment

We have just seen that productivity growth comes about through the reallocation of jobs and manpower. Thus the relation between productivity growth and unemployment is a priori ambiguous in sign.

There are a limited number of empirical studies dedicated to the relationship between the unemployment rate and the rate of productivity growth. They generally conclude that there is not a systematic and robust correlation between the different measures of the growth rate of productivity and the unemployment rate (see Bean and Pissarides, 1993, and Caballero, 1993, for example). To illustrate these results, figure 10.2 displays the correlation between the Solow residual and the unemployment rate over the period 1985–2009 for 20 OECD countries. These two variables are presented as five-year averages (1985–1989, 1990–1994…) for each country. Figure 10.2 brings out no linkage between the unemployment rate and technological progress measured by the Solow residual.

As shown by table 10.2, column III, this is confirmed by the regression of the unemployment rate onto the Solow residual. Table 10.2, column IV, shows that this result holds when country and period fixed effects are accounted for. At the aggregate level, technological progress is not correlated with unemployment. It is necessary to resort to more fine-grained analysis, taking special notice of the characteristics of the innovations that give rise to technological progress and labor market institutions, in order better to understand the impact of technological progress on unemployment.

1.2 The Capitalization Effect

Technological progress improves labor productivity and therefore increases the profit outlook due to job creation. This so-called capitalization effect changes the behavior of agents and influences labor market equilibrium. The basic model from chapter 9, slightly

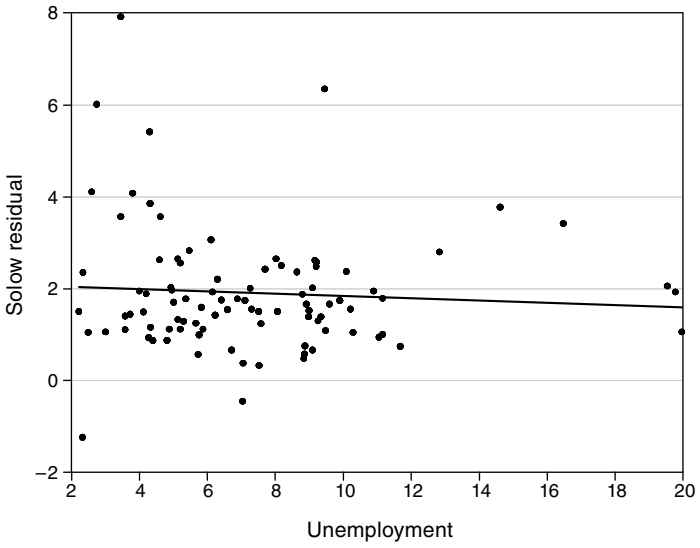


FIGURE 10.2
The relationship between the Solow residual and the unemployment rate in 20 OECD countries (Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany, Ireland, Italy, Japan, Korea, Netherlands, New Zealand, Portugal, Spain, Sweden, Switzerland, United Kingdom, and the United States) over the period 1985–2009.

Note: Variables averaged over 5-year periods. GDP per hour worked is measured in U.S. dollars, at constant prices and constant PPPs.

Source: OECD Productivity Database and Annual National Accounts.

modified, allows us to study the consequences of the capitalization effect. Technological progress can easily be brought into the basic model by assuming that an individual employee's (exogenous) production y grows at a constant rate denoted by g. We may note that there exists a relationship between the components of the Solow residual and the growth rate of labor productivity. Individual production $y \equiv Y/L$ grows at rate $g = g_Y - g_L$, and if we denote the Solow residual by $r_S = g_A + (1-\alpha)g_K + \alpha g_{A_L}$, equation (10.2) entails $g = r_S + (1-\alpha)(g_K - g_L)$. The individual productivity growth rate is equal to the Solow residual if the capital–labor ratio, K/L, and the share α of labor in total income remain constant. On the other hand, a reduction in the growth rate of the capital stock, which might for example occur as certain firms relocate to low-wage countries, leads to a reduction in the growth rate of labor productivity.

1.2.1 The Discount Rate and the Capitalization Effect

It turns out that productivity growth changes act like changes in the discount rate and thus play a part in intertemporal choices.

The "Effective" Discount Rate and Growth

If production grows at rate g, the incomes of agents increase at this rate as well along a balanced growth path (which we can also refer to as stationary equilibrium; in what

follows we use both expressions interchangeably). Consequently we need to modify the expressions of expected profit and utility, returning to chapter 9, section 3.2, and considering a short interval of time lying between dates t and $t + \mathrm{d}t$. If Π_e designates the profit expected from a job occupied at date t, at stationary equilibrium this profit will have increased by $g\mathrm{d}t$ % between dates t and $t + \mathrm{d}t$. Let w be the real wage and let Π_v be the profit expected from a vacant job at date t; relation (9.9) from chapter 9, giving the value of the profit expected from a filled job in the stationary state, will now be written:

$$\Pi_e = \frac{1}{1 + r\mathrm{d}t}\left[(y - w)\mathrm{d}t + q\mathrm{d}t(1 + g\mathrm{d}t)\Pi_v + (1 - q\mathrm{d}t)(1 + g\mathrm{d}t)\Pi_e\right] \qquad \textbf{(10.3)}$$

This equation indicates that the discounted expected profit from a job is equal to the discounted sum of the flow of instantaneous profit $(y - w)\mathrm{d}t$ over interval of time $\mathrm{d}t$ and of the discounted expected future profits. With a probability $q\mathrm{d}t$ these future profits will coincide with the expected profit $(1 + g\mathrm{d}t)\Pi_v$ from a vacant job, and with the complementary probability $(1 - q\mathrm{d}t)$ they will equal the expected profit $(1 + g\mathrm{d}t)\Pi_e$ from a filled job. After several rearrangements of terms, relation (10.3) takes this form:

$$(r - g)\Pi_e = (y - w) + q(1 + g\mathrm{d}t)(\Pi_v - \Pi_e)$$

Making $\mathrm{d}t$ go to 0, we get:

$$(r - g)\Pi_e = y - w + q\,(\Pi_v - \Pi_e) \qquad \textbf{(10.4)}$$

This equation[1] expresses the equality of the returns to different assets on a perfect financial market. An asset worth Π_e at date t, "invested" in the labor market, procures an instantaneous profit of $(y - w)$ to which is added the average gain $q\,(\Pi_v - \Pi_e)$ resulting from a possible change of state (a filled job can fall vacant at rate q). During this same interval of time, the value of this asset has risen by $g\Pi_e\mathrm{d}t$. In other words, the possessor of the asset can make a capital gain of $g\Pi_e\mathrm{d}t$ by selling his good at date $t + \mathrm{d}t$. Let us now suppose that this same asset is "invested" in a financial market offering a fixed interest rate r between dates t and $t + \mathrm{d}t$. It then earns $r\Pi_e\mathrm{d}t$ for its possessor. It turns out that there is an *opportunity cost*—precisely equal to $g\Pi_e\mathrm{d}t$—when the asset is invested in a financial market offering a fixed interest rate r in an environment characterized by regular growth at rate g. The *effective* return on the investment in the financial market is thus equal to $(r - g)\Pi_e\mathrm{d}t$. In sum, in an economy growing regularly at rate g, the effective rate of interest, that is, the discount rate actually used by agents to calculate the present discounted value of their gains, is equal to $(r - g)$. So the growth of the economy is simply accompanied by a capitalization effect equivalent to a reduction in the interest rate by an amount equal to the growth rate of productivity.

Labor Demand

On a balanced growth path, the exogenous parameters of the model all have to increase at the same rate. With no loss of generality, we may take the view that the costs arising

[1] Mathematical appendix D at the end of this book includes a rigorous proof of this type of formula, based on the assumption that certain well-specified random events follow Poisson processes.

from a vacant job are indexed to production y and can thus be written hy where h is a constant exogenous parameter. The expected profit from a vacant job is then written:

$$(r-g)\Pi_v = -hy + m(\theta)\,(\Pi_e - \Pi_v)$$

When the free entry condition $\Pi_v = 0$ is satisfied, the expected profit from a filled job Π_e should be equal to the average cost of a vacant job $hy/m(\theta)$, and relation (10.4) then gives labor demand:

$$\frac{y-w}{r-g+q} = \frac{hy}{m(\theta)} \tag{10.5}$$

For given wage w, the expected profit from an occupied job—represented by the left side of (10.5)—increases with g. Since the latter must exactly cover the average cost of an unfilled job, the average duration of a job remaining unfilled $1/m(\theta)$ increases, and consequently the labor market tightness θ rises too. In other words, for a given stock of unemployed persons and a given wage, firms open up more vacant jobs when g increases. Thanks to the capitalization effect, the growth in productivity exerts a *positive* effect on labor demand. In the (w,θ) plane, a rise in g appears as a shift upward of the (LD) curve. This shift is shown in figure 10.3.

1.2.2 When Technological Progress Reduces Unemployment

The capitalization effect alters the negotiated wage and through this channel influences the properties of the wage curve exhibited in the basic search and matching model (chapter 9, section 3.3).

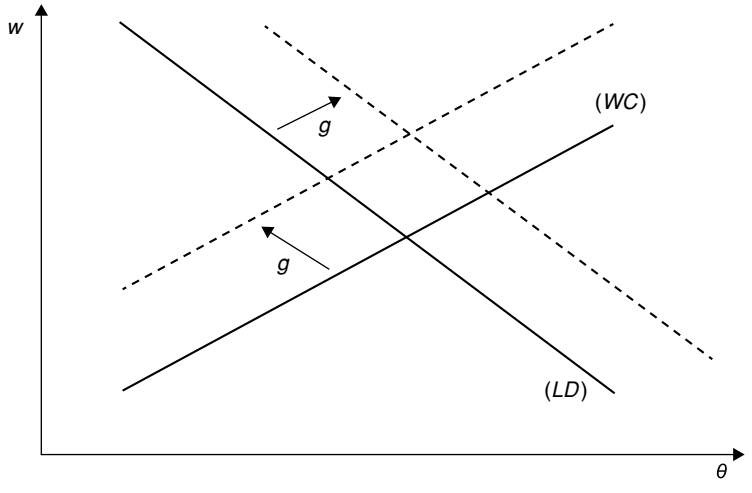


FIGURE 10.3
The effect of an increase in productivity.

Bargaining and the Wage Curve

With a line of reasoning analogous to that which brought us to condition (10.4) describing the expected profit of an occupied job, we find that the expected utility V_e of an employee receiving wage w satisfies:

$$(r-g)V_e = w + q(V_u - V_e) \tag{10.6}$$

In this relation, V_e and V_u designate respectively the expected utility of an employee and an unemployed person at date t. The existence of a balanced growth path entails that the gains of unemployed persons also increase at rate g. With no loss of generality, we assume that these gains are indexed to individual productivity and denote them by zy, where $z \in [0,1)$ is a constant exogenous parameter. The expected utility V_u of an unemployed person then solves:

$$(r-g)V_u = zy + \theta m(\theta)(V_e - V_u) \tag{10.7}$$

As regards wage bargaining, we note that this model is identical to the basic model of chapter 9, provided we simply change z to zy and r to $(r-g)$. If we make these substitutions in relation (9.21) from chapter 9, we get the equation of the wage curve describing the bargaining outcome in an economy growing regularly at rate g:

$$w = y[z + (1-z)\Gamma(\theta)] \quad \text{with } \Gamma(\theta) = \frac{\gamma[r-g+q+\theta m(\theta)]}{r-g+q+\gamma\theta m(\theta)} \tag{10.8}$$

We see that the capitalization effect entails that the strength Γ of an employee in bargaining increases with g. The reason for this result is that a rise in g corresponds to a reduction in the effective interest rate, which reduces the losses to the employed that ensue when a job is destroyed. So the employee has less fear of the prospect of unemployment, her bargaining position is strengthened, and in figure 10.3 the wage curve, denoted (WC), shifts upward. All other things being equal, productivity growth thus has a tendency to increase the negotiated wage.

Labor Market Equilibrium

The equilibrium values of θ and w correspond to the coordinates of the intersection of the (WC) and (LD) curves in figure 10.3. Knowing θ, the unemployment rate u on a balanced growth path can be deduced with the help of the relationship between θ and u compatible with equilibrium of flows in the labor market, expressed by the Beveridge curve: $u = (q+n)/[q+n+\theta m(\theta)]$, where n designates the growth rate of the labor force—see chapter 9, section 3.1. Note that the growth rate g of productivity does not come into the equation of this curve.

Figure 10.3 shows, first of all, that a rise in g has a positive effect on the equilibrium real wage. This result signifies that stronger productivity growth raises the *level* of the real wage. On the other hand, the effect of g on the equilibrium value of the labor market tightness θ turns out to be ambiguous a priori. By combining relations (10.5) and

(10.8), which define the (LD) and (WC) curves, however, we get an implicit equation that brings in θ alone:

$$\frac{(1-\gamma)(1-z)}{r-g+q+\gamma\theta m(\theta)} = \frac{h}{m(\theta)} \tag{10.9}$$

It is easy to verify that θ rises with g. Hence, stronger productivity growth increases the exit rate from unemployment $\theta m(\theta)$. The Beveridge curve being independent of g, we can deduce that stronger growth also reduces the unemployment rate. This conclusion springs from the fact that the profit from a filled job taking account of the negotiated wage—this profit is represented by the left side of equation (10.9)—rises with g.

This model describes a linkage between growth and unemployment. It has (at least) one major drawback, though: the source of job destruction is exogenous. Yet one of the strong tenets of the theory of growth is that technological innovations favor the creation, temporarily at least, of jobs that incorporate the most recent innovations and render certain existing jobs obsolete. This is the process of *creative destruction* described by Schumpeter (1934) and formalized by Aghion and Howitt (1992, 1998) and Mortensen and Pissarides (1998). Let us suppose that stronger productivity growth accelerates the destruction of jobs; we will then have $q = q(g)$ with $q'(g) > 0$. Relation (10.9) then shows that it is far from certain that the expected profit from a filled job increases with g. The acceleration of job destruction runs counter to the capitalization effect, and it is possible that a rise in unemployment will occur. The model developed in the next subsection throws light on these chains of causality and suggests that productivity growth could be positively linked to the level of unemployment.

1.3 Creative Destruction

In the previous model, the productivity of any job whatsoever increased regularly at rate g. To some extent, this hypothesis means that all jobs benefit uniformly, and at no cost, from the latest technological innovations. But in reality it is not, as a general rule, possible to apply the latest innovations to existing jobs without significant expense. In many areas, individual jobs continue to use more or less the same technology they began with, for as long as they last, and are finally destroyed precisely when the evolution of technology makes it unprofitable to keep them going. They are then "replaced," but not necessarily in the same firm, by a new job that incorporates the most recent technological innovations. In this process, the lifespan of each job, and thus the job destruction rate, are *endogenous* variables determined by, among other things, the rate of innovations.

We will illustrate this mechanism in a simple model, which shows how the emergence of new, more productive jobs can destroy older jobs and thus give rise to unemployment. In this model, there is no distinction between firm and job. More sophisticated models, developed notably by Lentz and Mortensen (2005, 2008, 2010), distinguish between job creation and destruction and between the entries and exits of firms.

1.3.1 A Model with Endogenous Job Destruction

In an economy that is growing regularly and that suffers no exogenous shocks, jobs disappear when the technology they employ no longer yields a positive surplus. This

condition allows us to characterize the lifespan of a job and therefore the rate at which jobs are destroyed.

The Lifespan of a Job

To give the simplest possible notion of the mechanism of job destruction and creation, we assume that the productivity of each *new* job increases at a constant exogenous rate g, but that all jobs keep their original productivity over the whole span of their existence. In other words, if y designates the productivity of a job created at date $t = 0$, that job keeps its productivity y permanently, whereas a job created at date $t \geq 0$ is assigned a productivity of $y(t) = ye^{gt}$ over its lifespan. In this model, the definition of job creation needs to be specified. We assume, for simplicity, that a job is created when an unemployed person and a vacant job are matched up. The productivity of a job thus incorporates the most recent innovations available at that moment, and not at the moment a vacant job is opened up.

To contrast better the lessons of this model with those of the preceding models, we assume further that there is no exogenous source of job destruction. A job disappears when the cost of keeping it going is greater than what it brings in, so the lifespan T of a job is an endogenous variable. The rate of job destruction, which we again denote by q, is thus also an endogenous variable, the stationary value of which is easily deduced from knowledge of T. If θ and U designate respectively the stationary values of the labor market tightness and the stock of unemployed persons present at every instant in the labor market, the number of jobs created per unit of time is equal to $\theta m(\theta)U$. As every job has a lifespan T, there are $L = \theta m(\theta)UT$ jobs occupied at every instant. If we assume, for simplicity, that the growth rate of the population is null, then at stationary equilibrium we have $qL = \theta m(\theta)U$, and so $q = 1/T$.

Expected Utilities and Profits

Let us consider a job created at date x the lifespan of which is equal to T, and let us denote by $w(x, s, T)$ the wage attached to this job after it has lasted for a period $s \in [0, T]$. Let us denote by $V_e(x, s, T)$ the expected utility of a worker at date $x + s$ who occupies a job created at date x with a lifespan equal to T. We can then define $V_e(x, 0, T)$ as follows:

$$V_e(x, 0, T) = \int_0^T w(x, s, T)e^{-rs}\mathrm{d}s + e^{-rT}V_u(x + T) \qquad \textbf{(10.10)}$$

where $V_u(x + T)$ designates the expected utility of an unemployed person whose job is destroyed at date $(x + T)$. The existence of a balanced growth path dictates that the gains of unemployed persons increase at rate g. For simplicity, we assume that these gains are indexed to productivity and denote them by $zy(t)$, where $z \in [0, 1)$ is an exogenous parameter. In these conditions, the equation describing the time path of the expected utility of an unemployed person on a balanced growth path takes the form:

$$(r - g)V_u(t) = zy(t) + \theta m(\theta)\left[V_e(t, 0, T) - V_u(t)\right] \qquad \textbf{(10.11)}$$

To lighten the notations from this point forward, we reason on the basis of a matchup occurring at date $x = 0$. As there is no exogenous source of job destruction

and the level of productivity is always equal to y, the expected profit at a date $t \in [0, T]$ flowing from a hire made at date 0, that is, $\Pi_e(0, t, T)$, is written as follows:

$$\Pi_e(0, t, T) = \int_t^T [y - w(0, s, T)] e^{-r(s-t)} \mathrm{d}s + e^{-r(T-t)} \Pi_v(T) \tag{10.12}$$

where $\Pi_v(t)$ designates the expected profit from a job that falls vacant at date t. Symmetrically, a person employed in a job created at date 0 attains at date $t \in [0, T]$ an expected utility $V_e(0, t, T)$ given by:

$$V_e(0, t, T) = \int_t^T w(0, s, T) e^{-r(s-t)} \mathrm{d}s + e^{-r(T-t)} V_u(T) \tag{10.13}$$

The Surplus

By definition, the surplus $S(0, t, T)$ yielded at date $t \in [0, T]$ by a match at date 0 is equal to:

$$S(0, t, T) = V_e(0, t, T) - V_u(t) + \Pi_e(0, t, T) - \Pi_v(t)$$

When the free entry condition $\Pi_v(t) = 0$ is satisfied at every date t, relations (10.12) and (10.13) allow us to write the surplus $S(0, t, T)$ in the following form:

$$S(0, t, T) = y \int_t^T e^{-r(s-t)} \mathrm{d}t + e^{-r(T-t)} V_u(T) - V_u(t), \quad \forall\, t \in [0, T] \tag{10.14}$$

Recalling that at stationary equilibrium $V_u(T) = V_u(t) e^{g(T-t)}$, after several simple calculations, we get:

$$S(0, t, T) = \frac{1 - e^{-r(T-t)}}{r} y - \left[1 - e^{-(r-g)(T-t)}\right] V_u(t) \tag{10.15}$$

The Optimal Lifespan of a Job

Let $\gamma \in [0, 1]$ again be the relative bargaining power of an employee; at each date $t \in [0, T]$ the outcome of bargaining corresponds to a shareout of the surplus $S(0, t, T)$ according to the usual formulas:

$$V_e(0, t, T) - V_u(t) = \gamma S(0, t, T) \quad \text{and} \quad \Pi_e(0, t, T) - \Pi_v(t) = (1 - \gamma) S(0, t, T), \quad \forall\, t \in [0, T] \tag{10.16}$$

This sharing rule shows that the employer and the employee both have an interest in staying together as long as the job yields a positive surplus. In other words, the job should be destroyed on the date the marginal surplus yielded by extending its lifespan becomes negative. Let $S_3(0, t, T)$ be the partial derivative of the surplus with respect to its third argument; the optimal lifespan of a job must then satisfy conditions $S_3(0, T, T) = 0$, and $S_{33}(0, T, T) < 0$. Using definition (10.15) of the surplus, we arrive at

$S_3(0, T, T) = y - (r - g)V_u(T)$. In consequence, the optimal lifespan of jobs is defined by the equality:[2]

$$y = (r - g)V_u(T) \tag{10.17}$$

This condition simply means that the employer and his employee have an interest in ending their relationship from the date at which, by looking for a new job, the worker will obtain a flow of gain $(r - g)V_u(T)$ greater than the flow of production y generated by the current job. Individual production y being an exogenous constant, and $V_u(T)$ being equal to $e^{gT}V_u(0)$, there exists a single value of T satisfying equation (10.17). Moreover, for this value of T, we find after several calculations that $S_{33}(0, T, T) = -gy < 0$. The marginal surplus due to an increase in the lifespan of the job at date T is thus indeed negative when this limit is extended.

1.3.2 THE BALANCED GROWTH PATH

It is possible to determine the equilibrium values of labor market tightness θ and the lifespan T of a job with the help of two relations that portray the conditions of job creation and job destruction.

Job Creation

The job creation equation results from free entry equilibrium, which indicates that the expected cost of a vacant job is equal to the expected gain of a filled one. Let us assume that the search costs arising from a vacant job increase at rate g, taking the form $hy(t)$ where h is a positive exogenous constant. At date t, the value $\Pi_v(t)$ of a vacant job will then be expressed as:

$$(r - g)\Pi_v(t) = hy(t) + m(\theta)\left[\Pi_e(t, 0, T) - \Pi_v(t)\right]$$

We obtain a relationship between T and θ, noting that in the context proper to this model, the free entry condition at $t = 0$, $\Pi_v(0) = 0$, entails that the expected profit $\Pi_e(0, 0, T)$ from a job created at date 0 must exactly cover the average cost $hy/m(\theta)$ of a vacant job posted at the same date $t = 0$. With the help of sharing rule (10.16), we will thus have $(1 - \gamma)S(0, 0, T) = hy/m(\theta)$. If we consider relation (10.15) at $t = 0$, and note that condition (10.17) characterizing the optimal lifespan of a job entails $V_u(0) = ye^{-gT}/(r - g)$, we arrive, after rearranging terms, at the following relation:

$$\frac{h}{m(\theta)} = \frac{(1 - \gamma)}{r}\left[1 + \frac{ge^{-rT} - re^{-gT}}{r - g}\right] \tag{10.18}$$

When $r > g$, it is easy to verify that the expected profit from a job at the time of its creation, represented by the right side of equation (10.18), rises with the lifespan of this job. As the average unit cost $h/m(\theta)$ is an increasing function of θ, equation (10.18) in sum defines an increasing relation between labor market tightness θ and the lifespan T

[2] We could, in like manner, have taken the view that the optimal lifespan of a job maximizes the surplus $S(0, t, T)$ for all $t \in [0, T]$. That would again give us relation (10.17).

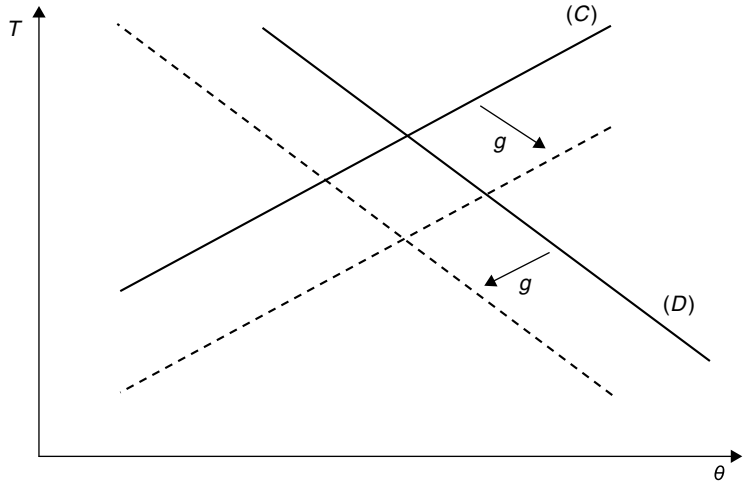


FIGURE 10.4
The equilibrium values of T and θ.

of a job which we can assimilate to a labor demand. We have identified it by the abbreviation (C) in figure 10.4. We can also verify that, for a given lifespan T, the expected profit from a new job increases with the rate of growth of productivity.[3] In figure 10.4, a rise in g shifts the (C) curve to the right.

For given T, that is, for a given job destruction rate $q = 1/T$, relation (10.18) is in fact analogous to relation (10.9) defining the equilibrium value of the labor market tightness θ in the previous model, where the rate of destruction q was exogenous. In the latter case, the capitalization effect entails that the profit expected from a filled job increases with g, and it is thus not surprising to find that θ rises with g for given T. In this model, however, the lifespan of jobs is an endogenous variable which, as we will prove below, *diminishes* with the growth rate g of productivity. In consequence, accelerated growth increases the destruction of jobs, running counter to the capitalization effect. The direction in which θ varies with g becomes a priori ambiguous. To overcome this ambiguity, we have to define the relationship between T and θ that corresponds to decisions to destroy jobs.

Job Destruction

We obtain a second relationship between θ and T using relation (10.11), which defines the expected utility of an unemployed person at instant $t = 0$, and applying the sharing rules (10.16). We thus get:

$$(r - g)V_u(0) = zy + \theta m(\theta)\frac{\gamma}{1-\gamma}\Pi_e(0, 0, T)$$

[3] Deriving equation (10.18), we note that $\partial\theta/\partial g$ is of the sign of $(r-g)rTe^{-gT} + r(e^{-rT} - e^{-gT})$. For given T, this expression amounts to zero if $r = g$. Moreover, the derivative of this expression with respect to g is equal to $-(r-g)T^2e^{-gT} < 0$ for $r > g$. In consequence $\partial\theta/\partial g$ is positive.

Following (10.17), $(r-g)V_u(0) = ye^{-gT}$, and since the expected profit $\Pi_e(0,0,T)$ is equal to the average cost $hy/m(\theta)$ we finally get:

$$e^{-gT} = z + \frac{\gamma h\theta}{1-\gamma} \tag{10.19}$$

This equation defines a decreasing relation between labor market tigthness θ and the lifespan of a job T. It is represented by the (D) curve in figure 10.4. Relation (10.19) indicates that high labor market tightness entails a strong exit rate from unemployment and a high expected utility for unemployed persons, which entails a weak surplus and consequently a shorter lifespan for jobs. We also see that an increase of g shifts this curve downward.

Equilibrium

Figure 10.4 shows that the lifespan of a job diminishes when growth accelerates. But the effect on θ is a priori ambiguous. In appendix 6.1 at the end of this chapter, however, we show that θ diminishes with the growth rate g of productivity, so an increase in g here lowers the exit rate $\theta m(\theta)$ from unemployment. When the labor force is constant, the unemployment rate is given by the formula:

$$u = \frac{q}{q+\theta m(\theta)} \quad \text{with } q = 1/T \tag{10.20}$$

Since an increase in g lowers the exit rate from unemployment and increases the rate q of job destruction, a stronger rise in productivity unambiguously increases unemployment.

In sum, technological progress increases the unemployment rate in this model with endogenous job destruction. But it must be understood that this result is not general. It follows from the fact that older jobs derive no benefit from technological progress and must necessarily be destroyed when they reach a certain age. This case is directly opposed to the one envisaged in the previous model, with exogenous destruction, in which all jobs benefit from technological progress, independently of the date at which they were created. Clearly an intermediate model, incorporating the two forms of technological progress, would show that technological progress is favorable to employment if and only if a sufficiently large share of technological progress is automatically incorporated into all jobs. The capitalization effect would then dominate the job destruction effect. From this perspective, Mortensen and Pissarides (1998) have built a model in which firms can overhaul jobs when their surplus becomes negative, at a certain cost. They then show that technological progress is favorable to employment if the costs of overhaul are slight, and unfavorable if they are not. Aghion and Howitt (1998, p. 129) present a model, similarly inspired, which yields similar results.

These analyses indicate that the impact of technological progress depends on the form it takes and the opportunities to reorganize available to firms. In this respect, it is important to know whether the market mechanisms at work in the previous model lead to an optimal reallocation of jobs.

1.3.3 The Efficiency of Creative Destruction

Under what circumstances is the restructuring caused by technological progress too rapid or, on the contrary, too slow? In a perfectly competitive economy, the answer to this question is evident: since the free play of competition leads to efficient allocations, the pace of technological progress is necessarily efficient too. In the presence of transaction costs in the labor market, the problem becomes thornier. Job destruction gives rise to reallocation unemployment, which may be thought to be socially inefficient. To answer this question, which has been studied by Caballero and Hammour (1996), it is necessary to characterize the social optimum—the values of labor market tightness, the unemployment rate, and the job destruction rate that maximize discounted aggregate production. For the sake of simplicity, we proceed as we did in chapter 9, leaving out preference for the present. In this model with growth, this hypothesis amounts to setting $r = g$. Moreover, we consider only stationary states.

The Planner's Problem

At date t, total output is equal to the sum of all the production achieved by all the jobs created between dates $t - T$ and t. As there are $\theta m(\theta)u$ jobs created at each date, and since a job created at date x produces $y(x) = ye^{gx}$, total production at date t is equal to $\int_{t-T}^{t} yu\theta m(\theta)e^{gx}\mathrm{d}x$. At this same date, unemployed persons produce $uzy(t) = uzye^{gt}$ and the cost of vacant jobs comes to $\theta uhy(t) = \theta uhye^{gt}$. Noting that $\int_{t-T}^{t} e^{gx}\mathrm{d}x = [e^{gT} - e^{g(t-T)}]/g$, aggregate production $\omega(t)$ at date t, equal by definition to the sum of all production minus the cost of vacant jobs, is therefore expressed as:

$$\omega(t) = ye^{gT}u\left[\theta m(\theta)\frac{1 - e^{-gT}}{g} + z - h\theta\right]$$

Following definition (10.20) of the stationary unemployment rate, we have $u = 1/[1 + T\theta m(\theta)]$, and the planner's problem can be written as:

$$\max_{(\theta,T)} \frac{1}{1 + T\theta m(\theta)}\left[\theta m(\theta)\frac{1 - e^{-gT}}{g} + z - h\theta\right]$$

Let us denote by $\eta(\theta) = -\theta m'(\theta)/m(\theta)$ the elasticity of the matching function with respect to the unemployment rate; after several calculations, we verify that the optimal values of labor market tightness, θ^*, and of the lifespan of jobs, T^*, are defined by the two following equations:

$$\frac{h}{m(\theta^*)} = [1 - \eta(\theta^*)]\left[\frac{1 - e^{-gT^*}}{g} - T^*e^{-gT^*}\right] \tag{10.21}$$

$$e^{-gT^*} = z + \frac{\eta(\theta^*)\theta^*h}{1 - \eta(\theta^*)} \tag{10.22}$$

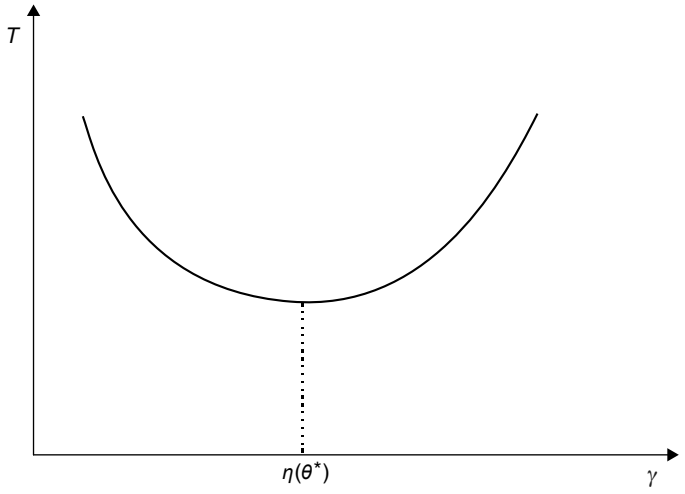


Figure 10.5
The relation between the lifespan of jobs and the bargaining power of workers.

We can compare the optimal values of labor market tightness and lifespan of jobs with those obtained at decentralized equilibrium by making r go to g in equation (10.18). In this configuration of the parameters, equation (10.18) is written:[4]

$$\frac{h}{m(\theta)} = (1-\gamma)\left[\frac{1-e^{-gT}}{g} - Te^{-gT}\right] \tag{10.23}$$

Comparison of the two systems of equations (10.19) and (10.23) on one hand, and (10.21) and (10.22) on the other, respectively defining decentralized equilibrium and the social optimum, shows that these two states coincide if and only if the Hosios condition $\gamma = \eta(\theta^*)$ is satisfied (see chapter 9, section 4, for more detail on this condition). Differentiating equations (10.19) and (10.23), we easily verify that the labor market tightness at decentralized equilibrium decreases with the bargaining power of workers, γ, and that the lifespan T of jobs reaches a minimum when $\gamma = \eta(\theta^*)$. The linkage between the lifespan of jobs and bargaining power is represented in figure 10.5.

Inefficiency and the Hosios Condition

We see that labor market tightness lies below its efficient level if and only if workers have bargaining power greater than $\eta(\theta^*)$. On the other hand, labor market equilibrium is always characterized by an *insufficient* reallocation of jobs when the Hosios condition is not met. This result, obtained by Caballero and Hammour (1996), suggests that the market imperfections resulting from an inefficient sharing of rents lead systematically

[4]Formula (10.23) is found by noting first that $ge^{-rT} - re^{-gT} = e^{-gT}[ge^{-(r-g)T} - r]$. In the neighborhood of 0, $e^{-(r-g)T}$ is equivalent to $1-(r-g)T$ and thus $ge^{-rT} - re^{-gT}$ is equivalent to $-(r-g)(1+gT)e^{-gT}$. When r tends to g, the bracketed expression in (10.18) is thus equal to $1 - e^{-gT} - gTe^{-gT}$.

to sclerosis in the process of job reorganization. We can understand this by going back to relation (10.17), which defines the optimal lifespan of jobs as a function of the expected utility of unemployed persons. As in the basic model of chapter 9, it is easy to verify here that the expected utility of unemployed persons reaches a maximum when the Hosios condition is satisfied. Relation (10.17) does indeed entail $(r-g)V_u(t) = e^{-g(T-t)}y$ on a balanced growth path, and since T reaches a minimum when $\gamma = \eta(\theta^*)$, the expected utility of unemployed persons is indeed maximal for $\gamma = \eta(\theta^*)$. Now the greater the expected utility of unemployed persons is, the less surplus a job generates—see relation (10.14)—and therefore the higher the gains of unemployed persons are, the shorter the lifespans of jobs. In sum, the insufficient reallocation of jobs in decentralized markets results from a very simple logic: when the labor market is inefficient, the gains from searching for a job are relatively slight, which tends to increase the rent of individuals holding jobs and thus gives them an incentive to keep their jobs as long as possible. Conversely, labor market efficiency ensures a maximal return to job search and produces a maximum of incentive to reorganize production units.

These results are obviously pertinent to economic policy. They suggest that measures to protect employment are ill suited to countering the effects of technological progress on unemployment. The model with endogenous job destruction, which in the simple form in which we have studied it in this section represents a situation where technological progress is embodied only in new jobs, indicates that more rapid growth increases unemployment. It also allows us to show that this source of unemployment ought not to be combated by putting in place measures to protect jobs. Caballero and Hammour (1996) suggest instead using subsidies to create employment. With this type of measure, market equilibrium can indeed be made to coincide with the social optimum. In our model, the values of labor market tightness and the job destruction rate defined by the systems (10.19) and (10.23), and (10.21) and (10.22) are identical if entrepreneurs receive a subsidy amounting to $h[\gamma - \eta(\theta^*)] / [1 - \eta(\theta^*)]$ per unit of time for each vacant job. The subsidy is thus positive if the bargaining power of workers is greater than the elasticity of the matching function with respect to the unemployment rate, and negative if not.

2 TECHNOLOGICAL PROGRESS AND INEQUALITY

What is the impact of technological progress on wage inequality and employment opportunity? For instance, computer-based information technologies have probably favored some categories of workers at the expense of others. So technological progress may be said to have been *biased* in favor of those with skills. This explanation of the evolution of wage inequality provided the impetus for a large body of research in the 1990s, summarized especially in Card and DiNardo (2002), Goldin and Katz (2008), and Acemoglu and Autor (2011). We begin by showing how technological progress influences wage inequality between workers with different skill levels. We will then see how theoretical and empirical works suggest that technological progress has played an important part in the development of inequality in the main OECD countries during the last two decades of the twentieth century and in the beginning of the twenty-first century. And finally, we will highlight the fact that the form taken by technological progress is not independent

of the structures of incentives within which agents act over time and that it might in part be determined by the composition of the labor force.

2.1 Facts on Wages and Occupations

Technological progress does not just affect the rate of global unemployment and the average wage. It also influences the distribution of the employment opportunities offered to different types of individual. So, technological progress alters the return to certain kinds of skills and educational investment. It also affects the occupational structure.

2.1.1 Wage Inequality

The end of the twentieth century constitutes a particularly interesting period for the analysis of inequality. Over this period the reward to workers of different skills changed a great deal in the industrialized countries of the OECD. This phenomenon is illustrated by figure 10.6, which describes changes in hourly wages for men in the United States from 1976 to 2012. This figure represents the changes in (real) hourly wages for different subperiods. For instance, the curve with diamonds indicates that the wages in the 20th quantile of the wage distribution decreased by about 15% between 1976 and 1990. More generally, this figure highlights a deformation of the wage distribution detrimental to low wages, which declined strongly over this period. High wages, above the 90th percentile, increased. Figure 10.7 shows that the same phenomenon arises for women. We will see that this trend was caused by the conjunction of interdependent elements. Technological progress and competition from low-wage countries contributed, in varying and

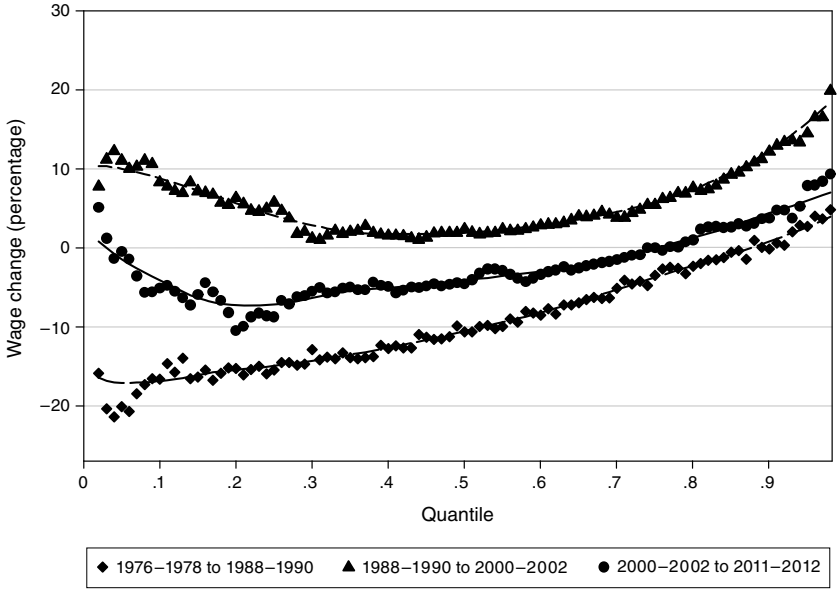


Figure 10.6

Changes in real hourly wages for males in the United States over the period 1976–2012.

Source: Firpo et al. (2011) data set.

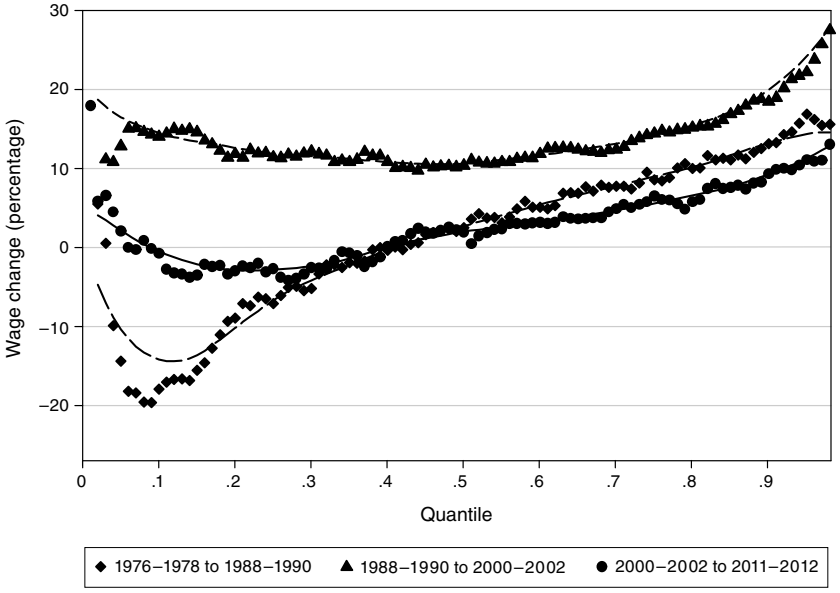


FIGURE 10.7
Changes in real hourly wages for females in the United States over the period 1976–2012.

Source: Firpo et al. (2011) data set.

much-debated degrees, to this increase of inequality. International migration, the evolution of labor market institutions, and certain organizational changes have also played a role, although probably a more marginal one. In different cases, this change led to a widened spread of earnings or a widened spread of unemployment rates across categories.

Figures 10.6 and 10.7 show that the wage distribution changed differently over subsequent periods. From 1990 to 2000, wages at the bottom of the distribution decreased less strongly than in the previous period and increased at a similar rate as that of wages in the middle of the distribution. Then, in the first decade of the twenty-first century, wages at the bottom and at the top of the distribution increased faster than those in the middle of the distribution. This is the phenomenon of "wage polarization" described by Autor, Levy, and Murnane (2003) for the United States, by Goos and Manning (2007) for Britain, by Dustmann, Ludsteck, and Schönberg (2009) for Germany, and by Goos, Manning, and Salomons (2009, 2014) for European countries.

2.1.2 CHANGES IN THE OCCUPATIONAL STRUCTURE

Changes in wage distribution are accompanied by changes in the occupational structure. This is illustrated by figure 10.8, borrowed from Acemoglu and Autor (2011), which plots changes over each of the last three decades from 1980 to 2010 in the share of U.S. employment accounted for by 318 detailed occupations encompassing all of U.S. employment. These occupations are ranked on the horizontal axis by their skill level

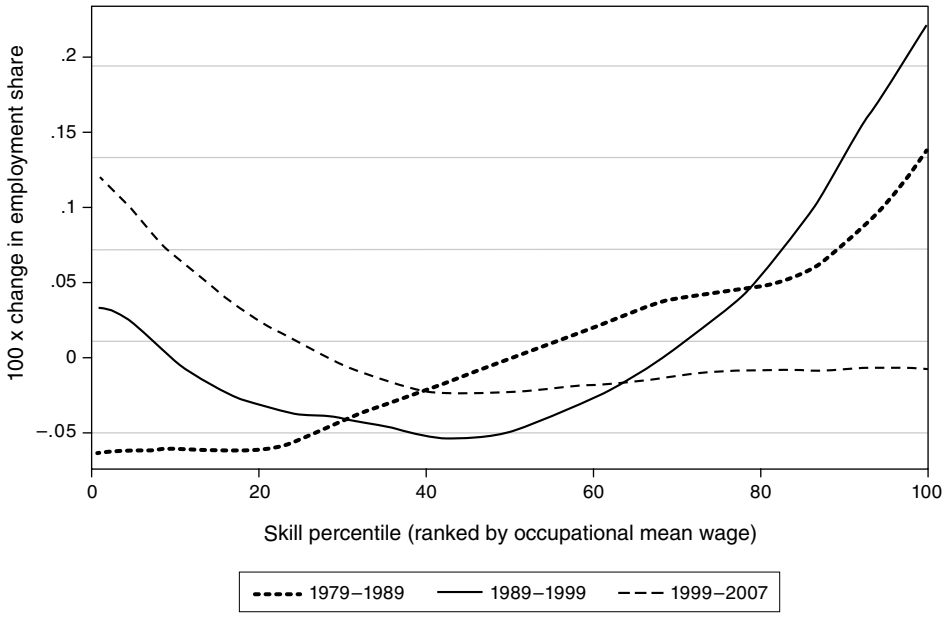


Figure 10.8
Changes in employment by occupational skill percentile. All occupation and earnings measures in these samples refer to prior year's employment. The figure plots log changes in employment shares by 1980 occupational skill percentile rank using a locally weighted smoothing regression, where skill percentiles are measured as the employment-weighted percentile rank of an occupation's mean log wage.

Source: Acemoglu and Autor (2011, figure 10).

from lowest to highest, where an occupation's skill rank is approximated by the average wage of workers in the occupation in 1980. It is apparent that the evolution of the occupational skill structure is similar to that of wage distribution. During the 1980s, there is a monotonous increasing relation between the growth of jobs and their skill contents. Then, during the 1990s, the share of jobs with low skill content began to increase. Job polarization began to emerge: the share of jobs with an intermediate level of skill grew at the slowest rate, while the number of jobs with low skill content grew faster. Then, job polarization continued during the 2000s, but with a higher increase in the jobs with the lowest skill content. All in all, figure 10.8 shows that the number of jobs with intermediate skill content has grown less than the other types of jobs since the beginning of the 1990s.

2.2 A Model with Skills and Tasks

In this section we present a model representing the impact of technological progress on the structure of jobs and on wages. In this model, workers have skills that enable them to accomplish tasks of different kinds. To each job there corresponds a set of particular tasks that can be carried out by workers with different skills. For example, the tasks that correspond to the job of house painter consist essentially of spreading paint over surfaces. These tasks can be carried out by professional painters. They could also be

carried out by economists, who are generally less efficient than professional painters at such tasks and who prefer to hold jobs as economists, for which they are generally better paid. Nonetheless, if technological progress were to make it possible to replace economists by computers, the remuneration available to economists would drop, and many would choose to become house painters instead.

2.2.1 The Assignment Model

The assignment model between jobs and skills allows us to study the impact of technological progress on the structure of jobs and on wages. It consists of an extension of the assignment model presented in chapter 3, in connection with the formation of the remuneration paid to superstars and CEOs (chief executive officers) of large firms, a model developed by Tinbergen (1974), Rosen (1974), Sattinger (1975), and Teulings (1995, 2005). The application of this model to the study of technological progress and wage inequalities was developed especially by Saint-Paul (2001, 2008), Garicano and Rossi-Hansberg (2006), Acemoglu and Autor (2011), as well as by Costinot and Vogel (2010), whose contribution we present below in a simplified version.

Preferences and Technology

The setup is this: an economy composed of a continuum of workers who each offer a unit of labor if the wage is positive. Each worker is endowed with a skill s falling in the interval or spread $[0,1]$. The apportionment of skills over this interval is described by a probability distribution $\phi > 0$, hence $\phi(s)\mathrm{d}s$ represents the supply of labor with skills falling in the interval $[s, s+\mathrm{d}s]$.

There is one final good, which is the numéraire and which is produced using a continuum of intermediate goods thanks to the following technology with constant returns:

$$Y = \int_{\underline{\sigma}}^{\bar{\sigma}} B(\sigma) Y(\sigma) \mathrm{d}\sigma$$

where Y designates the output of the final good and $Y(\sigma)$ represents the input of an intermediate good of type σ. Parameter σ is a real number representing a "task" performed by the workers. The tasks are ranked by increasing complexity over the interval $[\underline{\sigma}, \bar{\sigma}]$. The boundaries of this interval are endogenous. Quantity $B(\sigma) > 0$ designates an exogenous parameter linked to technological progress. It is also possible to assume that the economy comprises consumption goods $Y(\sigma)$ with a representative consumer whose preferences are portrayed by the linear utility function Y. In this interpretation, parameter $B(\sigma)$ reflects the preferences of the consumer.

Producing a quantity $Y(\sigma)$ of intermediate goods (or tasks) requires only workers. A person of skill s working in the production of good σ has a productivity equal to $A(s,\sigma)$. Workers are assumed to be perfect substitutes in the production of each task:

$$Y(\sigma) = \int_0^1 A(s,\sigma) L(s,\sigma) \mathrm{d}s, \quad \text{for all } \sigma$$

where $L(s,\sigma) \geq 0$ is the endogenous "number" of workers with skill s performing task σ.

Function $A(s,\sigma)$ marks a break with the manner in which production functions are usually represented, since it distinguishes between tasks and skills. Hitherto, the assumption was that there existed an exogenous one-to-one mapping between skills and tasks: a painter spreads paint and a professor of economics teaches courses in economics. Here the assumption is that a worker of a given skill level can perform a *variety* of tasks. A painter can paint, but he could also teach economics (why not?); likewise a professor of economics can teach, but he could also paint the walls of his university buildings. The point of representing the situation this way is that it allows us to explain the matchup between skills and tasks and thus to gain an understanding of how technological progress modifies this matchup.

The hypotheses made about function $A(s,\sigma)$ play a crucial role. In chapter 3, in a similar environment, we assumed that function $A(s,\sigma)$ was supermodular, which would amount in this case to hypothesizing that $A_{s\sigma} > 0$ ($A_{s\sigma}$ designates the partial second derivative of function A with respect to s and σ). This hypothesis means that function A_σ is increasing with skill s for every task σ, in which case more skilled workers are more productive than less skilled workers at all tasks. Such a hypothesis is not well suited to a description of technology in terms of skills and tasks, for it is possible, for example, that less-skilled workers are more productive than more-skilled workers at simple tasks. That is why the assumption made here is that more-skilled workers have a comparative advantage in more complex tasks, or, formally:

$$\frac{A(s_2,\sigma_2)}{A(s_2,\sigma_1)} > \frac{A(s_1,\sigma_2)}{A(s_1,\sigma_1)} \quad \text{for all } s_2 > s_1 \quad \text{and} \quad \sigma_2 > \sigma_1 \tag{10.24}$$

This assumption is also known as strict log supermodularity of function $A(s,\sigma)$, because it requires that function A_σ/A is increasing with skill s, as shown by the inequality:[5]

$$A_{s\sigma}A - A_s A_\sigma > 0 \quad \text{for all } s \quad \text{and} \quad \sigma \tag{10.25}$$

In the discussion of this model that follows, we will use the function:

$$A(s,\sigma) = \alpha s - \sigma, \alpha > 0 \tag{10.26}$$

This function is log supermodular, since it verifies condition (10.25). It also possesses the following properties: for a given skill level, production is less efficient when tasks grow more complex ($A_\sigma < 0$); more-skilled persons are relatively more efficient at performing more complex tasks ($A_s > 0$).

[5] To show this, let us remark that, subtracting 1 from each side of equation (10.24) and dividing by $(\sigma_2 - \sigma_1)$, we find that equation (10.24) is equivalent to:

$$\frac{A(s_2,\sigma_2) - A(s_2,\sigma_1)}{(\sigma_2 - \sigma_1)} \frac{1}{A(s_2,\sigma_1)} > \frac{A(s_1,\sigma_2) - A(s_1,\sigma_1)}{(\sigma_2 - \sigma_1)} \frac{1}{A(s_1,\sigma_1)} \quad \text{for all } s_2 > s_1 \quad \text{and} \quad \sigma_2 > \sigma_1$$

For $\sigma_2 \to \sigma_1$ this inequality can be written:

$$\frac{A_\sigma(s_2,\sigma_1)}{A(s_2,\sigma_1)} > \frac{A_\sigma(s_1,\sigma_1)}{A(s_1,\sigma_1)} \quad \text{for all } s_2 > s_1$$

where A_i denotes the partial derivative of function A with respect to variable i. This condition is satisfied only if $\frac{A_\sigma(s,\sigma)}{A(s,\sigma)}$ is a strictly increasing function of s, which is equivalent to $\partial^2 \log A(s,\sigma)/\partial s \partial \sigma > 0$.

The Behavior of Firms

Let $p(\sigma)$ denote the price of the intermediate good σ. Profit in the final good industry is then written:

$$\Pi = Y - \int_{\underline{\sigma}}^{\bar{\sigma}} p(\sigma) Y(\sigma) \mathrm{d}\sigma = \int_{\underline{\sigma}}^{\bar{\sigma}} [B(\sigma) - p(\sigma)] \, Y(\sigma) \mathrm{d}\sigma$$

All markets are in a state of perfect competition, meaning primarily that there is free entry to all markets and that agents take prices as given. In this context, profits are null and the maximization of profit with respect to $Y(\sigma)$ entails that $p(\sigma) = B(\sigma)$.

As markets are perfectly competitive, workers with the same skill s all necessarily obtain the same wage, denoted $w(s)$. The functions of wage $w(s)$ are regarded as given by firms. Profit in the firm producing intermediate good σ is written:

$$\Pi(\sigma) = p(\sigma) Y(\sigma) - \int_0^1 w(s) L(s, \sigma) \mathrm{d}s = \int_0^1 [p(\sigma) A(s, \sigma) - w(s)] \, L(s, \sigma) \mathrm{d}s$$

This firm chooses the skill intensity s and the corresponding number of workers $L(s, \sigma)$ that maximize its profit. As $p(\sigma) = B(\sigma)$, the firm's problem σ takes the form:

$$\max_{s, \, L(s,\sigma)} \Pi(\sigma) = \int_0^1 [B(\sigma) A(s, \sigma) - w(s)] \, L(s, \sigma) \mathrm{d}s$$

Deriving profit $\Pi(\sigma)$ with respect to $L(s, \sigma)$, we get:

$$B(\sigma) A(s, \sigma) - w(s) = 0 \tag{10.27}$$

We have reverted to the classic condition of null profit (or free entry) for technologies with constant returns.

The first-order condition of the profit maximization program with respect to s takes this expression:

$$B(\sigma) A_s(s, \sigma) - w'(s) = 0 \tag{10.28}$$

and the second-order condition is verified if:

$$B(\sigma) A_{ss}(s, \sigma) - w''(s) < 0 \tag{10.29}$$

The first-order condition shows that the firm which produces the intermediate good σ chooses the skill intensity s for which the marginal gain from an increase in s, or $B(\sigma) A_s(s, \sigma)$, equals the marginal cost, $w'(s)$, corresponding to the variation in wages needed to boost skill intensity.

Equilibrium

Let us denote by $M(\sigma)$ the assignment (or the matching) function which defines the skill linked to task σ. In accordance with (10.27) and (10.28), at competitive equilibrium, this matching function verifies the two following relations:

$$B(\sigma)A\left[M(\sigma),\sigma\right]-w\left[M(\sigma)\right]=0 \tag{10.30}$$

$$B(\sigma)A_s\left[M(\sigma),\sigma\right]-w'\left[M(\sigma)\right]=0 \tag{10.31}$$

Deriving (10.30) with respect to σ, we arrive (with the obvious notations) at $B'A+BA_\sigma+M'(BA_s-w')=0$. With (10.31) this gives: $B'A+BA_\sigma=0$. The matching function is thus completely defined by equation:

$$\frac{A_\sigma\left[M(\sigma),\sigma\right]}{A\left[M(\sigma),\sigma\right]}=-\frac{B'(\sigma)}{B(\sigma)} \tag{10.32}$$

When the hypothesis of log supermodularity is satisfied, it is possible to show that this relation adequately defines a function M that links to each task σ a single skill $s=M(\sigma)$. As well, this function is increasing, which means that there is positive assortative matching: the more-skilled workers are matched to the more complex tasks. These two general results are set out in full in the appendix to this chapter, section 6.2. We illustrate them here by assuming that the production function of the intermediate goods is defined by equation (10.26), or $A(s,\sigma)=\alpha s-\sigma$, and that the technological parameter takes the form $B(\sigma)=b\sigma^\beta, b>0, \beta>0$. That being the case, condition (10.32) entails:

$$M(\sigma)=\frac{1+\beta}{\beta\alpha}\sigma \tag{10.33}$$

This relation shows that the matching function is indeed increasing: more-skilled persons carry out more complex tasks. The interval (or spread) of tasks is itself determined by the spread of workers' skills and by technological parameters: since each worker offers a unit of labor inelastically within the skill spread $[0,1]$, the lower and upper boundaries of the tasks utilized by firms are defined respectively by $\underline{\sigma}=M^{-1}(0)=0$ and $\bar{\sigma}=M^{-1}(1)=\beta\alpha/(1+\beta)$.

Knowledge of the matching between skills and tasks permits us to calculate the production of each intermediate good, as well as wages. The production of an intermediate good utilizing task σ produced by workers of skill $s=M(\sigma)$ is equal to the production of each worker of type s multiplied by the mass of workers of skill s, or:

$$A\left[s,M^{-1}(s)\right]\phi(s)=\frac{\alpha s}{1+\beta}\phi(s)$$

Following condition (10.30), the wage of workers of skill s is given by:

$$w(s)=B\left[M^{-1}(s)\right]A\left[s,M^{-1}(s)\right]$$

After several simple calculations, we arrive at:

$$w(s) = \frac{b}{\beta}\left(\frac{\alpha\beta}{1+\beta}s\right)^{1+\beta} \tag{10.34}$$

As the reader will see, the wage rises with the level of skill. Hence more-skilled workers carry out more complex tasks that are rewarded by higher wages.

The Social Optimality of Technological Choices

To grasp fully how this model functions, let us study the options facing a social planner whose problem is to assign workers of different skills to a range of tasks in order to maximize aggregate welfare, equal to aggregate production Y. This will allow us to understand clearly the consequences of matching skills to tasks. It will also allow us to observe that the matching of skill to tasks determined by competitive equilibrium is socially efficient.

The planner knows that any person of skill $s \in [0,1]$ working in the production of good σ has a productivity equal to $A(s,\sigma)$. If the planner decides to assign $L(s,\sigma)$ workers of skill s to task σ, the production of good σ is equal to:

$$Y(\sigma) = \int_0^1 A(s,\sigma)L(s,\sigma)\mathrm{d}s, \quad \text{for all } \sigma$$

The total product then attains the value

$$Y = \int_{-\infty}^{+\infty} B(\sigma)Y(\sigma)\mathrm{d}\sigma = \int_{-\infty}^{+\infty}\int_0^1 B(\sigma)A(s,\sigma)L(s,\sigma)\mathrm{d}s\mathrm{d}\sigma$$

Since we assume that working time is indivisible and that each worker can be assigned to just one task, the planner must choose a function that to each value of s assigns a task $\sigma = m(s)$. She must also choose the number of workers $L[s,m(s)]$ that will be assigned to task $\sigma = m(s)$. As the planner has an obvious interest in utilizing the whole available labor force, at the optimum we have $L[s,m(s)] = \phi(s)$. Total production may thus be written:

$$Y = \int_0^1 \left[\int_{-\infty}^{+\infty} B(\sigma)A(s,\sigma)\mathrm{d}\sigma\right]\phi(s)\mathrm{d}s$$

This form makes it clear that function $m(s)$ corresponds to the value of σ that maximizes the quantity $B(\sigma)A(s,\sigma)$ for a given s. The derivation of $B(\sigma)A(s,\sigma)$ with respect to σ yields the relation (10.32), which characterizes the matching function at market equilibrium. Competitive equilibrium is thus indeed socially efficient. Relation (10.32) may also be written:

$$A_\sigma(s,m(s))B(m(s)) + B'(m(s))A(s,m(s)) = 0$$

This formula is easy to interpret by considering our example where $A(s,\sigma) = \alpha s - \sigma$. We have here a situation in which the productivity of a worker with a given skill level drops

when the complexity of the tasks he is called upon to perform rises. Correspondingly, the value added to intermediate goods produced using more complex tasks is greater, since $B(\sigma) = b\sigma^{\beta}$ rises with σ. In this case, for each skill level, the planner increases the complexity of tasks to the point where the marginal cost, represented here by the term $-A_{\sigma}(s,\sigma)B(\sigma)$, equals the marginal gain $B'(\sigma)A(s,\sigma)$.

2.2.2 The Consequences of Technological Changes

Technological progress induces changes in the technological schedule B. Such changes modify the assignment of skills to tasks and then the equilibrium wage distribution. We consider two different changes that allow us to explain the increase in the relative wages of highly skilled workers on one hand, and the phenomenon of wage and job polarization on the other. These results are arrived at by adopting particular forms for functions $B(\sigma)$ and $A(s,\sigma)$. In appendix 6.2 we show that they are in fact general.

The Increase in Relative Wages of Highly Skilled Workers

We consider a biased technological progress: the more skilled a worker is, the more it boosts her productivity. For this purpose, let us define the relative productivity of a schedule B_2 with respect to a schedule B_1 for the task σ by the ratio $B_2(\sigma)/B_1(\sigma)$. We say that B_2 is skill biased relative to B_1 if the ratio $B_2(\sigma)/B_1(\sigma)$ is an increasing function of σ. This definition captures the idea that skill-biased technological change increases the relative productivity of tasks with high skill intensity. When function $B_2(\sigma)/B_1(\sigma)$ is increasing, its derivative is positive, or in formal terms:

$$\frac{B_2'(\sigma)}{B_2(\sigma)} \geq \frac{B_1'(\sigma)}{B_1(\sigma)} \quad \text{for all } \sigma \tag{10.35}$$

If we assume that $B_i(\sigma) = b_i\sigma^{\beta_i}$, this condition is equivalent to $\beta_2 > \beta_1$. The analysis of the impact of biased technological progress in favor of skilled workers then comes down to the study of the impact of an increase in β on the equilibrium of the assignment model. Matching function (10.33) shows that workers of skill s are assigned to tasks of intensity $\sigma = [\beta\alpha/(1+\beta)]\, s$. We see that an increase in β leads to workers being assigned to more complex tasks, since the derivative of σ with respect to β is positive for given s. The increase in β also modifies the spread of tasks lying between $\underline{\sigma} = 0$ and $\bar{\sigma} = \beta\alpha/(1+\beta)$: it leads to the utilization of more complex tasks.

For that matter, equation (10.34) entails:

$$\frac{w'(s)}{w(s)} = \frac{1+\beta}{s}$$

If we designate by $w_i(s)$ the wage linked to value β_i of the parameter, we immediately deduce:

$$\frac{w_2'(s)}{w_2(s)} > \frac{w_1'(s)}{w_1(s)} \quad \text{for all } s$$

This inequality signifies that the ratio $w_2(s)/w_1(s)$ is an increasing function of s. That being the case, when the technology becomes more skill biased—when β

increases—the wage of an individual with few skills rises less in relative terms than the wage of a highly skilled one. Accordingly, biased technological progress increasing the relative productivity of more skill-intensive tasks has a positive impact on inequalities.

Job and Wage Polarization

The assignment model with tasks and skills also helps to shed light on the phenomenon of job polarization, which corresponds to a drop in relative wages in the middle of the wage distribution. Such a phenomenon can be explained by a biased technological progress that decreases the relative productivity of tasks with intermediate levels of skill intensity. Following Costinot and Vogel (2010), let us consider a shift from schedule B_1 to schedule B_2 such that:

$$\frac{B_2'(\sigma)}{B_2(\sigma)} \geq \frac{B_1'(\sigma)}{B_1(\sigma)} \quad \text{for } \sigma > \hat{\sigma} \quad \text{and} \quad \frac{B_2'(\sigma)}{B_2(\sigma)} \leq \frac{B_1'(\sigma)}{B_1(\sigma)} \quad \text{for } \sigma < \hat{\sigma} \tag{10.36}$$

In these inequalities $\hat{\sigma}$ is an exogenous threshold such that the shift from B_1 to B_2 increases the relative demand for tasks with low skill intensities over the range $\sigma < \hat{\sigma}$ and increases the relative demand for tasks with high skill intensities over the range $\sigma > \hat{\sigma}$.

It is possible to represent this type of technological progress by assuming that the technological parameter takes the expression $B_i(\sigma) = (\sigma + b\beta_i - \hat{\sigma})^{\beta_i}$, with $\beta_i > 0, b > 0$. That being the case, a modification of the technology that increases β_i corresponds to a biased technological progress like that portrayed in equation (10.36). So in order to see the impact, we need only examine the impact of an increase in β on the equilibrium model. With this definition of the technological parameter $B(\sigma)$, equations (10.26) and (10.32) permit us to find the expression of the matching function. Inverting this function, we observe that workers of skill s are assigned to the task σ defined by:

$$\sigma = \frac{\beta(\alpha s - b) + \hat{\sigma}}{1 + \beta}$$

It is easily verified that the derivative of σ with respect to β is of the sign of $\alpha s - b - \hat{\sigma}$. In consequence, an increase in β entails that highly skilled workers, above threshold $(b + \hat{\sigma})/\alpha$, are assigned to more complex tasks, while workers with skills below this threshold are assigned to less complex tasks. Hence increasing β induces workers to reallocate out of intermediate tasks and toward more extreme tasks. Workers move away from tasks that use intermediate skill intensity to concentrate on tasks with low and high skill intensities. Such a reallocation corresponds to the phenomenon of job polarization that we described in figure 10.8. This phenomenon of job polarization arises from the fact that when β increases, the price of the intermediate goods, $p(\sigma) = B(\sigma)$, produced through tasks of high or low complexity, rises relative to the price of intermediate goods produced through tasks of intermediate complexity. Workers specialize in tasks the relative productivity of which, and thus the relative price of which, is increasing.

Job polarization is accompanied by wage polarization. We can observe this by remarking that wage increase as a function of skill may be written, using equations (10.30) and (10.31):

$$\frac{w'(s)}{w(s)} = \frac{A_s\left[s, M^{-1}(s)\right]}{A\left[s, M^{-1}(s)\right]}$$

which after several simple calculations yields:

$$\frac{w'(s)}{w(s)} = \frac{\alpha(1+\beta)}{\alpha s + b\beta - \hat{\sigma}}$$

This expression shows that the derivative with respect to β of the rate of wage increase according to skill is of the sign of $\alpha s - b - \hat{\sigma}$. Consequently, an increase of β causes the rate of wage growth with respect to skill for skills lying below the threshold $(b+\hat{\sigma})/\alpha$ to diminish, whereas it augments the rate of wage rise with respect to skill for skills lying above this threshold. Formally, we thus have:

$$\frac{w_2'(s)}{w_2(s)} > \frac{w_1'(s)}{w_1(s)} \quad \text{for } s > \frac{b+\hat{\sigma}}{\alpha} \quad \text{and} \quad \frac{w_2'(s)}{w_2(s)} < \frac{w_1'(s)}{w_1(s)} \quad \text{for } s < \frac{b+\hat{\sigma}}{\alpha}$$

These inequalities indicate that the ratio $w_2(s)/w_1(s)$ is an increasing (or decreasing) function of skill s when this skill lies above (or below) threshold $(b+\hat{\sigma})/\alpha$. That being the case, when the technology becomes more skill biased—when β increases—the wage of a low-skilled person—a person with skills lying below threshold $(b+\hat{\sigma})/\alpha$—increases relatively less than the wage of a more highly skilled person—one with skills lying above threshold $(b+\hat{\sigma})/\alpha$. This is a characteristic of wage polarization described in figures 10.6 and 10.7.

2.3 What Empirical Research Tells Us

A number of studies have attempted to estimate the technological progress bias. It is possible to distinguish two strands of research. The first strand uses a simple supply-and-demand model to show that the increase in wage inequality between high- and low-skilled workers is compatible with a biased technological progress. We illustrate this approach by presenting the analysis of Katz and Murphy (1992). Data and programs provided by Acemoglu and Autor (2011), available at www.labor-economics.org, allow us to apply this approach to the United States over the period 1963–2008. The second research strand is more focused on the impact of technological changes on occupational structure and wages.

2.3.1 Wage Inequality Between High- and Low-Skilled Workers

Much research on the increased inequality between high-skilled and low-skilled workers has sought the reason for this increase in the interaction between the evolution of technological progress, biased in favor of high-skilled workers, and the evolution of the labor they supply, related to opportunities to invest in education. Figure 10.9 displays

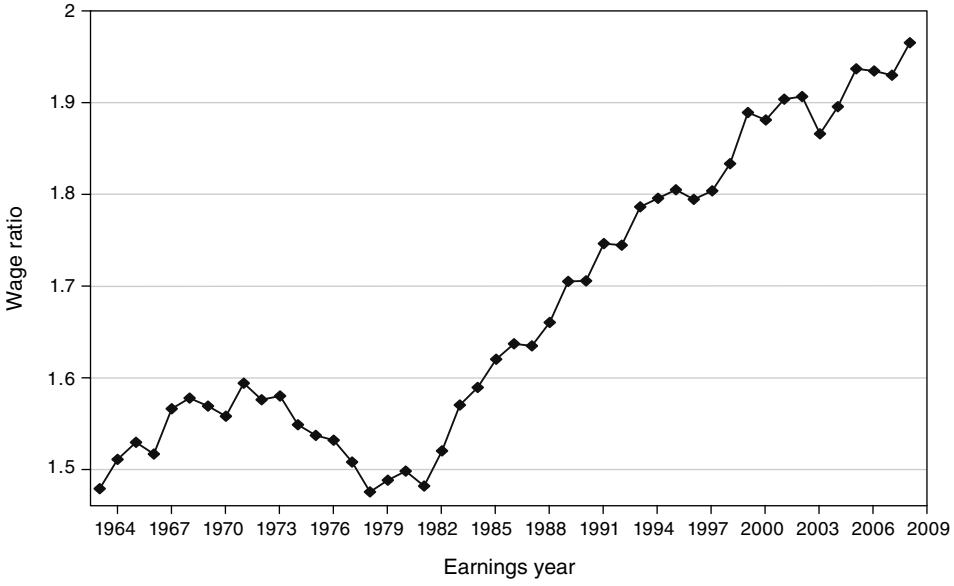


FIGURE 10.9
College/high school weekly wage ratio in the United States, 1963–2008.

Source: Acemoglu and Autor (2011) data set.

the evolution of the ratio between the average wage of workers with, at minimum, a college degree (at least 16 years of schooling) and that of high school graduates (12 years of schooling) from 1963 to 2008. It is apparent that this ratio was similar in 1963 and 1982, approximately equal to 1.5. However, after 1982, there is a strong positive trend driving this ratio up to about 2 in 2008.

Estimating the Technological Progress Bias

To account for the evolution of the ratio of wages represented in figure 10.9, a number of studies that follow on from the contribution of Katz and Murphy (1992) limit themselves to two categories of jobs and workers. Katz and Murphy consider a simplified version of the assignment model presented above, in which there are two kinds of job, skilled and unskilled, and two corresponding kinds of worker. In this version, the skilled workers, in quantity L_h, can hold only skilled jobs. They produce an intermediate good in quantity Y_h with a linear technology $Y_h = L_h$. The unskilled workers hold unskilled jobs and produce an intermediate good in quantity Y_ℓ with technology $Y_\ell = L_\ell$.

There is a final good produced out of the intermediate goods following a production function of the CES type:

$$Y = \left[B_h (L_h)^{\frac{\varepsilon-1}{\varepsilon}} + B_\ell (L_\ell)^{\frac{\varepsilon-1}{\varepsilon}} \right]^{\frac{\varepsilon}{\varepsilon-1}} \tag{10.37}$$

where B_h and B_ℓ are positive parameters that affect the productivity of each kind of labor and $\varepsilon > 0$ is the elasticity of substitution between skilled and unskilled labor. With

this CES production function, it is easy to verify that the relative demand $\lambda = L_h/L_\ell$ for skilled labor is given by the formula:

$$\lambda = \left(\frac{\omega}{\beta}\right)^{-\varepsilon} \tag{10.38}$$

where $\beta = B_h/B_\ell$ and $\omega = w_h/w_\ell$. This simple relationship between the relative demand for skilled labor, the wage differential, and the technological progress bias can be used to evaluate the technological progress bias.

Let us assume that the labor market is competitive, so that parameter λ is equal to the relative supply of skilled workers, $N_h/N_\ell = \nu$, since equality of labor supply and demand entails, $L_h/L_\ell = N_h/N_\ell$. Then, let us write equation (10.38) in logarithm, as follows:

$$\ln \omega = -\frac{1}{\varepsilon} \ln \nu + \ln \beta \tag{10.39}$$

With data for ν and for ω at our disposal, we can then estimate the elasticity of substitution ε and the technological progress bias β. Katz and Murphy (1992) estimated such a relation for the United States over the period 1963–1987. The dependent variable ω represents the ratio between the average wage of workers with, at minimum, a college degree (at least 16 years of schooling) and that of high school graduates (12 years of schooling). Running the same regression as Katz and Murphy over the period 1963–2008 using the data provided by Acemoglu and Autor (2011), we get:

$$\ln \omega = \underset{(0.043)}{-0.339} \ln \nu + \underset{(0.001)}{0.016} t - 0.059, \qquad R^2 = 0.93 \tag{10.40}$$

In this equation, t designates a time trend and R^2 is the adjusted coefficient of determination, and the figures in parentheses designate the standard errors of the coefficients. These results allow us, first of all, to give an estimate of the elasticity of substitution, $\varepsilon = 1/0.339 \simeq 2.9$. A number of studies carried out on various OECD countries using similar methodology obtain results closely similar, with an elasticity of substitution lying between one and three (see Goldin and Katz, 2008, and Acemoglu and Autor, 2011). In this setting, the time trend can be interpreted as the effect of technological progress on wage differences. The positive coefficient linked to the trend then signifies that the technological progress is skill biased: it increases the relative wage of the most highly skilled workers. Figure 10.10 shows that this approach provides a good explanation of the increase in the wage ratio between high-skilled and low-skilled workers. This figure displays the Katz-Murphy prediction of the college/high school wage ratio and the actual wage ratio over the period 1963–2008. It turns out that the Katz-Murphy assumption of a constant trend in technological bias allows us to predict fairly well the stability in the college/high school wage ratio until 1982, and its steady increase afterwards. All these elements point to the conclusion that technological bias may have played an important role in reshaping the demand for labor between high-skilled workers and low-skilled workers.

However, according to this model, the increase in inequality between high-skilled and low-skilled workers from the beginning of the 1980s is induced by an insufficient

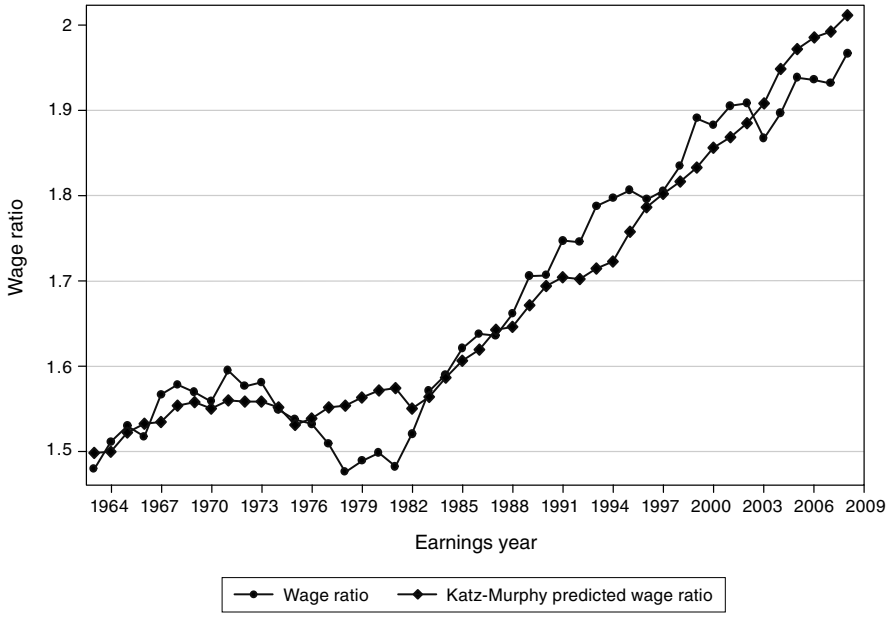


FIGURE 10.10
Predicted and observed college/high school wage ratio in the United States over the period 1963–2008.

Source: Acemoglu and Autor (2011) data set.

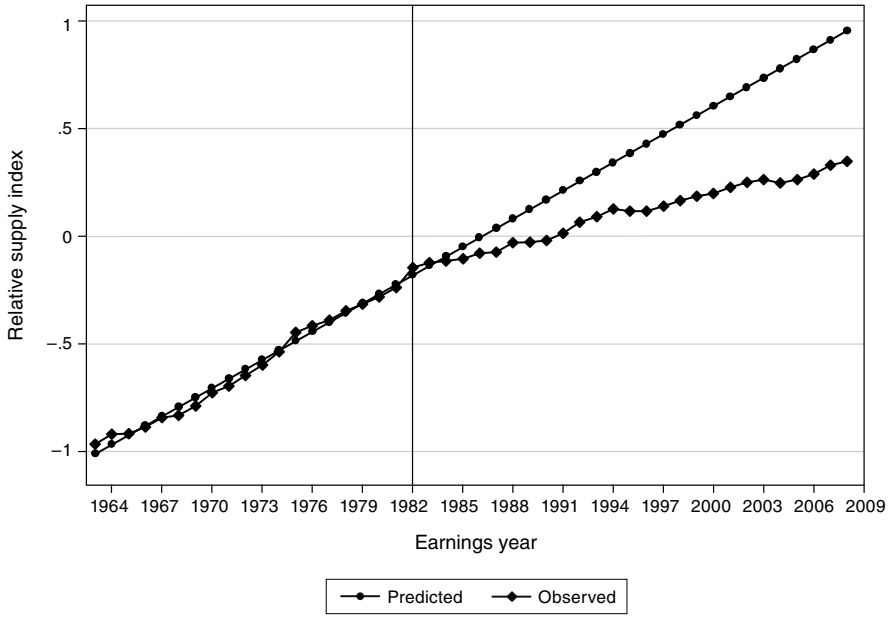


FIGURE 10.11
Predicted and observed college/high school log relative labor supply in the United States over the period 1963–2008. The predicted ratio corresponds to the extrapolation of a linear trend from years 1963 to 1982.

Source: Acemoglu and Autor (2011) data set.

supply of skilled labor. This phenomenon is illustrated by figure 10.11, which shows that the college/high school log relative labor supply (the term $\ln \nu$ in equation (10.40)) increased at a lower rate after 1982. As argued by Goldin and Katz (2008) in their book *The Race Between Education and Technology*, the simple supply-and-demand framework does a good job of explaining changes in inequalities between high-skilled and low-skilled workers. And this framework shows that if the relative supply of college-educated workers had increased at the same rate from 1980 to 2008 as it did from 1960 to 1980, the college premium would not have increased in the United States after 1980. The college wage premium increased because the United States lost the race to technology in the late twentieth century.

Overall this research shows that the evolution of wage inequalities between skilled and unskilled workers since the 1960s in the United States can be explained by a bias in technological progress and by choices concerning investment in education. The technological progress bias, however, only shows up as a temporal trend. It cannot be excluded a priori that this trend is explained by other factors, such as changes in consumer preferences or in international trade. To accurately evaluate the impact of technological progress, it is absolutely necessary to have more precise information about the form this progress takes.

2.3.2 Computerization

Many contributions have highlighted a relationship between the adoption of new technologies, the occupational structure, and wages. Thus, research on U.S. data generally finds that the introduction of new technologies (investment in computerization, expenditures on research and development, changes to the capital–labor ratio, employment of scientists and engineers, etc.) is accompanied by alterations to the structure of employment at the expense of unskilled manpower. For example, Berman et al. (1994) estimate, on sectoral U.S. data, that the relative growth of skilled labor is positively correlated with investment in computer equipment and research and development. Autor et al. (1998) show that, in every sector, the bias of technological progress is linked to the utilization of computers. Machin and Van Reenen (1998) show that this relationship turns up in the principal industrialized countries of the OECD. In addition, they emphasize that the reshaping of labor demand has spread through all sectors of the American economy.

Still, the simple observation of a relationship between technology, occupations, and wages does not allow us to pin down the impact of technology. Some of the research on the impact of the spread of computer technology illustrates this problem perfectly. Krueger (1993) observed that more intensive use of computer technology goes hand in hand with rising earnings inequality. He took the view that the increasing use of computers in the 1980s was essentially restricted to more-skilled workers and contributed to widening the wage gap in their favor. The wage bonus associated with the use of computers was thought to be on the order of 20% in 1989. Research by Entorf and Kramarz (1997) and Entorf et al. (1999) on French data indicated, however, that too much may be read into estimates of this type. These authors emphasized the possibility of a selection bias: firms may have chosen the most productive employees to work with the new equipment. For that matter, their estimates suggested that this selection bias explained the largest part of the wage bonus. When this selection bias is corrected for, it turns out that the wage bonus linked to the use of the computer amounted only to 2%. This result was confirmed by a study on German data by DiNardo and Pischke (1997), which

showed that pens, pencils, and even the sitting (as opposed to standing) position exert positive effects on wages, similar to those induced by computers. Users of computers, pens, and pencils, or even persons who work in a sitting position, likely possess unobservable characteristics that favor high productivity. Therefore, individuals receiving relatively high wages would have been the first to be provided with computers.

In this context, it is essential to account for the causes of the adoption of new technologies in order to grasp their interactions with wages and the occupational structure. Autor et al. (2003) have shown that within industries, occupations, and education groups, computerization is associated with reduced labor input of routine manual and routine cognitive tasks of the sort that can be accomplished by following explicit rules and with increased labor input of nonroutine cognitive tasks. This phenomenon has had a significant impact on the U.S. labor market, as can be seen in figure 10.12, which plots changes in the tasks performed by the U.S. labor force over the period 1960 to 2002.

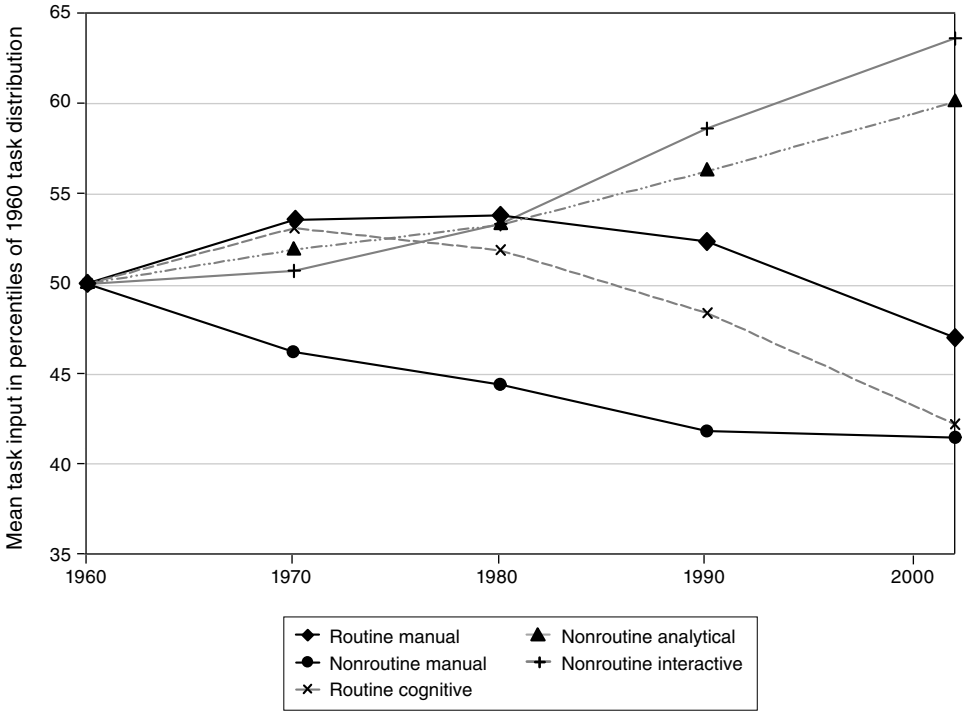


FIGURE 10.12
Trends in routine and nonroutine task input, 1960 to 2002.

Note: The figure is constructed using *Dictionary of Occupational Titles* (1977) task measures by gender and occupation paired to employment data from the Census and and the Current Population Survey (CPS) samples. Data are aggregated to 1,120 industry-gender-education cells by year, and each cell is assigned a value corresponding to its rank in the 1960 distribution of task input (calculated across the 1,120, 1960 task cells). Plotted values depict the employment-weighted mean of each assigned percentile in the indicated year. By construction, each task variable has a mean of 50 centiles in 1960. Subsequent points depict the employment-weighted mean of each assigned percentile over each decade.

Source: Autor et al. (2003, figure 1).

In this context, computer capital substitutes for workers performing routine cognitive and manual tasks, but computer capital complements workers in performing nonroutine problem-solving and complex communications tasks. On the basis of this observation, we ought to expect more computerization in firms employing a high proportion of wage earners performing routine tasks. This prediction is documented by Autor and Dorn (2013). To this end, they dispose of data concerning 722 commuting zones in the United States over the period 1950–2005. These commuting zones, which correspond to local labor markets, are clusters of counties that are characterized by strong commuting ties within commuting zones and weak commuting ties across commuting zones. The Census IPUMS and the American Community Survey (ACS) provide information on 330 detailed occupations encompassing all of U.S. nonfarm employment. Furthermore, the *Dictionary of Occupational Titles* allows them to measure routine, abstract, and manual task content by occupation. This makes it possible to build an index of routine-task intensity by occupation k, calculated as:

$$T_k = \ln T_k^R - \ln T_k^A - \ln T_k^M$$

where $\ln T_k^R, \ln T_k^A$, and $\ln T_k^M$ are, respectively, the routine, abstract, and manual task inputs in occupation k.

Using this index, it is possible to compute a routine employment share measure by commuting zone. For each commuting zone, this routine employment share measure is equal to:

$$RSH = \frac{\sum_k L_k \mathbb{I}\left(RTI_k > \overline{RTI}\right)}{\sum_k L_k}$$

where L_k is employment in occupation k in the corresponding commuting zone and $\mathbb{I}$ is the indicator function equal to 1 if $RTI_k > \overline{RTI}$ and to zero otherwise. $\overline{RTI}$ is a threshold value assumed to be equal to the 66th percentile of the distribution of RTI_k.

Autor and Dorn then estimate a relation between the computer capital measured by the average number of personal computers per employee and the routine employment share measure by commuting zone, or:

$$\Delta PC_{jst} = a_0 + a_1 RSH_{jst_0} + \gamma_s + \delta_t + \varepsilon_{kst}$$

where ΔPC_{jst} denotes changes in computer capital in commuting zone j of state s from decade t_0 to decade t, γ_s is a state-fixed effect, δ_t is a time-fixed effect, and ε_{kst} is an error term. Table 10.3 presents the results of this estimation for different time periods. It shows that coefficient a_1, which measures the relation between the increase in computer capital in decade t and the routine employment share in the previous decade, is positive and highly significant. This indicates that computerization is more widespread in commuting zones where the initial share of routine occupations was larger. Beaudry et al. (2010) found similar results: they document that cities that were initially relatively skill-abundant as of 1980 differentially adopted computer technology thereafter. Autor and Dorn also show that changes in routine employment shares (i.e., ΔRSH_{jst}) are negatively correlated with routine employment shares in the previous decade (i.e., RSH_{jst_0}). All in all, these results indicate that computer capital displaces labor from routine tasks.

TABLE 10.3
Computer adoption and task specialization within commuting zones, 1980–2005.

	1980–1990	1990–2000	1980–2000
Share of routine occupations in previous decade	0.695 (0.061)	0.490 (0.076)	0.619 (0.044)
R^2	0.577	0.332	0.385
Number of observations	675	660	1335

Note: The dependent variable is 10 × annual change in adjusted personal computers per employee. Robust standard errors in parentheses are clustered on state.

Source: Autor and Dorn (2013, table 3).

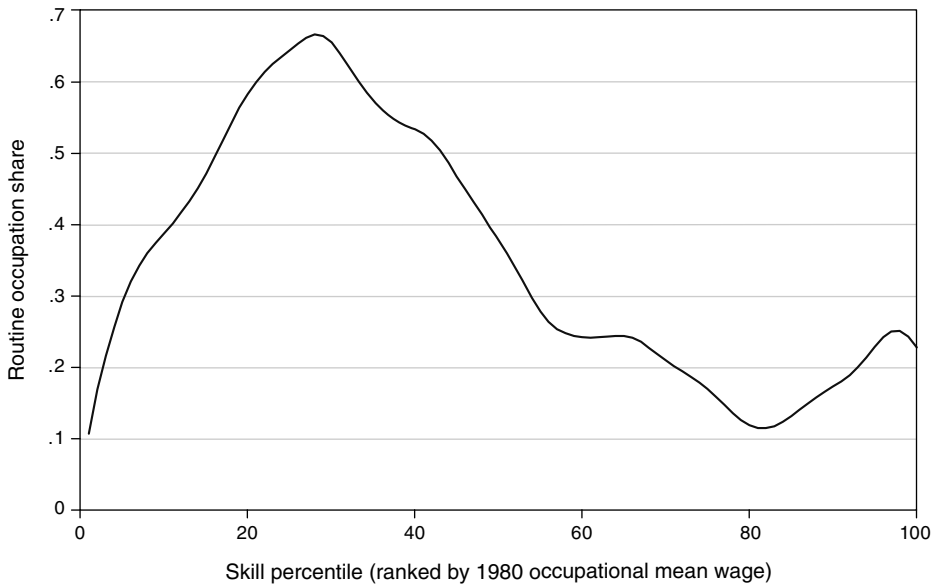


FIGURE 10.13
Share of "routine" occupations by occupational skill percentile.

Source: Autor and Dorn (2013, figure 4).

Autor and Dorn then show that workers with an intermediate level of skill are more frequently assigned to routine tasks than workers with either high or low levels of skill. This phenomenon is illustrated by figure 10.13, which displays the relation between the skill level of occupations (measured by occupational mean wage) and the routine employment share (measured by the *RTI* index) in 1980. It depicts an inverted U-shape between the skill level and the routine employment share. Therefore, according to the assignment model, it should be expected that computerization induces job and wage polarization. It should reduce the share of occupations with intermediate skill levels and decrease the relative wage of these jobs. Figure 10.14 shows that job polarization over the period 1980–2005 is more pronounced in commuting zones where the share of routine occupations was bigger in 1980. Figure 10.15 shows that wage polarization is also more marked in these commuting zones. Accordingly, these figures suggest that

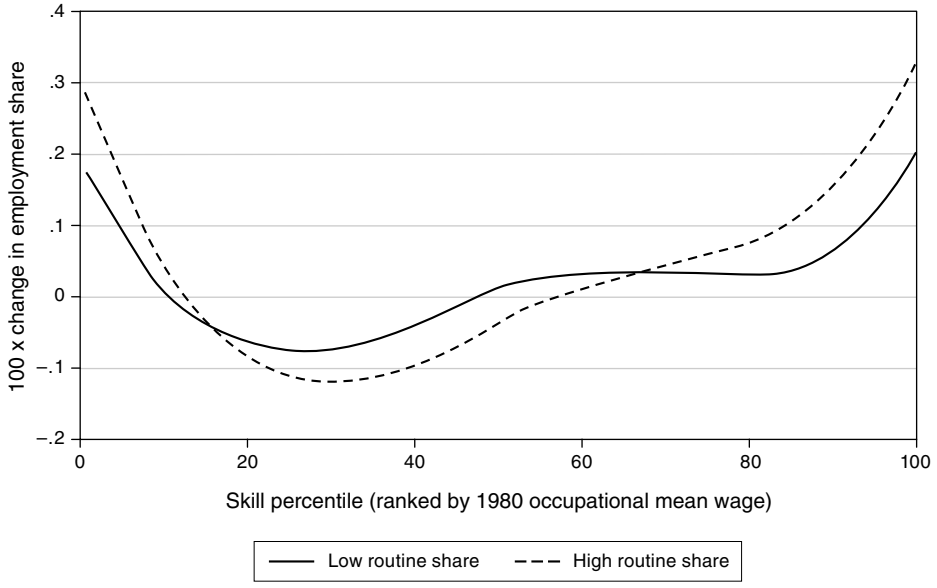


Figure 10.14
Smoothed changes in employment by skill percentile in commuting zones with high and low routine employment shares in 1980.

Source: Autor and Dorn (2013, figure 5A).

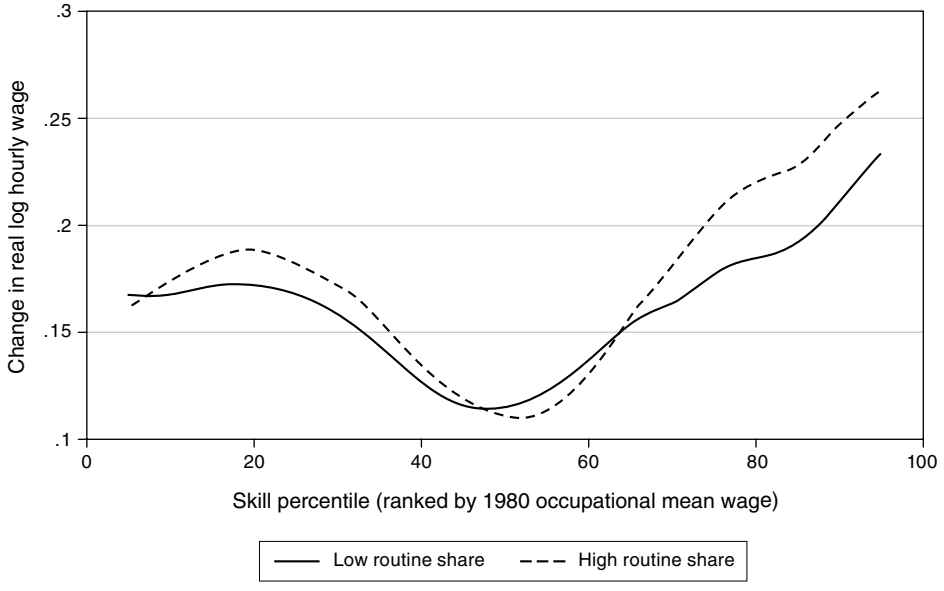


Figure 10.15
Smoothed changes in hourly wages by skill percentile in commuting zones with high and low routine employment shares in 1980.

Source: Autor and Dorn (2013, figure 5B).

the declining price of computer technology induced job and wage polarization. Autor and Dorn (2013) also show that this phenomenon has induced low-skilled workers to reallocate their labor supply to service occupations, which are difficult to automate because they rely heavily on flexible interpersonal communication and direct physical proximity.

Overall, theoretical and empirical studies suggest that technological bias has a significant impact on the occupational structure and on inequality between workers with different skill levels. It also shows that the type of technological bias can change over time.

2.4 THE ROLE OF INSTITUTIONS

The fall in demand for low-skilled labor led, during the last two decades of the twentieth century, to an increase in wage inequality in the "Anglo-American" economies, while in continental Europe it led to a heightened incidence of unemployment. To simplify somewhat, we can distinguish two types of behavior in response to the reshaping of labor demand. On one hand there is the model we are calling "Anglo-American," characterized by wage flexibility and resulting in an increase in wage inequality. Katz and Autor (1999) emphasize that the increase in the wage spread between workers with different skill levels in the 1980s was indeed greater in the United States and the United Kingdom than it was in other European countries. On the other hand, the model we are calling "European" (especially Germany until the mid-2000s, France, Italy, and to a lesser extent Sweden), marked by refusal to accept increasing wage inequality, saw heightened disparity in the incidence of unemployment. To understand the impact of the reshaping of labor demand on unemployment and wage inequality, it is useful to analyze the consequences of biased technological progress when there is a binding minimum wage. It then becomes possible to assess the evolution of a global inequality index—discounted average gains—for the two types of worker in the Anglo-American and European models. It will be shown that controlling the spread of remunerations by means of a minimum wage can lead, in the end, to a decrease in wage inequality but to an increase in inequality in terms of discounted average gains.

2.4.1 WAGE INEQUALITY AND UNEMPLOYMENT

Let us consider a model similar to that described above in section 2.3. We thus assume that there are two labor markets, perfectly sealed off from each other, and corresponding to skilled labor ($i = h$) and unskilled labor ($i = \ell$). The productive sector produces three goods: a final good, consumed by agents, and two intermediate goods that serve to produce the final good. The final good is the numéraire, and the price of a unit of intermediate good of type i is denoted by p_h, $i = h, \ell$. Intermediate good h is produced using skilled labor alone, while intermediate good ℓ is produced using unskilled labor alone. Each employee is capable of making one unit of intermediate good per unit of time. Production of the final good is represented by the CES production function defined by equation (10.37), where L_h and L_ℓ can designate either the quantities of intermediate goods produced by the skilled and the unskilled respectively, or the number of skilled and unskilled jobs respectively. In this setup, parameters B_h and B_ℓ then measure technological progress that increases the efficiency of skilled and unskilled labor. All markets are assumed to be perfectly competitive.

The demands for the intermediate goods maximize profits in the final good industry. This implies:

$$\frac{p_h}{p_\ell} = \frac{B_h}{B_\ell}\left(\frac{L_h}{L_\ell}\right)^{-1/\varepsilon} \tag{10.41}$$

At equilibrium, the ratio of the prices of the intermediate goods thus depends on technological progress and the number of jobs in each of the two worker categories. Zero profit condition implies that, in each intermediate good sector, the wage w_i equals the price p_i since one unit of labor produces one unit of good in those sectors. Therefore, at labor market equilibrium, where labor supply N_i equals labor demand L_i for both worker types, equation (10.41) implies that the relative wage of skilled workers is given by:

$$\omega = \beta \nu^{-1/\varepsilon} \tag{10.42}$$

where $\omega = w_h/w_\ell, \beta = B_h/B_\ell$, and $\nu = N_h/N_\ell$. We may take the view that this relation describes the Anglo-American model, characterized by flexible wages. This flexibility ensures that biased technological progress which increases the relative productivity of skilled workers does not induce unemployment but does, as a counterpart, increase wage inequality.

On a European labor market, the wage of unskilled workers is no longer flexible. This entails that technological bias affects unemployment for this category of worker. To show this, let us suppose that unskilled workers are paid the minimum wage, and that the minimum wage is indexed to the wage of skilled personnel, so that $w_\ell = \mu w_h$, where μ is an exogenous parameter lying between zero and one. When the minimum wage is binding, equation (10.42) implies that $1/\mu = \beta\left(N_h/L_\ell\right)^{-1/\varepsilon}$, and then that the unemployment rate of unskilled workers, $u_\ell = \left(N_\ell - L_\ell\right)/N_\ell$, is defined by:

$$u_\ell = 1 - \nu\left(\mu\beta\right)^{-\varepsilon}$$

A technological bias unfavorable to unskilled workers, corresponding to an increase in β, increases their unemployment rate. This model well describes the situation of a country like France, where the minimum wage is de facto indexed to the average wage. If wages could be adjusted, the technological bias would actually lead to an adjustment of remunerations without changing the unemployment rate. But the indexation of the minimum wage to the wage of skilled workers prevents these adjustments from taking place and, in sum, the technological bias entails a rise in unemployment among the unskilled. It is also worth noting that the minimum wage reduces the wage of skilled workers because their marginal productivity increases with unskilled labor (the cross derivative of the production function with respect to skilled and unskilled labor is positive). Since the minimum wage reduces unskilled labor demand, it has a negative impact on the wage of skilled workers.

2.4.2 Does the Minimum Wage Reduce Inequality?

One of the purposes of the minimum wage is to reduce inequality of income. However, the minimum wage increases inequality of exposure to the risk of unemployment when

the economy is affected by a reshaping of labor demand. So the minimum wage has an ambiguous effect on the average gains of unskilled workers. Let us denote the minimum wage by w and demand for the labor of low-skilled workers by $L = L^d(w)$. We get:

$$\frac{\mathrm{d}(wL)}{\mathrm{d}w} = L\left(1 + \frac{w\mathrm{d}L}{L\mathrm{d}w}\right)$$

This relation shows that the minimum wage drives up the average gains of low-skilled workers if and only if (the absolute value of) the elasticity of the demand for labor by low-skilled workers is less than unity. Since we saw in chapter 2 on labor demand that the elasticity of the demand for labor by low-skilled workers is approximately equal to one, it is verisimilar that the minimum wage has a weak impact on the average gains of low-skilled workers. Thus the minimum wage does not yield a reduction in the inequality between the average gains of skilled and unskilled workers. It does, however, increase income inequality within unskilled workers as a group, some of whom benefit from higher wages while others lose their jobs.

Flinn (2002) has shown that comparison of the Italian with the U.S. experience provides an illustration of this type of result.[6] Using a job search model and individual-level data for Italy and the United States, he shows that while the cross-sectional wage distributions of young Italian males are much more compressed than are the comparable distributions for young white U.S. males, it turns out that the distribution of lifetime welfare is no more dispersed in the United States than in Italy. Bowlus and Robin (2012) have obtained similar results in a comparison of the United States, Canada, France, and Germany in the 1990s.

Overall, these results suggest that the minimum wage may be a very poor instrument for the redistribution of income when the labor market is impacted by biased technological progress. We will see in chapter 12 that fiscal measures are probably a better way to counteract the effects of the reshaping of labor demand but that certain categories of the population may be opposed to using the fiscal system as an instrument of income redistribution.

2.5 Endogenous Technological Progress

To this point, technological progress has been considered exogenous. But the fact is that the form an innovation takes is not independent of the capacities of those who will be assigned to make use of it. It is likely that a relative abundance of workers with low skills will spur the invention of technologies that complement this input. This seems to have been the case at the end of the eighteenth and early in the nineteenth century, when the rural exodus of low-skilled manpower was accompanied by new kinds of machinery that workers of that sort could operate to carry out repetitive manufacturing tasks (see Acemoglu, 2002). So it is entirely possible, on that basis, that the increase in the supply of skilled labor in the second half of the twentieth century spurred innovations of the kind that complement skilled labor.

We can illustrate the determinants of technological progress by assuming that firms choose not only quantities of skilled and unskilled labor but also technology,

[6] There is no national minimum wage in Italy, but wage floors are defined by collective agreements at the industry level.

represented by parameters B_h and B_ℓ in the model with two categories of workers used in section 2.3 above. Let us consider a simplified limit case in which one unit of output is required to produce one unit of technological factor h or ℓ. The problem of the representative firm is then written:

$$\max_{(A_h,L_h,A_\ell,L_\ell)} G(B_h,L_h,B_\ell,L_\ell) - w_h L_h - w_\ell L_\ell - B_h - B_\ell$$

The production function $G(B_h,L_h,B_\ell,L_\ell)$ has to have constant returns to scale with respect to all inputs, B_h, L_h, B_ℓ, and L_ℓ. Assuming, for the sake of simplicity, that the production function is of the CES type, it reads:

$$G(B_h L_h, A_\ell L_\ell) = \left[B_h (L_h)^{\frac{\varepsilon-1}{\varepsilon}} + B_\ell (L_\ell)^{\frac{\varepsilon-1}{\varepsilon}}\right]^{\frac{\varepsilon}{2\varepsilon-1}}$$

It can easily be verified that the first-order conditions entail that the relative demand for skilled labor satisfies equation (10.38) and that the choice of technological factors must satisfy:

$$\beta = \lambda^{\frac{1-\varepsilon}{\varepsilon}} \tag{10.43}$$

If we assume competitive labor markets, the relative employment of skilled workers is equal to the relative supply of skilled labor, that is, $\lambda = \nu$. Equation (10.43) then shows that the relative productivity $\beta = B_h/B_\ell$ of skilled workers increases with the supply of skilled workers if and only if the elasticity of substitution is smaller than unity. We can also eliminate the technological bias β from equations (10.38) and (10.43) to find a relationship between the structure of the labor supply and the wage structure; what we get is:

$$\omega = \nu^{\frac{\varepsilon^2}{1-2\varepsilon}} \tag{10.44}$$

This relation shows that the increase in the relative supply of skilled labor, ν, leads to an increase in the relative wage of skilled workers if $\varepsilon < 2$, and to a decrease if not. So the endogenous response of technological progress can lead to an increasing relation between the relative supply of skilled labor and the relative wage of skilled workers, for a sufficiently low value of the elasticity of substitution; and such a value is plausible according to the empirical studies presented above. This rising relation, which does not exist when technological progress is exogenous, arises from the choice by firms of technologies complementary to skilled labor when the quantity of this input grows. Note however that the model presented here is very simple and leaves out the dynamic aspects of the adoption of new technologies. In reality, the installation of new technology is generally accompanied by adjustment costs that can reduce the incomes of the individuals least adaptable to change (on the dynamics of inequality and its links with technological progress, see Caselli, 1999, and Aghion, 2002).

This rudimentary model does nevertheless allow us to understand why an increase in the proportion of highly skilled workers may, on its own, support technological bias and steep wage inequalities. It also highlights the potential ambiguity of the impact of government aid for education on inequality: the general rise in educational level achieved by prolonging compulsory schooling does not always lead to a

reduction in inequality. The response of innovators affects the direction of technical progress and may on the contrary help to increase the inequality between those who succeed in accumulating enough knowledge and know-how to master the new technologies and the rest. In these circumstances, a rise in the supply of skilled labor may increase inequality and have the opposite effect to the one intended. Accordingly the interactions between education and inequality are complex: in order to reduce wage inequality, it is not enough just to increase the proportion of skilled workers, for the direction of technological progress itself depends on the economic environment.

3 SUMMARY AND CONCLUSION

- Growth in labor productivity improves the profit outlook. This *capitalization effect* is favorable to employment.
- As a general rule, technological progress does not apply to all jobs in a uniform manner. Jobs based on obsolete technologies are destroyed, and only those capable of integrating the latest innovations survive. This process of *creative destruction* can be unfavorable to employment.
- Empirical studies suggest that, overall, technological progress has an ambiguous effect on unemployment. The impact of technological progress on unemployment depends on the type of innovation that underpins it and on labor market institutions.
- During the last two decades of the twentieth century, most industrialized countries were faced with technological bias that altered labor demand in favor of skilled workers. The new information technologies, which have taken the place of jobs of intermediate skill involving routine tasks, have induced a phenomenon of job polarization, entailing a diminution in the proportion of jobs of intermediate skill and a reduction in the remuneration to such jobs.
- The Anglo-American model is characterized by high wage flexibility. Conversely, in the European model wages are most often downward rigid, and a large portion of adjustment occurs through variation in employment. The existence of a high minimum wage is a major element in this type of regulation. More severe technological bias may entail more inequality in terms of *average gains* (that is, over the whole of the life cycle) in the presence of a minimum wage.

4 RELATED TOPICS IN THE BOOK

- Chapter 2, section 1.2: The substitution of capital for labor
- Chapter 3, section 3: Assortative matching
- Chapter 4, section 1: The theory of human capital
- Chapter 4, section 5: The returns to education
- Chapter 9: Equilibrium unemployment

- Chapter 11, section 1: International trade and labor market: Facts and theories
- Chapter 12: Income redistribution
- Chapter 13, section 2: Employment protection

5 FURTHER READINGS

Acemoglu, D., & Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 4B, pp. 1043–1171). Amsterdam: Elsevier Science.

Autor, D., & Dorn, D. (2013). The growth of low-skill service jobs and the polarization of the U.S. labor market. *American Economic Review*, *103*(5), 1553–1597.

Costinot, A., & Vogel, J. (2010). Matching and inequality in the world economy. *Journal of Political Economy*, *118*, 747–786.

Goldin, C., & Katz, L. (2008). *The race between education and technology*. Cambridge, MA: Harvard University Press.

Mortensen, D., & Pissarides, C. (1998). Technological progress, job creation, and job destruction. *Review of Economic Dynamics*, *1*, 733–753.

6 APPENDIX

6.1 The Relation Between θ and g

Differentiating the two sides of relation (10.19) defining the wage curve brings us to:

$$\frac{\partial T}{\partial g} = -\left(T + \frac{\gamma h e^{gT}}{1-\gamma}\frac{\partial \theta}{\partial g}\right)/g \tag{10.45}$$

Equation (10.18) defining labor demand can be written in the following manner:

$$\frac{h}{m(\theta)} = \frac{1-\gamma}{r}H(g,T) \quad \text{with} \quad H(g,T) = 1 + \frac{ge^{-rT} - re^{-gT}}{r-g}$$

Differentiating this equation with respect to g, we get:

$$-\frac{hm'(\theta)}{m^2(\theta)}\frac{\partial \theta}{\partial g} = \frac{1-\gamma}{r}\left(H_g + H_T\frac{\partial T}{\partial g}\right)$$

Bringing the value of $\frac{\partial T}{\partial g}$ that issues from (10.45) into this last inequality, we find:

$$\left[\frac{\gamma h H_t e^{gT}}{rg} - \frac{hm'(\theta)}{m^2(\theta)}\right]\frac{\partial \theta}{\partial g} = \frac{1-\gamma}{r}\left(H_g - \frac{TH_T}{g}\right) \tag{10.46}$$

with:

$$H_T = \frac{rg}{r-g}\left(e^{-gT} - e^{-rT}\right) > 0$$
$$H_g = \frac{1}{(r-g)^2}\left[(r-g)\left(e^{-rT} + rTe^{-gT}\right) + ge^{-rT} - re^{-gT}\right]$$

After several rearrangements, we see that $H_g - TH_T/g$ is of the same sign as $e^{-rT} - e^{-gT} + T(r-g)e^{-rT}$, and a second-order expansion of $e^{-rT} - e^{-gT}$ then shows that this expression is negative. Equation (10.46) then entails $\partial\theta/\partial g < 0$.

6.2 Properties of the Assignment Model

6.2.1 Unicity and Monotony of the Matching Function $M(\sigma)$

When (10.25) is verified, function $A_\sigma(s,\sigma)/A(s,\sigma)$ is monotonously increasing with s for all σ, which entails that equation (10.32) defines a unique value of M for every value of σ. Condition (10.32) thus well defines a function M that associates with each task σ a unique skill $s = M(\sigma)$.

Deriving relation (10.31) with respect to σ, we get (simplifying certain notations in obvious fashion):

$$\left(BA_{ss} - w''\right)M' + A_sB' + A_{s\sigma}B = 0$$

Following (10.32), $A_sB' + A_{s\sigma}B$ is equal to $\frac{B}{A}(A_{s\sigma}A - A_\sigma A_s)$, which is a strictly positive quality following hypothesis (10.25) of log supermodularity. For that matter, $BA_{ss} - w''$ is strictly negative following the second-order condition (10.29). The result is that $M'(\sigma) > 0$ and the matching function is thus strictly increasing.

6.2.2 The Increase in Relative Wages of Highly Skilled Workers

According to relation (10.32), defining the matching function, to each schedule B_i ($i = 1, 2$) there corresponds a matching function M_i defined by $F[M_i(\sigma),\sigma] = -B_i'(\sigma)/B_i(\sigma)$ with $F(s,\sigma) \equiv A_\sigma(s,\sigma)/A(s,\sigma)$. When the inequality (10.35) is verified, then we have $F[M_2(\sigma),\sigma] < F[M_1(\sigma),\sigma]$. We have already noted that function $F(s,\sigma)$ is increasing with s when we admit the hypothesis of log supermodularity (10.25); the result is that $M_2(\sigma) < M_1(\sigma)$ for all σ. This inequality signifies that skill-biased technological progress reduces the skills necessary to perform a given task. The result will be that workers will move toward tasks with higher skill intensity. To show the last point formally, we apply function M_2^{-1} (which is increasing) to the two sides of inequality $M_2(\sigma) < M_1(\sigma)$. We get $\sigma < M_2^{-1}[M_1(\sigma)]$. Let us now consider skill s associated to skill intensity σ by schedule M_1, or $\sigma = M_1^{-1}(s)$. We then arrive at $M_1^{-1}(s) < M_2^{-1}(s)$, which signifies precisely that each worker will have to migrate toward tasks with higher skill intensity. Inserting $\sigma = M^{-1}(s)$ in equations (10.30) and (10.31), we arrive at:

$$\frac{w'(s)}{w(s)} = \frac{A_s\left[s, M^{-1}(s)\right]}{A\left[s, M^{-1}(s)\right]} \tag{10.47}$$

Let us denote $w_i(s)$ $(i = 1, 2)$ the wage function corresponding to the technological schedule $B_i(s)$. We have just seen that $M_1^{-1}(s) < M_2^{-1}(s)$ when condition (10.35) is verified. Let us now consider function $G(s, \sigma) = A_s(s, \sigma)/A(s, \sigma)$. It is easily verified that this function is increasing with σ when function $A(s, \sigma)$ satisfies condition (10.25) of log supermodularity. Equation (10.47) then entails:

$$\frac{w_2'(s)}{w_2(s)} > \frac{w_1'(s)}{w_1(s)}$$

This inequality signifies that the ratio $w_2(s)/w_1(s)$ is an increasing function of s.

6.2.3 Job and Wage Polarization

The same reasoning we set out following condition (10.35) leads to $\sigma < M_2^{-1}[M_1(\sigma)]$ for $\sigma > \hat{\sigma}$ and $\sigma > M_2^{-1}[M_1(\sigma)]$ for $\sigma < \hat{\sigma}$. Inserting $\sigma = M_1^{-1}(s)$ in these last two inequalities, we find:

$$M_2^{-1}(s) > M_1^{-1}(s) \quad \text{for } s > M_1(\hat{\sigma}) \quad \text{and} \quad M_2^{-1}(s) < M_1^{-1}(s) \quad \text{for } s < M_1(\hat{\sigma})$$

It could be shown by the same line of reasoning that wages undergo a change characterized by:

$$\frac{w_2'(s)}{w_2(s)} > \frac{w_1'(s)}{w_1(s)} \quad \text{for } s > M_1(\hat{\sigma}) \quad \text{and} \quad \frac{w_2'(s)}{w_2(s)} < \frac{w_1'(s)}{w_1(s)} \quad \text{for } s < M_1(\hat{\sigma})$$

REFERENCES

Acemoglu, D. (2002). Technical change, inequality, and the labor market. *Journal of Economic Literature*, *40*(1), 7–72.

Acemoglu, D., & Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 4B, pp. 1043–1171). Amsterdam: Elsevier Science.

Aghion, P. (2002). Schumpeterian growth theory and the dynamics of income inequality. *Econometrica*, *70*, 855–882.

Aghion, P., & Howitt, P. (1992). A model of growth through creative destruction. *Econometrica*, *60*, 323–351.

Aghion, P., & Howitt, P. (1998). *Endogenous growth theory*. Cambridge, MA: Harvard University Press.

Autor, D., & Dorn, D. (2013). The growth of low-skill service jobs and the polarization of the U.S. labor market. *American Economic Review*, *103*(5), 1553–1597.

Autor, D., Katz, L., & Krueger, A. (1998). Computing inequalities: Have computers changed the labor market? *Quarterly Journal of Economics*, *113*, 1169–1213.

Autor, D., Levy, F., & Murnane, R. (2003). The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics*, *118*(4), 1279–1334.

Bauer, T., & Bender, S. (2004). Technological change, organizational change and job turnover. *Labour Economics*, *11*(3), 265–291.

Bean, C., & Pissarides, C. (1993). Unemployment, consumption and growth. *European Economic Review*, *37*, 837–854.

Beaudry, P., Doms, M., & Lewis, E. (2010). Should the personal computer be considered a technological revolution? Evidence from U.S. metropolitan areas. *Journal of Political Economy*, *118*, 988–1036.

Behaghel, L., Caroli, E., & Walkowiak, E. (2012). Information and communication technologies and skill upgrading: The role of internal vs external labour markets. *Oxford Economic Papers*, *64*(3), 490–517.

Berman, A., Bound, J., & Griliches, Z. (1994). Changes in the demand for skilled labor within U.S. manufacturing: Evidence from the annual survey of manufactures. *Quarterly Journal of Economics*, *109*, 367–397.

Bloom, N., Sadun, R., & van Reenen, J. (2012). Americans do IT better: US multinationals and the productivity miracle. *American Economic Review*, *102*(1), 167–201.

Bowlus, A. J., & Robin, J.-M. (2012). An international comparison of lifetime labor income values and inequality. *Journal of the European Economic Association*, *10*(6), 1236–1262.

Caballero, R. (1993). Comment on the Bean and Pissarides paper. *European Economic Review*, *37*, 855–859.

Caballero, R., & Hammour, M. (1996). On the timing and efficiency of creative destruction. *Quarterly Journal of Economics*, *111*, 805–852.

Card, D., & DiNardo, J. (2002). Skill biased technological change and rising wage inequality: Some problems and puzzles (NBER Working Paper No. 8769). www.nber.org/papers/w8769.

Caselli, F. (1999). Technological revolutions. *American Economic Review*, *87*, 78–102.

Costinot, A., & Vogel, J. (2010). Matching and inequality in the world economy. *Journal of Political Economy*, *118*, 747–786.

DiNardo, J., & Pischke, J. (1997). The returns to computer use revisited: Have pencils changed the wage structure too? *Quarterly Journal of Economics*, *114*, 291–303.

Dustmann, C., Ludsteck, J., & Schönberg, U. (2009). Revisiting the German wage structure. *Quarterly Journal of Economics*, *124*, 809–842.

Entorf, H., & Kramarz, F. (1997). Does unmeasure ability explain the higher wages of new technology workers? *European Economic Review*, *41*, 1489–1509.

Entorf, H., Gollac, M., & Kramarz, F. (1999). New technologies, wages, and worker selection. *Journal of Labor Economics*, *17*, 464–491.

Firpo, S., Fortin, N., & Lemieux, T. (2011). Occupational tasks and changes in the wage structure. Mimeo, University of British Columbia, Canada.

Flinn, C. (2002). Labor market structure and inequality: A comparison of Italy and the U.S. *Review of Economic Studies*, *69*, 611–645.

Foster, L., Haltiwanger, J., & Krizan, C. (2001). Aggregate productivity growth: Lessons from microeconomic evidence. In E. Dean, M. Harper, & C. Hulten (Eds.), *New development in productivity analysis* (pp. 303–363). Chicago, IL: University of Chicago Press.

Foster, L., Haltiwanger, J., & Krizan, C. (2006). Market selection, reallocation, and restructuring in the U.S. retail trade sector in the 1990s. *Review of Economics and Statistics*, *88*(4), 748–758.

Garicano, L., & Rossi-Hansberg, E. (2006). Organization and inequality in a knowledge economy. *Quarterly Journal of Economics*, *121*(4), 1383–1435.

Givord, P., & Maurin, E. (2004). Changes in job insecurity and their causes: An empirical analysis for France, 1982–2002. *European Economic Review*, *48*, 595–615.

Goldin, C., & Katz, L. (2008). *The race between education and technology*. Cambridge, MA: Harvard University Press.

Goos, M., & Manning, A. (2007). Lousy and lovely jobs: The rising polarization of work in Britain. *Review of Economics and Statistics*, *89*(1), 118–133.

Goos, M., Manning, A., & Salomons, A. (2009). The polarization of the European labor market. *American Economic Review Papers and Proceedings*, *99*(2).

Goos, M., Manning, A., & Salomons, A. (2014). Explaining job polarization: Routine-biased technological change and offshoring. *American Economic Review*, forthcoming.

Griliches, Z., & Regev, H. (1995). Firm productivity in Israeli industry, 1979–1988. *Journal of Econometrics*, *65*, 175–203.

Katz, L., & Autor, D. (1999). Changes in the wage structure and earnings inequality. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 3A, chap. 26, pp. 1463–1559). Amsterdam: Elsevier Science.

Katz, L., & Murphy, K. (1992). Changes in relative wages, 1963–1987. *Quarterly Journal of Economics*, *107*, 35–78.

Krueger, A. (1993). How computers have changed the wage structure: Evidence from microdata 1984–1989. *Quarterly Journal of Economics*, *108*, 33–60.

Lentz, R., & Mortensen, D. (2005). Productivity growth and worker reallocation. *International Economic Review*, *46*, 731–751.

Lentz, R., & Mortensen, D. (2008). An empirical model of growth through product innovation. *Econometrica*, *76*, 1317–1373.

Lentz, R., & Mortensen, D. (2010). Labor market friction, firm heterogeneity, and aggregate employment and productivity. *Annual Review of Economics*, *2*, 577–602.

Lynch, L., & Black, S. (1998). Beyond the incidence of employer provided training. *Industrial and Labor Relations Review*, *52*(1), 64–81.

Machin, S., & Van Reenen, J. (1998). Technology and changes in skill structure: Evidence from seven OECD countries. *Quarterly Journal of Economics*, *113*, 1215–1244.

Mortensen, D., & Pissarides, C. (1998). Technological progress, job creation, and job destruction. *Review of Economic Dynamics*, *1*, 733–753.

OECD. (2003). The sources of economic growth in OECD countries. Paris: OECD Publishing.

Rifkin, J. (1995). *The end of work: The decline of the global labor force and the dawn of the post-market era*. New York, NY: Tarcher and Putnam's Sons.

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, *82*, 34–55.

Saint-Paul, G. (2001). On the distribution of income and workers assignment under intrafirm spillovers, with an application to ideas and networks. *Journal of Political Economy*, *110*, 1–37.

Saint-Paul, G. (2008). *Innovation and inequality: How does technical progress affect workers?* Princeton, NJ: Princeton University Press.

Sattinger, M. (1975). Comparative advantage and the distributions of earnings and abilities. *Econometrica*, *43*, 455–468.

Schumpeter, J. (1934). *The theory of economic development*. Cambridge, MA: Harvard University Press.

Sismondi, J. (1991). *New principles of political economy* (R. Hys, Trans.). New Brunswick, NJ: Transaction Publishers.

Solow, R. (1957). Technical change and the aggregate production function. *Review of Economics and Statistics*, *39*, 312–320.

Teulings, C. (1995). The wage distribution in a model of assignment of skills to jobs. *Journal of Political Economy*, *103*, 280–315.

Teulings, C. (2005). Comparative advantage, relative wages, and the accumulation of human capital. *Journal of Political Economy*, *113*(2), 425–461.

Tinbergen, J. (1974). Substitution of graduate by other labor. *Kyklos*, *27*(2), 217–226.

CHAPTER 11

Globalization, Employment, and Inequality

In this chapter we will:

- Review facts about the rise in the volume of trade and its consequences
- Analyze the effects of trade on wage inequality and unemployment
- See how trade selects firms according to their productivity
- Estimate the impact of trade on unemployment using the approach of Dutt, Devashish, and Priya (2009) (The related data for 90 countries over the 1985–2004 period, as well as programs, are available at www.labor-economics.org.)
- Identify the characteristics of exporting firms
- Review facts about migratory flows
- Learn what the economic consequences of immigration are
- Estimate the impact of internal migration flows, based on the analysis of Boustan, Fishback, and Kantor (2010) for the United States during the Great Depression of the 1930s (Data and programs are available at www.labor-economics.org.)
- Observe how rapidly the labor market adjusts to international migration shocks

INTRODUCTION

Over the last 40 years international trade as a percentage of GDP has risen threefold in the United States, Germany, and Japan, and twofold in France, Sweden, and many other OECD countries. This increase has been driven in part by imports from the developing economies, and notably China, which is now the main source of imports in these countries and also a growing destination of exports. In the meantime migration flows also rose twofold in the advanced economies, notably due to growing inflows into European countries that used to be countries of emigration until the 1960s or the 1970s. These changes are at the origin of the growing attention to what is called "globalization."

This attention has also focused on other phenomena that paralleled the growing flows of trade and people across the world. Notably, the share of employment in manufacturing has continuously declined in the advanced economies: since the 1970s

it was halved, from 22% to about 9% of total employment, in the United States, but also France, the United Kingdom, and even in Germany, while outsourcing—imports of intermediate goods and services—has been growing steadily over the same period. At the same time, the performance of labor markets has changed a lot. In some countries, such as the United States, but also in many other OECD countries, wage inequality kept growing. In other countries, notably in continental Europe, high unemployment became a persistent phenomenon.

Are rising inequality and persistent unemployment the consequences of globalization? This question attracts a lot of attention in the media and in political debate, maybe even more in recent years than the role of technological progress. The image of plants closing at home and reopening in low-wage countries is repeatedly on the front page of newspapers. Multinationals exploiting workers in poor countries is another image frequently invoked to explain the rising tide of unemployment, or increasing inequality. But it is a stretch to go from observing the long-term trends in globalization and labor market performances to drawing conclusions about the existence of a causal relationship between globalization and labor market performances. This chapter is about carefully identifying the nature of this relationship, if any.

Changes in the economic environment, like international competition, the organization of production, and the cost of labor, do not just affect the rate of global unemployment and the average wage. They also influence the distribution of employment opportunities offered to different types of individual within countries. Competition with low-wage countries producing goods highly substitutable for those made by low-skilled workers in industrialized countries may prove unfavorable to the latter. We can discover the determinants of the evolution of wage inequalities and employment opportunities among workers of different skill levels by studying the evolution of the labor supply and demand for each category of worker, the impact of trade on the characteristics of firms, and the role of labor market institutions. An increase in the demand for a given type of labor is favorable to the opportunities of individuals who can supply this type of labor, while an increase in supply is unfavorable to them. But the supply and the demand for each type of labor are themselves influenced by many factors other than globalization, such as technological progress, demographic phenomena, and labor market institutions as a whole. Trade may also change the composition of firms because it favors more productive firms, which tend to be larger and offer higher wages for both blue-collar and white-collar workers. All these factors need to be taken into account before drawing a conclusion.

In section 1, we lay out the salient facts regarding the evolution of trade in conjunction with unemployment and inequality during the last four decades, then present the main explanations for them. We see in particular that shifts in the structure of labor demand, induced by competition from low-wage countries but also by the new market opportunities offered by trade, have undoubtedly played a major role but not necessarily along the lines suggested by newspapers. Notably, there is growing evidence that trade is associated with more employment and less unemployment in the long run, while evidence on the role of trade in the development of inequalities is at best mixed. But different OECD countries have reacted in sharply different ways to this alteration in the structure of labor demand, and in all countries the nature of jobs has changed substantially. Section 2 provides empirical evidence, at both the macro and micro level, concerning the effects of trade on the labor market. Finally, section 3 is devoted to

migration. In particular, we see that empirical studies on the impact of exceptional migration events do not reveal any substantial impact on the wage or level of unemployment. It is possible for labor markets to absorb large flows of immigrants without altering substantially the employment and wage outcomes of the native-born.

1 INTERNATIONAL TRADE AND LABOR MARKETS: FACTS AND THEORIES

Since World War II, international trade has been on the rise, and the poorer countries with large volumes of low-skilled labor ready to accept low pay are playing a larger part. The theory of international trade teaches us that this expanded participation by low-wage countries can lead, in certain circumstances illustrated by the Stolper and Samuelson theorem (1947), to a fall in the demand for low-skilled labor in the rich countries. This result does not, however, hold true in all circumstances. There are situations, it may plausibly be argued, in which stronger competition from low-wage countries benefits low-skilled workers in rich countries. More recent theories stress that trade favors the most productive firms, leading to different predictions on labor market outcomes.

1.1 The Rise in the Volume of Trade and Its Consequences

The integration of the world economy designated by the term "globalization" advanced at some periods and retreated at others during the twentieth century (see Temin, 1999). During recent years, however, the volume of trade between the industrialized countries and the emerging economies has risen in terms of both exports and imports. The gap in the cost of low-skilled manpower between the rich and the poor countries suggests that the latter have an advantage in the export of goods produced by this type of labor.

1.1.1 The Evolution of Trade Between Industrialized Countries and Developing Countries

Since the end of the 1970s, the fall in demand for unskilled labor in the developed countries has gone along with a strong advance in international trade, and in particular, commercial exchange between rich countries and poor ones. Figure 11.1 presents the openness rate (calculated as the average share of GDP of combined imports and exports of goods and services) of several OECD countries. It shows that on average these rates have grown considerably since 1970, notably for the United States, where the openness rate was multiplied by 3, from 10% in 1970 to 30% in 2008. The trend is similar in Japan, where trade reached 27% in 2008 at the onset of the Great Recession. Actually, this trend is common to almost all OECD countries and has been gaining strength since the early 1990s. In the European countries, the relative importance of trade appears to be larger at the end of the 2000s than in the United States or Japan, due to the strong economic integration with other European Union member countries (exports by European countries to other European countries represent about 70% of total cumulated exports in

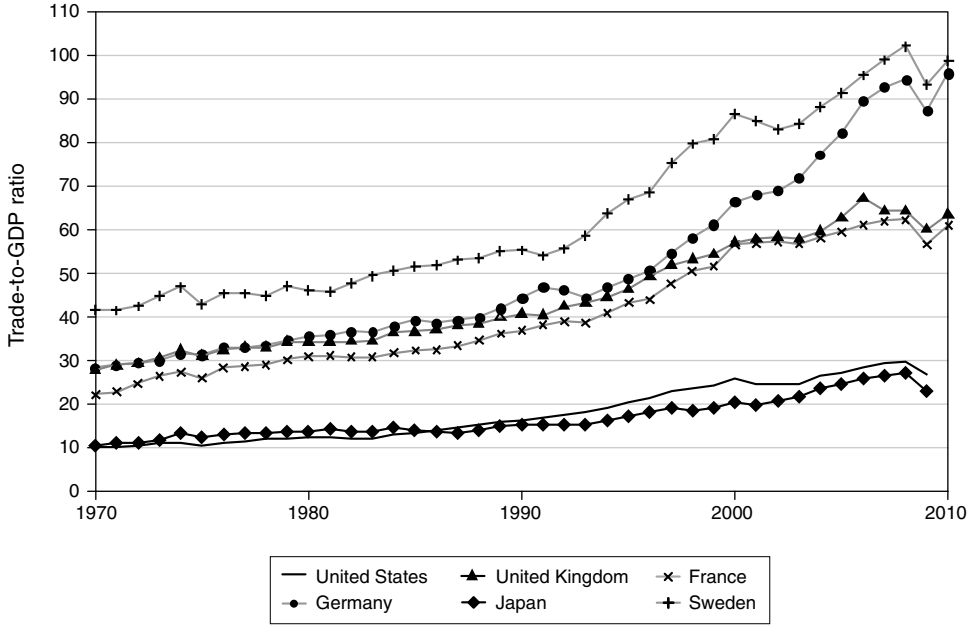


FIGURE 11.1
Trade openness, 1970–2010.

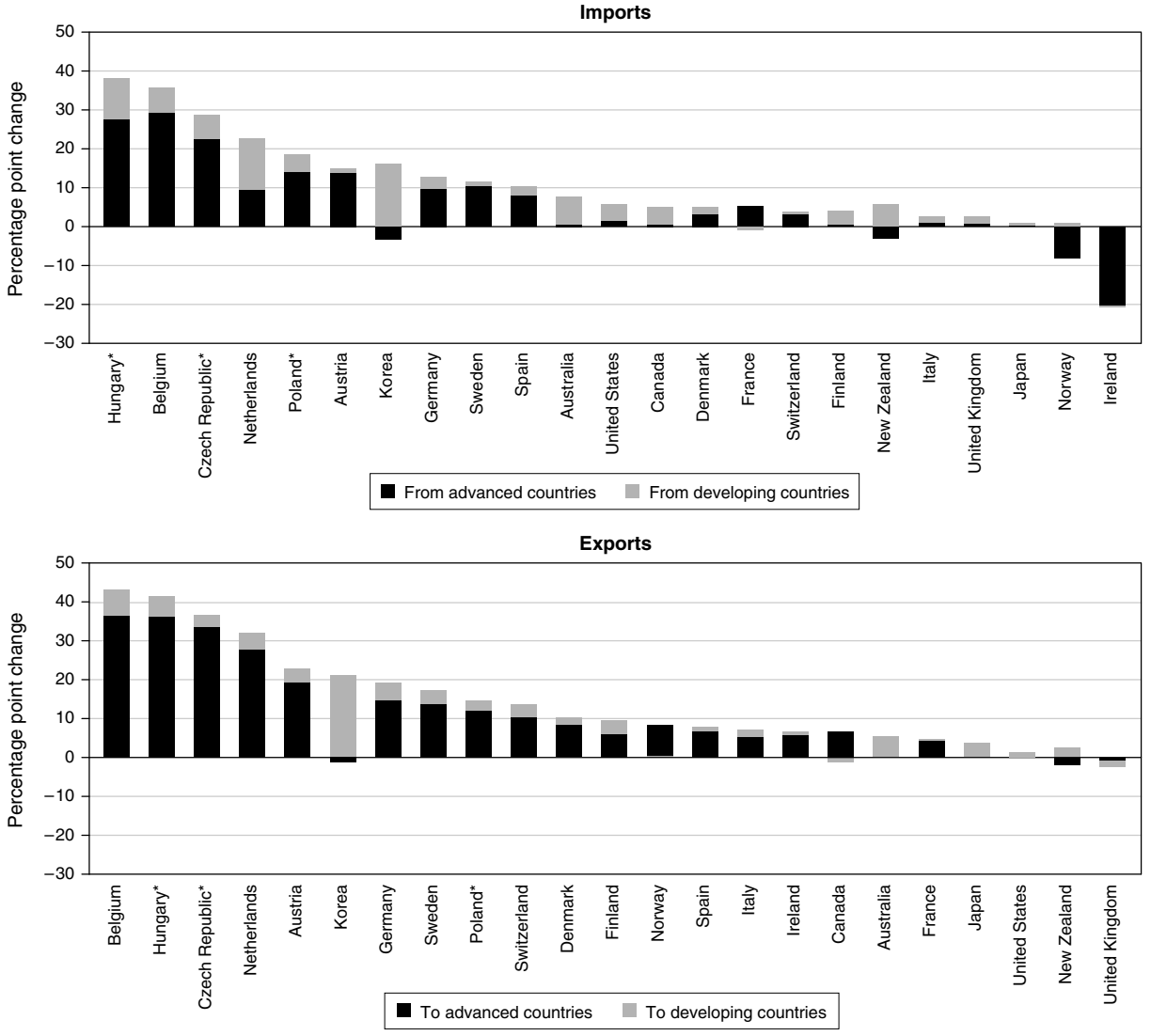
Note: Trade openness is defined by (exports + imports)/GDP. Values at constant prices, constant exchange rates (base year 2000).

Source: OECD Macro trade indicators.

this area (see World Trade Organization, 2012, table I.4). The rise is particularly marked in Sweden and Germany, where openness is now close to 100%.

At the same time, the shares of exports (and also, but to a lesser extent, of imports) from the largest economies have shrunk over time, reflecting the strong development of trade in other parts of the world, notably in Asia. For instance, the value of exports from the European Union and United States represented respectively 24% and 28% of total exports worldwide in 1958 but only 16% and 11% respectively in 2010 (see Eurostat, 2011, table 1A), while China's exports went from 2% in 1990 to 11% in 2010. This exceptional expansion of trade in China has no equivalent among other developing economies.

This is reflected in the origin of imports and the destination of exports in the most advanced economies. The importation of manufactured goods coming from emerging economies has regularly risen but still it contributes to only part of the total increase in trade. Figure 11.2 shows how exports and imports of merchandise evolved in the OECD countries over the period 1980–2008 as a percentage of GDP, distinguishing destination/origin between advanced and developing countries. In most OECD countries growth in trade intensity from developing countries contributed less than a quarter of the total increase in merchandise imports. This means that most of the increase of imports stemmed from the development of trade between advanced economies over that period. The extent of economic integration between advanced and developing economies was much stronger in the non–European Union countries: strikingly, nearly



*Data from early 1990s to 2008.

FIGURE 11.2
Change in import and export intensity by region of origin, 1980–2008. Trade in services is not included.

Note: "Import intensity" is imports/GDP; "export intensity" is exports/GDP.

Source: OECD (2012, figure 1.3).

all the increase in merchandise imports and exports in Australia, New Zealand, Korea, and Japan over this period can be attributed to a rise in trade with developing countries. As for exports, they appear to have evolved primarily towards the most advanced countries, except in Korea and Japan where trade is particularly integrated with other Asian countries.

TABLE 11.1
The origin of imports into the European Union countries, the United States, Japan, and China in 2011.

European Union		U.S.		Japan		China	
1. China	17.3	China	18.4	China	21.5	European Union	12.1
2. Russian Federation	11.8	European Union	16.6	European Union	9.4	Japan	11.2
3. United States	10.9	Canada	14.1	United States	8.9	Korea, Rep. of	9.3
4. Norway	5.5	Mexico	11.7	Australia	6.6	Taipei	7.2
5. Switzerland	5.5	Japan	5.9	Saudi Arabia	5.9	United States	7.1

How to read: 18.4% of the imports of the United States come from China.

Note:"European Union" included all 27 countries in 2011.

Source: World Trade Organization, www.wto.org.

TABLE 11.2
The destination of exports from the European Union countries, the United States, Japan, and China in 2011.

European Union		U.S.		Japan		China	
1. United States	17.0	Canada	19.0	China	19.7	European Union	18.8
2. China	8.9	European Union	18.2	United States	15.5	United States	17.4
3. Switzerland	8.0	Mexico	13.3	European Union	11.7	Hong Kong	14.1
4. Russian Federation	7.1	China	7.0	Korea, Rep. of	8.0	Japan	7.8
5. Turkey	4.7	Japan	4.5	Taipei	6.2	Korea, Rep. of	4.4

How to read: 19% of the exports of the United States go to Canada.

Note: "European Union" included all 27 countries in 2011.

Source: World Trade Organization, www.wto.org.

As a result, China has gained a substantial share of imports into the European Union countries and the United States, as shown in table 11.1. The share held by the United States and the European Union in their respective imports is now smaller than that of China. But the same is also true, although to a lesser extent, for exports. Table 11.2 shows that China is now the first exporting market for Japan, the second exporting market for Europe, and the fourth for the United States.

1.1.2 INTERNATIONAL TRADE, UNEMPLOYMENT, AND INEQUALITIES

If we examine the structure of employment in the developing countries, we find that they do indeed have large pools of unskilled labor. Skilled labor, in contrast, is relatively rare there. The level of education is much lower in the emerging economies than in the industrialized countries. In 2010, the proportion of adult individuals (aged 24–65) with at least upper secondary schooling is around 20% in China and Indonesia, about 30% in South Africa, and 40% in Brazil, whereas it is 75% in the OECD zone and close to 90% in the United States. Of the latter two, more than 30% have tertiary education in the OECD (40% in the United States), whereas the comparable figure is only around 5% in China, Indonesia, and South Africa, and 10% in Brazil (OECD, 2012b, table A1.1a).

International Differences in the Cost of Labor in Manufacturing Industry

Table 11.3 compares the cost per hour per blue-collar worker in industry in the United States with that of certain developing countries in 1997 and 2011. We see that

Table 11.3
The cost of labor in the manufacturing industry in U.S. dollars, 1997 and 2011.

	In U.S. dollars		U.S. = 100	
	1997	2011	1997	2011
Sweden	25.0	49.1	108.6	138.3
Germany	29.2	47.4	126.6	133.4
France	24.9	42.1	107.9	118.5
Italy	19.8	36.2	85.7	101.8
Japan	22.0	35.7	95.4	100.5
United States	23.0	35.5	100.0	100.0
United Kingdom	19.3	30.8	83.7	86.6
Spain	14.0	28.4	60.5	80.1
Korea, Republic of	9.2	18.9	40.0	53.2
Brazil	7.1	11.6	30.7	32.8
Taiwan	7.0	9.3	30.6	26.3
Poland	3.2	8.8	13.7	24.9
Mexico	3.5	6.5	15.1	18.3
Philippines	1.3	2.0	5.6	5.7

Source: Bureau of Labor Statistics, www.bls.gov/fls/.

the differences are considerable. The cost of labor is about three times lower in Brazil, four times lower in Taiwan, and fifteen times lower in the Philippines. Note however that cost differences expressed in dollars do not reflect purchasing power differences. In reality, the currencies of developing countries are generally undervalued. Since workers in poor countries consume products produced locally for the most part, the differences in purchasing power are less than the differences in cost. Still, even if the developing countries have a technological lag in many areas, the size of the cost difference for low-skilled labor gives them an advantage in the production of goods requiring intensive utilization of this type of labor.

International Trade and Unemployment

The regular and massive growth in trade over the last 30 years seemed little affected by economic cycles. In all countries, unemployment varied much more substantially over that period than trade did. At first glance, OECD countries with more openness to trade also tend to feature higher unemployment rates in cross section (figure 11.3). Although weak, this correlation is often put forward in the public debate in support of the idea that trade causes unemployment. Nevertheless, this cross-section evidence is not verified when considering a larger set of countries (see figure 11.7), or longitudinal data. As shown in figure 11.4, which plots changes in unemployment against changes in trade openness over the period 1980 to 2010 across the OECD countries, we observe a negative correlation between unemployment and international trade flows in the long run. The same holds good for developing economies, and post-communist economies, for which data are available since the early 1990s (figure 11.5). This correlation observed at the macroeconomic level is difficult to interpret, as explained in greater detail below in

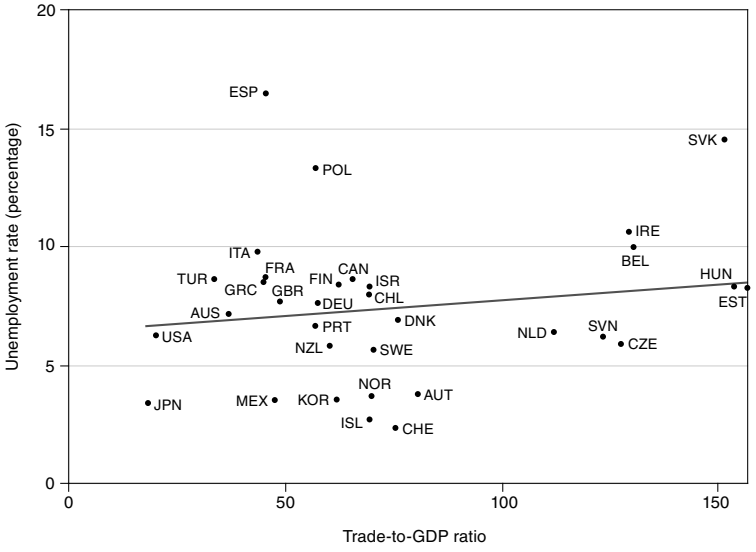


FIGURE 11.3
Unemployment and openness in the 34 OECD countries, over the period 1980–2010.

Note: Averages of unemployment rates and trade-to-GDP ratios (exports and imports/GDP) over the period 1980–2010, except for Chile, the Czech Republic, Estonia, Hungary, Israel, Poland, the Slovak Republic, and Slovenia for which the period starts between 1989 and 1996.

Source: OECD Labor Force and Trade Indicators databases.

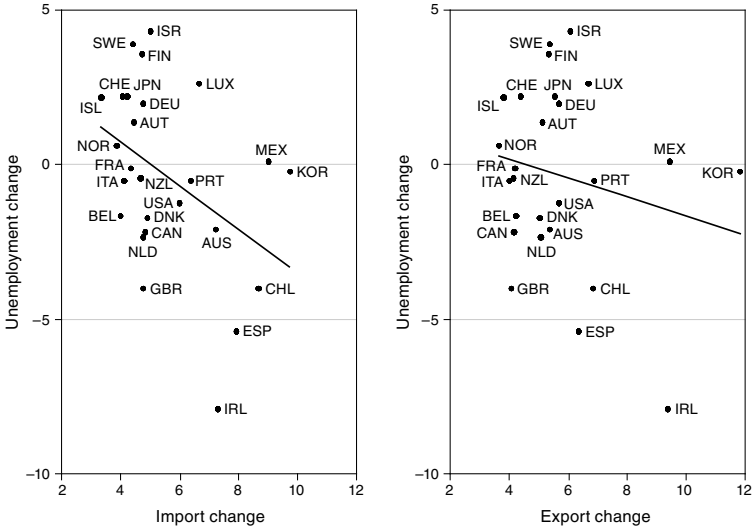


FIGURE 11.4
Change in unemployment rates and change in exports/imports in the OECD countries (excluding former communist countries), over the period 1980–2012 (percentage points for unemployment, percentage for import and export change).

Note: For imports and exports: average annual percentage change of volume of imports/exports of goods and services over the period 1980–2010; for unemployment: difference between the averages of unemployment rates over the periods 2000–2010 and 1980–1990.

Source: IMF World Economic Outlook data.

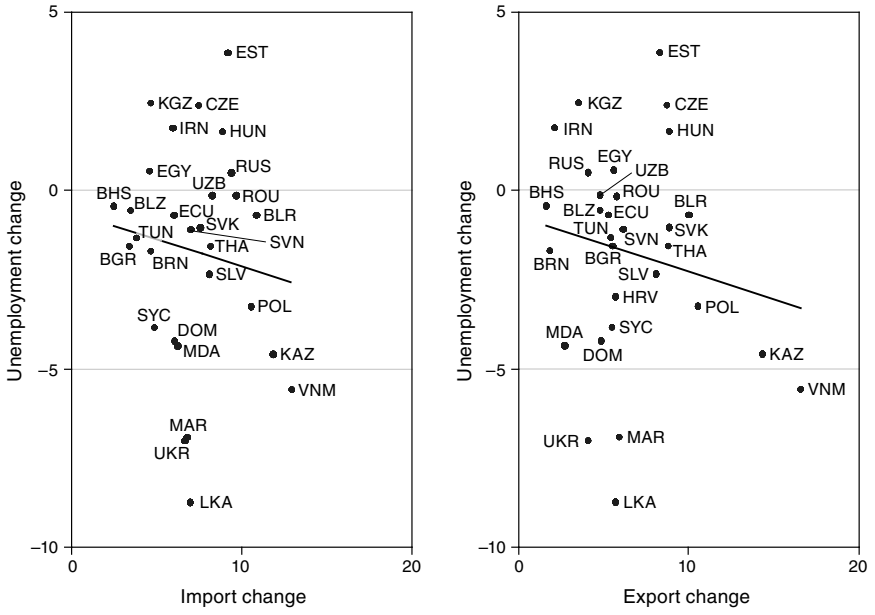


Figure 11.5

Change in unemployment rates and change in exports/imports in the non-OECD countries and in former communist countries, over the period 1990–2012 (percentage points for unemployment, percentage for import and export change).

Note: For imports and exports: average annual percentage change of volume of imports/exports of goods and services over the period 1990–2012; for unemployment: difference between the averages of unemployment rates over the periods 2007–2012 and 1990–1995. Countries selected based on the availability of data for that period.

Source: IMF World Economic Outlook data.

section 2.1. Indeed many other macroeconomic changes could be correlated with both employment/unemployment and trade, such as the development of financial markets and the development of new technologies. Furthermore, trade could be caused by unemployment as much as unemployment could be caused by trade if, for instance, policy makers react to changes in unemployment by reforming trade policies. The same type of difficulty attends the positive correlation observed in the OECD countries between the growth in trade openness and the rise in wage inequality over the last 40 years (see figure 11.8). The development of wage inequalities, possibly due to other factors such as technological change, could influence trade policies as much as trade policies could affect wage inequalities. To establish a causal link between changes in unemployment or inequality and trade, we need both a theory to figure out what kind of relationship to expect and an empirical strategy to disentangle the facts and make sure we can identify a causal relationship between these variables.

1.2 The Stolper and Samuelson Theorem

In international trade theory, each country should export goods the production of which demands the relatively intensive use of the factors of which it has a relatively abundant supply (see Krugman and Obstfeld, 2009). So increased participation by poor countries in international trade should entail an increased supply of the kind of goods that use

unskilled labor intensively and a fall in the price of those goods. International trade theory also establishes that movements in the prices of the traded goods have an impact on the prices of the inputs needed to produce these goods. The Stolper and Samuelson theorem (1947) establishes that in every country, trade liberalization entails that the real remuneration of the scarce factor is liable to decline and that of the abundant factor to rise. So, according to this theorem, the wages of the unskilled should decline in the developed countries and rise in the poor countries, whereas the wages of the skilled should rise in the rich countries and decline in the less developed ones. Yet, after reviewing the Stolper and Samuelson theorem, we see that it only holds good in particular circumstances. We also see that empirical work suggests that these circumstances may not actually come about.

1.2.1 The Closed Economy

To examine the impact of international competition on the price of the inputs, let us begin by considering a closed economy, and then open its borders. Three goods are produced: a final good, consumed by agents, and two intermediate goods used in making the final one. The final good is the numéraire, and the price of a unit of intermediate good of type i is denoted by p_i, $i = h, \ell$. Intermediate good h is produced using skilled labor alone (index h stands for high skilled), whereas intermediate good ℓ is produced using unskilled labor alone (index ℓ stands for low skilled). One unit of labor is needed to produce one unit of intermediate good in every sector. The supply of each kind of labor, denoted $N_i, i = h, \ell$, is assumed to be given. Production of the final good is represented by a concave function with constant returns $F(A_h Y_h, A_\ell Y_\ell)$, where Y_h and Y_ℓ designate the quantities of intermediate goods produced respectively by the skilled and the unskilled workers. Parameters A_h and A_ℓ are measures of technological progress that increases the efficiency of, respectively, skilled and unskilled labor.

Assuming that the market for the final good is perfectly competitive, the demands for the intermediate goods are found using the maximization problem of the representative firm in this sector:

$$\max_{\{Y_h, Y_\ell\}} F(A_h Y_h, A_\ell Y_\ell) - p_h Y_h - p_\ell Y_\ell \tag{11.1}$$

The solutions to this problem are:

$$p_i = A_i F_i(A_h Y_h, A_\ell Y_\ell), \quad i = h, \ell \tag{11.2}$$

In this expression, F_i, $i = h, \ell$, designate respectively the partial derivative of function F with respect to its first and second arguments. Assuming that the markets are perfectly competitive, we have $Y_i = N_i$, for $i = h, \ell$, and in every sector, the wage w_i equals the price p_i. Using the equilibrium conditions, $Y_i = N_i$ and $w_i = p_i$ together with the homogeneity of degree zero of the partial derivatives of function F, we arrive, with the help of (11.2), at:

$$w_i = p_i = A_i F_i(\alpha\nu, 1),\ i = h, \ell \quad \text{with} \quad \alpha = A_h/A_\ell \quad \text{and} \quad \nu = N_h/N_\ell \tag{11.3}$$

This relation entails that any increase in the relative supply of skilled labor ν reduces the price p_h and the wage w_h in the skilled sector, and has an effect of the opposite sign in the other sector.[1] The result is that the relative price p_h/p_ℓ of the good produced by the skilled workers, and the relative wage of these workers, diminish with the relative supply of skilled labor. So in countries richly endowed with skilled labor, skilled workers should get a lower relative wage than in countries poorly endowed with this type of labor.

1.2.2 The Open Economy

Let us now open up the economy, and assume that the rest of the world produces the same goods with the same technologies and is endowed with skilled and unskilled labor in quantities $\tilde{N}_h$ and $\tilde{N}_\ell$. Since all the technologies yield constant returns, production of the final good and the demand for the intermediate goods can always be obtained from the behavior of a representative firm as formalized by the optimization problem (11.1). Relation (11.2) thus continues to hold. But equilibrium in each labor market entails that the total supply of intermediate good i now equals $\tilde{N}_i + N_i$. The equilibrium conditions in the markets for goods are thus written $Y_i = \tilde{N}_i + N_i$, which, with the help of relation (11.2), gives us the equilibrium values of wages, $\bar{w}_i$, and prices, $\bar{p}_i$:

$$\bar{w}_i = \bar{p}_i = A_i F_i(\alpha\bar{\nu}, 1), \quad i = h, \ell \quad \text{with} \quad \bar{\nu} = \left(N_h + \tilde{N}_h\right) / \left(N_\ell + \tilde{N}_\ell\right) \tag{11.4}$$

Comparison of $\bar{p}_i$ and p_i tells us that the price of good h is higher after the opening of the economy if $\bar{\nu} < \nu$, that is, if the rest of the world is less intensively endowed with skilled labor. The relative price of good h does indeed rise with the ratio of unskilled to skilled labor. If the relative supply of skilled labor in the world market is inferior to that in the closed economy, then opening it up leads to an increase in the relative price of the good produced using skilled labor. Relation (11.4) illustrates the Stolper and Samuelson theorem (1947). It indicates that the increase in trade reduces the remuneration of the factors that are scarce (relative to other countries) and increases that of the factors that are abundant. According to this theorem, liberalizing trade with low-wage countries ought to increase the wage of skilled workers and reduce that of low-skilled workers in the rich countries that are well endowed with skilled labor.

1.2.3 The Limitations of the Stolper and Samuelson Theorem

The validity of the Stolper and Samuelson theorem is grounded in quite specific assumptions. This theorem assumes that all goods are traded, that the markets are perfectly competitive, and that countries have access to the same technologies. If these hypotheses are not fulfilled, the results may turn out differently. To confirm this, let us assume that countries do not use the same technologies to produce the final good: for example, let the rest of the world make use of a production function $F(\tilde{A}_h\tilde{Y}_h, \tilde{A}_\ell\tilde{Y}_\ell)$.

[1] The concavity of F entails $F_{ii} < 0$, and deriving equation (11.3), we get $\mathrm{d}w_h/\mathrm{d}\nu = \alpha A_h F_{hh}(\alpha\nu, 1) < 0$ and $\mathrm{d}w_\ell/\mathrm{d}\nu = \alpha A_\ell F_{\ell h}(\alpha\nu, 1) = -A_h(1/\alpha\nu)^2 F_{\ell\ell}(1, 1/\alpha\nu) > 0$.

In that case, the same line of reasoning as the one followed above entails that the equilibrium values of the price and the wage are now defined by:

$$\bar{w}_i = \bar{p}_i = A_i F_i \left(\frac{A_h N_h + \tilde{A}_h \tilde{N}_h}{A_\ell N_\ell + \tilde{A}_\ell \tilde{N}_\ell}, 1 \right), \quad i = h, \ell \tag{11.5}$$

As we see, the wage of skilled workers (which decreases with respect to the first argument of function F_h) no longer depends exclusively on the relative proportions of skilled and unskilled workers but also on the technologies of the two countries. If the rest of the world has a relative abundance of low-skilled labor ($\tilde{\nu} < \nu$), but if this labor is relatively less efficient than in the domestic economy ($A_\ell / A_h > \tilde{A}_\ell / \tilde{A}_h$), then it is possible to arrive at situations in which liberalization of trade with countries that abound in low-skilled manpower will lead to a rise in the wages of low-skilled workers in the domestic economy and a fall in the wages of skilled ones (if $\alpha\nu < \tilde{\alpha}\tilde{\nu}$). This example illustrates a situation in which the developed countries complement low-skilled labor with technologies more capital-intensive than the ones used in the developing countries. In this case, trade liberalization may be favorable to low-skilled workers in the industrialized countries and may help to reduce wage inequality in these countries.

These points suggest that the impact of international trade on the welfare of unskilled workers depends strongly on the structure of the economies in which they live and work. So it is not an ascertained fact that the shift in labor demand at the expense of workers with fewer skills observed in the industrialized countries is the consequence of increased participation by low-wage countries in international trade. To find out more, we must turn our attention to empirical research.

1.3 Firms' Selection and Trade

In the previous model, trade stems from differences between countries. This model predicts that exporting firms should belong to different industries across countries. Now, since the 1970s and the seminal contributions by Krugman (1979, 1980), there is abundant evidence that trade happens mostly within industries. For instance, France and Germany trade cars even though the two countries are major car producers. Krugman's model explains this prominent feature of trade by the preference of consumers for diversity in goods and services and by the existence of fixed costs of production, which lead firms to look for larger markets. In this model, even exactly identical countries would trade with one another and would benefit from it.

In Krugman's model, all firms are identical. In actuality, exporting firms have specific features that could influence the labor market: notably, they are bigger, more productive, and pay higher wages (export wage premium) than nonexporting firms (Bernard and Jensen, 2004). This is the result of the "selection effect" highlighted by Melitz (2003). In what follows, we present the basic Melitz model and discuss its consequences.

1.3.1 A Model of Trade with Monopolistic Competition

The Melitz model assumes a heterogeneity among firms with respect to productivity. Under this hypothesis, when there are fixed costs to export, only firms with sufficient productivity can reach foreign markets.

Consumers

To highlight this selection effect, let us consider a country where consumers have a preference for diversity. Assume that there is a continuum of diversified and substitutable goods denoted by I. This continuum includes goods produced domestically and goods produced abroad when there is trade openness. The preferences of a representative consumer are described by a CES utility function of the Dixit-Stiglitz (1977) kind, taking the form:

$$U = \left(\int_I C_i^{\frac{\sigma-1}{\sigma}} \mathrm{d}i \right)^{\frac{\sigma}{\sigma-1}} \tag{11.6}$$

where $\sigma > 1$ is the elasticity of substitution between two goods and C_i is the consumption of good i. The higher σ is, the more substitutable goods are, and the less market power firms have.

Maximizing U with respect to C_i subject to the budget constaint $\int_I p_i C_i \mathrm{d}i = R$, where p_i is the price of good i and R is the total available income of the representative consumer of the country in question gives the following demand function for good i:

$$Y_i = \left(\frac{p_i}{P} \right)^{-\sigma} \frac{R}{P} \tag{11.7}$$

In this expression, P represents a price index defined by:

$$P = \left(\int_I p_i^{1-\sigma} \mathrm{d}i \right)^{\frac{1}{1-\sigma}} \tag{11.8}$$

Equation (11.7) simply states that consumers break their final consumption down among all available goods, and the higher the price of the goods under consideration, the lower the demand for them. The reaction to prices is more sensitive when the elasticity of substitution is larger.

Firms

On the production side, there is a continuum of firms. Each firm produces one differentiated good i. Firms produce these differentiated goods with a linear technology using only labor but entailing a fixed overhead cost $f > 0$. The production function then reads:

$$Y_i = A_i (L_i - f) \iff L_i = f + \frac{Y_i}{A_i} \tag{11.9}$$

where A_i is the productivity parameter of firm i, while Y_i and L_i represent respectively the output and the employment level of firm i. The productivity parameters are distributed according to a distribution with positive support and no mass point over $[0; +\infty)$. The cumulative distribution function is denoted by $G(.)$. The result is that to each firm i there corresponds a level of productivity A_i, so that a good may be specified by the parameter A designating the productivity of the firm producing it. We adopt this convention in what follows.

The process by which firms make their choice is represented with a static model in two stages. In the first stage, there is a large pool of prospective entrants into the economy. Prior to entry, firms are identical. To enter, firms must pay a free entry sunk

cost, which allows them to draw their productivity parameter A and to produce a good i. In a second stage, each firm decides to produce for the domestic market and/or for the foreign market or not to produce at all. To characterize the equilibrium of this economy, we must reason by backward induction, which means beginning with the behavior of firms at the second stage.

The "Zero Cutoff Conditions"

In the second stage, a firm having drawn productivity A from within the distribution $G(.)$ has the possibility of producing a quantity $Y_d(A)$ to meet domestic demand for good A and a quantity $Y_x(A)$ to meet the demand for this good from abroad. We assume that production for foreign markets is subject to what Krugman (1980) calls an "iceberg cost" that needs to be paid for each unit of production exported and that reflects the cost incurred by shipping the merchandise, as well as tariffs and duties. Hence in order to sell quantity $Y_x(A)$, firm A must produce $\tau Y_x(A)$, with $\tau \geq 1$. Parameter τ is an indicator of the physical or tariff barriers to openness in international trade; $\tau = 1$ corresponds to a borderless world (all nations are totally integrated into international trade). Conversely, $\tau = \infty$ corresponds to a world in which all nations live in autarky. The ratio $(1/\tau)$ may be viewed as a measure of the degree of integration (or openness) of international trade. It amounts to 0 for pure autarky and to 1 for a borderless world.

The labor market is perfectly competitive. Workers are thus perfectly mobile among the different firms, and there is a single wage in the economy, denoted W. Let us designate the domestic price of good A by $p_d(A)$ and its export price by $p_x(A)$. The domestic profit, $\Pi_d(A)$, and the export profit, $\Pi_x(A)$, of firm A are then written:

$$\Pi_d(A) = p_d(A)Y_d(A) - WL_d(A) \tag{11.10}$$

$$\Pi_x(A) = p_x(A)Y_x(A) - WL_x(A) \tag{11.11}$$

In these expressions, $L_d(A)$ and $L_x(A)$ designate the levels of employment corresponding to productions $Y_d(A)$ and $Y_x(A)$. The variables L and Y are therefore linked by the production function (11.9). Note that in the expression (11.11) of the export profit, on account of the iceberg costs, $L_x(A) = f + \tau Y_x(A)/A$ represents the quantity of labor necessary to produce $\tau Y_x(A)$ so that the firm can sell $Y_x(A)$. Firm A chooses its prices $p_d(A)$ and $p_x(A)$, its levels of production $Y_d(A)$ and $Y_x(A)$, and its levels of employment $L_d(A)$ and $L_x(A)$ in such a way as to maximize its total profit $\Pi(A) = \Pi_d(A) + \Pi_x(A)$.

For ease of calculation, we will assume that the world is made up of two nations (home and abroad) that do not produce the same goods but that are otherwise identical (this assumption is not crucial to the result but greatly facilitates calculations to arrive at closed form solutions). So in our two nations, the population of labor market participants is the same size, the distribution of productivities is the same, and consumers everywhere have the same preferences, represented by the utility function (11.6). Under these conditions, the equilibrium values of the variables R, P, and W will be identical at home and abroad, and the demands for goods are written:

$$Y_d(A) = \left[\frac{p_d(A)}{P}\right]^{-\sigma} \frac{R}{P} \quad \text{and} \quad Y_x(A) = \left[\frac{p_x(A)}{P}\right]^{-\sigma} \frac{R}{P} \tag{11.12}$$

Taking into account these demand functions and the production functions described by (11.9), the maximization of the domestic profit, $\Pi_d(A)$, and the export profit, $\Pi_x(A)$, leads to the following price-setting rules:

$$p_d(A) = \frac{\sigma}{\sigma - 1}\frac{W}{A} \quad \text{and} \quad p_x(A) = \tau p_d(A) = \frac{\sigma}{\sigma - 1}\frac{\tau W}{A} \tag{11.13}$$

These rules show that firms with monopsonistic power charge a constant markup over the marginal cost. The less substitutable goods are (the closer σ is to 1), the higher this markup is. The higher the productivity A, the lower the marginal cost and the lower the price. Under this pricing rule, and given the demands (11.12) facing firms at this price, the domestic and export profits are given by:

$$\Pi_d(A) = \frac{R}{\sigma}\left(\frac{\sigma - 1}{\sigma}A\frac{P}{W}\right)^{\sigma-1} - Wf \quad \text{and} \quad \Pi_x(A) = \tau^{1-\sigma}\frac{R}{\sigma}\left(\frac{\sigma - 1}{\sigma}A\frac{P}{W}\right)^{\sigma-1} - Wf \tag{11.14}$$

The firm with productivity A engages in export if and only if $\Pi_x(A) \geq 0$. The expression of the export profit described by (11.14) shows that the latter is the case if and only if productivity is superior to a threshold A_x defined by:

$$\frac{R}{\sigma}\left(\frac{\sigma - 1}{\sigma}A_x\frac{P}{W}\right)^{\sigma-1} = \tau^{\sigma-1}Wf \tag{11.15}$$

This firm does business domestically if and only if $\Pi_d(A) \geq 0$, in other words, if and only if its productivity is superior to a threshold A_d defined by:

$$\frac{R}{\sigma}\left(\frac{\sigma - 1}{\sigma}A_d\frac{P}{W}\right)^{\sigma-1} = Wf \tag{11.16}$$

Equations (11.15) and (11.16) are known as the "zero cutoff conditions." Since $\tau \geq 1$, these equations entail $A_x \geq A_d$. In consequence, firms engaged in the export market are equally engaged in the domestic market.

A firm which has productivity such that $A \geq A_x$ realizes a total profit $\Pi(A) = \Pi_d(A) + \Pi_x(A)$. Conversely, a firm for which $A_x \geq A \geq A_d$ will do business only in the domestic market and will realize a profit $\Pi_d(A)$. Last, a firm for which $A_d \geq A$ does business in neither market. To sum up, the profit of a firm with productivity A takes the form:

$$\Pi(A) = \begin{cases} \Pi_d(A) + \Pi_x(A) & \text{if } A \geq A_x \\ \Pi_d(A) & \text{if } A_x \geq A \geq A_d \\ 0 & \text{if } A_d \geq A \end{cases} \tag{11.17}$$

Trade Equilibrium

During the first stage, firms do not know their productivity, but they do know that if they draw a productivity A at the second stage, they will obtain the levels of profit given

by (11.17). Thus they can calculate the expected profit they would obtain should they decide to enter the economy. With the help of (11.17) we find:

$$E(\Pi) = \int_{A_d}^{+\infty} \left[\frac{R}{\sigma} \left(\frac{\sigma - 1}{\sigma} A \frac{P}{W} \right)^{\sigma - 1} - Wf \right] \mathrm{d}G(A) + \int_{A_x}^{+\infty} \left[\tau^{1-\sigma} \frac{R}{\sigma} \left(\frac{\sigma - 1}{\sigma} A \frac{P}{W} \right)^{\sigma - 1} - Wf \right] \mathrm{d}G(A) \quad (11.18)$$

To enter the economy and acquire the right to a draw from the distribution of productivities, each firm must pay a fixed cost proportional to the wage, or Wh, where h designates an exogenous parameter. This fixed nonrefundable cost takes in all the expenses a firm must meet before it even knows whether it is effectively going to be able to produce and eventually export. They might include outlays on installation, recruitment, and training for example. Parameter h also takes into account the cost of a distribution network abroad or any other cost that makes it difficult to operate at a distance. The free entry condition then dictates that firms enter the economy as long as the profit outlook is positive. At the equilibrium of free entry, we thus have $E(\Pi) = Wh$.

Let N be the size, common to both nations, of the workforce and let M be the mass of goods (or firms) common to the domestic economy and the foreign one. In each economy, the mass of wages distributed is equal to WN while the mass of net profits distributed, equal to $M[E(\Pi) - Wh]$, is null, since at free entry equilibrium $E(\Pi) = Wh$ obtains. In both economies the total gains of the representative consumer then amount to WN.

Let us posit $\alpha = \frac{N}{\sigma} \left(\frac{\sigma - 1}{\sigma} \right)^{\sigma - 1}$ and let us denote the real wage by $w = (W/P)$. If we replace R by WN in equations (11.15), (11.16), and (11.18), we arrive at a system of three equations with three unknowns, w, A_d, and A_x. This system is written as follows:

$$\alpha \left(\frac{A_d}{w} \right)^{\sigma - 1} = f \quad (11.19)$$

$$\alpha \left(\frac{A_x}{w} \right)^{\sigma - 1} = \tau^{\sigma - 1} f \quad (11.20)$$

$$\int_{A_d}^{+\infty} \left[\alpha \left(\frac{A}{w} \right)^{\sigma - 1} - f \right] \mathrm{d}G(A) + \int_{A_x}^{+\infty} \left[\tau^{1-\sigma} \alpha \left(\frac{A}{w} \right)^{\sigma - 1} - f \right] \mathrm{d}G(A) = h \quad (11.21)$$

Equations (11.19) and (11.20) are the cutoff conditions and equation (11.21) is the free entry equation. It is easy to eliminate the real wage from these three equations in order to arrive at a system where the only unknowns are the thresholds A_d and A_x.

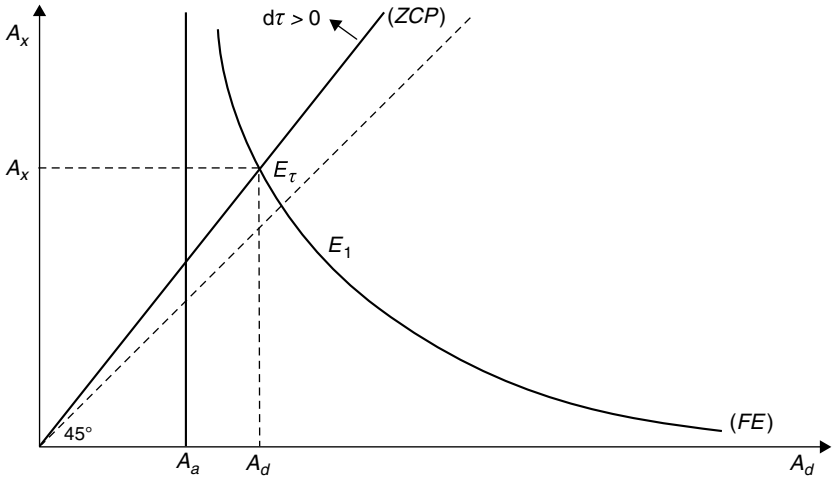


Figure 11.6
Equilibrium values of the productivity thresholds.

From (11.19) we deduce $\alpha \left(\frac{A}{w}\right)^{\sigma-1} = \left(\frac{A}{A_d}\right)^{\sigma-1} f$ and from (11.20) we deduce $\alpha \left(\frac{A}{w}\right)^{\sigma-1} = \tau^{\sigma-1} \left(\frac{A}{A_x}\right)^{\sigma-1} f$. Carrying these expressions into (11.21), we arrive at:

$$\int_{A_d}^{+\infty} \left[\left(\frac{A}{A_d} \right)^{\sigma-1} - 1 \right] \mathrm{d}G(A) + \int_{A_x}^{+\infty} \left[\left(\frac{A}{A_x} \right)^{\sigma-1} - 1 \right] \mathrm{d}G(A) = \frac{h}{f} \qquad \textbf{(11.22)}$$

Dividing equations (11.19) and (11.20) member by member, we find:

$$\frac{A_x}{A_d} = \tau \qquad \textbf{(11.23)}$$

In figure 11.6 we represent the determination of the equilibrium values of the thresholds of productivity A_d and A_x with the help of two curves. The first curve, labeled *ZCP* for Zero Cutoff Profit condition, is a straight line having (11.23) for its equation. The second, labeled *FE* for Free Entry, is described by equation (11.22). It is easy to verify that *FE* is a decreasing curve possessing a vertical asymptote at the abscissa point A_a defined by equation:

$$\int_{A_a}^{+\infty} \left[\left(\frac{A}{A_a} \right)^{\sigma-1} - 1 \right] \mathrm{d}G(A) = \frac{h}{f}$$

The value A_a represents the threshold of productivity in an autarkic economy (characterized by $\tau = +\infty$ and $A_x = +\infty$). The equilibrium corresponding to a degree of integration $1/\tau$ in international trade is represented by point E_τ. Point E_1 situated at the intersection of the first bisectrix of the curve *FE* corresponds to totally integrated international trade ($\tau = 1$). In this situation, we have $A_d = A_x$, which signifies that all the firms remaining in the market are engaged in export.

1.3.2 THE CONSEQUENCES OF OPENNESS

The most interesting result of this model is to show that opening up to international trade causes firms to undergo a selection according to their productivity. This selection effect redistributes the workforce toward firms that are open to international trade, but it also destroys less productive firms.

The Selection Effect

First, more trade liberalization, through lower trade cost τ, is represented in figure 11.6 by a rotation of the straight line (*ZCP*) toward the bottom. We see that that entails a lowering of the productivity threshold for operating in foreign markets, A_x, because incentives to export are strengthened. As a result, existing high-productivity firms expand and hire more workers to serve foreign markets, and new high-productivity firms enter, attracted by the prospect of higher profits. Second, trade liberalization increases the productivity threshold for operating on the domestic market, A_d, because the drop in trade costs increases the demand for labor and then increases wages for all firms (including for those with low productivity that might have to exit the market); at the same time it increases competition, with more productive foreign firms entering the domestic market. This drives the less productive firms in the domestic market out of business. As a result, market shares are reallocated towards exporters in all trading economies: there are more exporters at the right end of the productivity distribution. Employment is therefore reallocated towards exporters.

Another particularly interesting result is that the average productivity of the economy rises when integration into world trade increases. By definition the average productivity of all firms (exporters and nonexporters) is given by:

$$\bar{A} = \frac{1}{1 - G(A_d)} \int_{A_d}^{+\infty} A \mathrm{d}G(A)$$

The right member of this last equation being an increasing function[2] of A_d, and A_d being decreasing with τ, we deduce that average productivity $\bar{A}$ diminishes with τ, which means that it increases with the degree that the economy is opened up (equal to $1/\tau$). This result is the direct upshot of the process of selection that occurs in the wake of increased openness to international trade, which eliminates the less productive firms.

To close the model, we must still characterize the number M of firms (for each country) that enter the economy in stage 1. To that end, it suffices to return to the definition (11.8) of the price index. In this definition, I represents the sum of domestic and foreign goods consumed in each country. Given that only the firms within a country having a productivity superior to A_d produce, and that the inhabitants of a given

[2] Let us consider the function:

$$F(x) = \frac{1}{1 - G(x)} \int_{x}^{+\infty} A \mathrm{d}G(A)$$

We have:

$$F'(x) = \frac{g(x)}{1 - G(x)} \left[\frac{1}{1 - G(x)} \int_{x}^{+\infty} A \mathrm{d}G(A) - x \right]$$

Since $\frac{1}{1-G(x)} \int_{x}^{+\infty} A \mathrm{d}G(A) > x$, we get $F'(x) > 0$.

country consume foreign goods produced by firms that have productivity superior to A_x, we have:

$$P^{1-\sigma} = \int_I p_i^{1-\sigma} \mathrm{d}i = \int_{A_d}^{+\infty} [p_d(A)]^{1-\sigma} M \mathrm{d}G(A) + \int_{A_x}^{+\infty} [p_x(A)]^{1-\sigma} M \mathrm{d}G(A)$$

Taking into account the equilibrium values of domestic prices and export prices given by equation (11.13), we arrive after several simple calculations at:

$$P^{1-\sigma} = M \left(\frac{\sigma}{\sigma - 1} W \right)^{1-\sigma} \left[\int_{A_d}^{+\infty} \left(\frac{1}{A} \right)^{1-\sigma} \mathrm{d}G(A) + \tau^{1-\sigma} \int_{A_x}^{+\infty} \left(\frac{1}{A} \right)^{1-\sigma} \mathrm{d}G(A) \right]$$

With the help of relations (11.19) and (11.20), and recalling that $\alpha = \frac{N}{\sigma} \left(\frac{\sigma - 1}{\sigma} \right)^{\sigma - 1}$, we arrive finally at:

$$1 = \sigma f \frac{M}{N} \left[\int_{A_d}^{+\infty} \left(\frac{A}{A_d} \right)^{\sigma - 1} \mathrm{d}G(A) + \int_{A_x}^{+\infty} \left(\frac{A}{A_x} \right)^{\sigma - 1} \mathrm{d}G(A) \right] \qquad \textbf{(11.24)}$$

Knowing the values of A_d and A_x determined by the system (11.23) and (11.22), equation (11.24) then permits us to know the value of M. A priori, trade has an ambiguous impact on the total number of firms, M, since every potential entrepreneur anticipates that greater openness to international trade will eliminate less productive firms but increase the market share of the more productive ones. Melitz (2003) shows nonetheless that trade increases the total number of goods available for consumers thanks to foreign firms which produce goods that are not to be found in an autarkic economy.

Trade Liberalization and Employment

This simple model tells us nothing about the effects of opening up to international trade on unemployment, since by construction, the whole workforce is employed. We may discern however that wage flexibility plays a crucial role as the labor market adjusts to more openness in international trade. To do so, let us revert to the system of three equations (11.19), (11.20), and (11.21), and let us assume henceforth that the real wage w is exogenous, but that the level of employment N becomes an endogenous variable. That being the case, this system determines the thresholds of productivity A_d and A_x and the level of employment N by means of the variable $\alpha = \frac{N}{\sigma} \left(\frac{\sigma - 1}{\sigma} \right)^{\sigma - 1}$. We see that the thresholds of productivity are the same as under the hypothesis of wage flexibility. They continue to be defined by equations (11.22) and (11.23). Equation (11.19) then shows that, if the real wage is exogenous, the level of employment N varies inversely with respect to threshold A_d. Greater openness to international trade then tends to diminish the level of employment. If wages cannot adjust, the destruction of jobs in low-productivity firms is not offset by job creation in high-productivity exporting firms. For the effects of greater openness to international trade not to be negative, it is necessary for wages to be able to adjust in all sectors of the economy.

Several authors have looked more closely at the effect on the level of unemployment of opening up to international trade; they integrate a labor market with friction

where wages are bargained over at the moment firms and workers match up (see chapter 9) into the model of Melitz (2003). With this approach, Felbermayr et al. (2011a) find that trade liberalization lowers unemployment and raises real wages as long as it improves aggregate productivity (net of transport cost) due to the selection effect. In the model of Helpman and Itskhoki (2010), the economy is composed of two sectors with specific labor markets. In the nonexporting, homogeneous good sector, the labor market features low search frictions, while higher frictions obtain in the exporting, differentiated good sector. In this setting, Helpman and Itskhoki (2010) show that the relationship between trade and unemployment can be hump shaped: if the labor market of the exporting sector is particularly "rigid" compared with that of the nonexporting sector, that is, if it features a slower adjustment of the labor force, unemployment is higher than in the nonexporting sector. Then trade liberalization decreases unemployment in the exporting sector but it will also tend to reallocate the workforce to this higher-unemployment segment of the economy. The composition effect can dominate if trade liberalization is not substantial enough to compensate by lowering the unemployment rate sufficiently in the exporting sector.

The impact of trade on unemployment may be less clear-cut when labor markets differ across trading partners. Our simple model assumes symmetric trading partners. But labor market institutions often vary significantly among countries. Frictions in the labor market impeding a rapid adjustment of employment can be more pronounced in some countries. Resolving trade models with asymmetric countries is typically more complex. In their model, Helpman and Itskhoki (2010) show that, despite the labor market frictions, both countries gain from trade, but a country's welfare level depends on labor market frictions in its trade-partner country in addition to its own.

Also, the size of countries matters in explaining the influence of labor market frictions in relation to trade. Felbermayr et al. (2013) show that relatively larger and less open economies are harmed more by their own labor market frictions, whereas smaller and more open economies are hit relatively harder by foreign labor market frictions and less by their own. This is because of an income effect: when a country's domestic demand falls due to high unemployment (i.e., larger labor market frictions), so must its demand for foreign goods. The larger the country (and the lower the trade barriers), the stronger the effect on trade partners. This is not the case for small, open economies whose labor market has little impact on trade partners, and hence on its ability to export.

Impact on Wage Inequality

This selection effect also has implications for wage inequality, since differences in efficiency could in principle imply differences in wages. In the Melitz (2003) model studied previously, the heterogeneity of productivities does not generate a distribution of wages because the labor market is competitive: more-productive firms expand and pay higher wages but less-productive firms must pay the same wage or exit the market. This feature is at odds with the finding that more-productive firms pay higher wages.

To bypass this limitation, Helpman et al. (2010a) have introduced search and matching frictions into the Melitz model. In this context, the opening of a closed economy to trade raises wage inequality because larger firms pay higher wages and the opening of trade increases the dispersion of firm revenues, which in turn increases the dispersion of firm wages. However, once the economy is open to trade, the relationship between wage inequality and trade openness is at first increasing and then

decreasing. Indeed, when no firm exports, a small reduction in trade costs increases wage inequality because it induces some firms to export and raises the wages paid by these exporting firms relative to domestic firms. At the other extreme, when all firms export, a small increase in trade costs raises wage inequality because it induces some firms to cease exporting and reduces the wages paid by these domestic firms relative to exporting firms. Using the same framework, Helpman et al. (2010b) show further that workers with intermediate ability lose more than others when trade is liberalized, because the opening of economies to trade leads to a reallocation of employment toward exporting firms: such firms select their employees more rigorously, which reduces the wages workers with intermediate ability can receive. In comparison, trade liberalization does not affect low-productivity workers (hired only by low-productivity firms, not involved in trade) or high-productivity workers (who will always enjoy priority in hiring by high-productivity firms, with or without trade). The selection effect at work in the Melitz model associated with search frictions suggests that wage inequality ought to occur within sectors or occupations. This is in line with empirical evidence presented in the following section (see also chapter 10).

2 INTERNATIONAL TRADE AND LABOR MARKETS: EMPIRICAL EVIDENCE

Empirical research on the labor market effects of international trade has long been influenced by the Heckscher-Ohlin–Stolper-Samuelson model, which yields predictions about changes in relative wages across skill groups and also unemployment (once search frictions are introduced). Since the developed economies of the OECD are well equipped with capital (human, financial, physical) compared to emerging countries, whereas the (relatively) scarce factor is unskilled labor, the Stolper and Samuelson theorem predicts that whenever developed economies engage in trade with emerging or developing economies, the unskilled workers of developed economies are expected to lose, while owners of human, physical, and financial capital are expected to gain. If labor markets are rigid, the loss will make itself felt through rising unemployment, while if labor markets are flexible, the loss will take the form of lower incomes.

The Stolper-Samuelson model does not fit the evidence very well. Empirical studies tell us that at the macro level, more trade is associated with less unemployment, not more, at least in the long run. Moreover, empirical studies into the 2000s have typically concluded that the impact of trade on wage inequality is modest at best, and it happens not across sectors or countries but across plants and firms within sectors, and in both developed and developing countries. This is consistent with the fact that trade is mostly intra-industry and driven by product differentiation (Krugman, 1980; Melitz, 2003), inducing reallocation of factors between firms within a sector.

2.1 Empirical Evidence at the Macro Level

A first approach is to relate macroeconomic outcomes directly to trade openness using macroeconomic data. Two types of data can be used for macroeconomic evaluation depending on availability: cross-section databases, where information is only observed

at one point in time across many countries, and panel databases, where the information is available for different points in time across many countries. We start with cross-section data and then review the studies that use panel data. To illustrate the method, we use the framework of Dutt et al. (2009). The related data and programs are available at www.labor-economics.org.

2.1.1 The Basic Regression

Cross-national data are the most common set of information available in international databases, notably when we want to include in our analysis non-OECD countries for which reliable statistics are less frequent. A basic regression with cross-section data is:

$$y_i = \alpha + \beta T_i + \mathbf{X}_i \boldsymbol{\gamma} + \varepsilon_i \tag{11.25}$$

where y_i is a measure of unemployment or income/wage inequality in country i, T_i is a measure of trade such as trade openness (i.e., imports and exports as a percentage of GDP), import penetration (imports as a percentage of domestic demand), import duties, or offshoring (ratio of total imported intermediate goods purchased to GDP). The vector $\mathbf{X}_i$ represents a set of controls, such as labor market institutions, demography (total population or working-age population), and the business cycle (output gap). These variables are usually averaged over several years, which has the advantage of neutralizing the effect of the business cycle. The variable ε_i is an error term and β and $\boldsymbol{\gamma}$ are a set of parameters to be estimated. This equation only yields a nonbiased estimate of the impact of trade on the outcome if $\mathbb{E}(T|\varepsilon) = 0$. This might not be the case for several reasons. Consider for instance the case where y_i measures unemployment.

- First, some important variables influencing both trade and unemployment may have been omitted from the regression (11.25). For instance, good macroeconomic policies might lead to more trade openness and less unemployment. The same holds good for product or labor market regulations. Omitting these variables would yield a spurious positive relationship between trade and unemployment, and we would attribute to trade what ought to be attributed to other policies. This can be dealt with by including some key variables accounting for macroeconomic policies or product and labor market policies as part of the set of control variables.
- Second, there may be reverse causality; trade may be caused by unemployment rather than unemployment being caused by trade. For instance, if policy makers erect trade barriers as a response to unemployment shocks, this would yield a spurious negative correlation between the level of trade and unemployment. In this cross-section setting, this can be dealt with by instrumenting trade with variables that are not related to unemployment but that explain trade. This may be the case for instance with certain geographical variables such as the distance between trade partners.

Moreover, the unemployment rate and trade, like any other economywide outcomes, are subject to measurement error, especially with a large set of countries including a number of developing countries. Measurement error tends to attenuate the

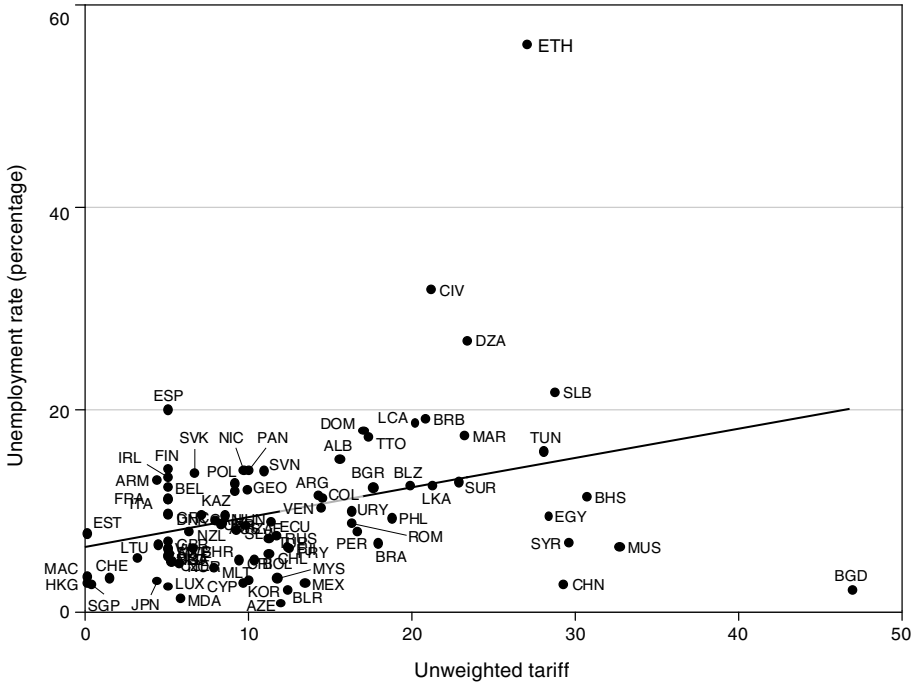


Figure 11.7
Tariffs and unemployment rate (1990s, 90 countries).

Note: "Unweighted tariff" is the unweighted average of rates of customs duty.

Source: Dutt et al. (2009, figure 1).

relationship between trade and unemployment. This risk can be reduced by instrumenting trade (if trade is the variable that is poorly measured) or by running a number of robustness checks with various measures of trade and unemployment. Using period average also tends to reduce the problems of measurement errors.

Estimating a basic regression with cross-section data such as equation (11.25), and considering tariffs as a measure of openness to international trade (i.e., an unweighted average of rates of customs duty on merchandise imports), Dutt et al. (2009) obtain a negative and significant effect of openness to international trade on the unemployment rate. This is apparent in figure 11.7 but comes out even more clearly in the estimations of Dutt et al. reproduced in table 11.4. In addition to tariffs, the authors add two alternative measures of international trade: openness (the sum of imports and exports as a percentage of GDP) and import duties, which is a weighted average of the duties collected on imports for each good as a percentage of total imports. In the first estimation, the unemployment rate is regressed on an indicator of trade and in the following estimations control variables are added in conformity with equation (11.25). Finally, in the last two estimations of table 11.4, the measure of trade is instrumented according to the following equation:

$$T_i = \mathbf{Z}_i\boldsymbol{\delta} + \eta_i$$

TABLE 11.4
The effect of trade policies on the unemployment rate across countries.

	OLS			IV			IV	
T_i = Unweighted tariff	.351***			.750**			.659*	
T_i = Openness		−.024*			−.065**			
T_i = Import duty			.492***			.664***		.453**
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Participation	No	No	No	No	No	No	Yes	Yes
Observations	55	55	54	44	55	43	44	43
R^2	.28	.20	.33	.	.	.	.	.

Note: Controls include the GDP, the output volatility, EPL index, labor union power index, working-age population, civil liberties, and black market premium; participation includes the labor market participation rate and the female labor market participation rate. *, **, ***: significant at the 10%, 5%, and 1% level respectively.

Source: Dutt et al. (2009, tables 2, 3, 4, and 5).

where $\mathbf{Z}$ is a set of instrumental variables that influence trade but are not correlated with unemployment, such as country size (the smaller the country, the more trade openness), distance between trade partners (the closer they are, the greater the trade), and other geographical determinants, while η_i is an error term. In addition, Dutt et al. control for variations in labor market participation, for as we saw in chapter 9, the latter can influence variations in the rate of unemployment.

The key identifying assumption is that the instruments are uncorrelated with the residuals in equation (11.25), that is, $\mathbb{E}(\mathbf{Z}|\varepsilon) = 0$. For the measures of trade based on tariffs and import duties collected, the authors use as instruments the number of years the country has remained outside GATT (General Agreement on Tariffs and Trade) since its inception in 1948 and outside the World Trade Organization (WTO) since its inception in 1995, and also the proportion of tax revenues obtained from taxes on domestic activities. In fact, the longer a country is outside the GATT, the larger its protectionism, and more developed countries are better able to collect tax on domestic income than less advanced countries that rely more on duties—and thus have a motive for maintaining a certain degree of protectionism. For the measure of openness, the authors make use of geographical variables like the distance between trade partners, the size of the domestic market, and other geographical variables that have no a priori connection to the unemployment rate (Frankel and Romer, 1999).

Whatever the indicator of trade, the use of instrumental variables does not change the significance of the results obtained by OLS. Neither does the significance of the results change if labor force participation is instrumented by the mortality rate or the rate of prevalence of HIV. This verification is not pointless, for participation in the labor market may be influenced by trade, and for that matter Dutt et al. observe that more trade is associated with higher participation. As a general rule, in these regressions the control variables relative to demography, GDP, and the labor market (with the exception of labor union power, which tends to increase unemployment) are rarely significant.

To verify whether trade has a differential impact on the labor market outcomes of advanced and developing economies, one possibility is to allow the coefficient of the trade variable to vary according to the level of capital per capita, which is also a close

TABLE 11.5
The effect of trade policies on the unemployment rate depending on the capital-to-labor ratio.

OLS	T_i = Unweighted tariff	T_i = Openness	T_i = Import duty
Trade measure	.227	.158	3.824**
Trade measure × capital-labor ratio	.015	−.017	−.349**
Capital-labor ratio	1.427	1.350	4.521**
Controls	Yes	Yes	Yes
Observations	48	48	47
R^2	.31	.27	.42

Note: Controls include the GDP, the output volatility, EPL index, labor union power index, working-age population, civil liberties, and black market premium. **: significant at the 5% level.

Source: Dutt et al. (2009, table 6).

measure of the share of skilled labor in the economy (skilled labor being a complement to capital). Let us denote by K_i/L_i the level of capital per head in country i; equation (11.25) is then written:

$$y_i = \alpha + \beta_1 T_i + \beta_2 T_i (K_i/L_i) + \beta_3 K_i/L_i + \mathbf{X}_i \boldsymbol{\gamma} + \boldsymbol{\varepsilon}_i \quad \textbf{(11.26)}$$

In this setting the marginal impact of trade on unemployment is $\beta_1 + \beta_2 (K_i/L_i)$. If the Stolper-Samuelson theorem is verified, then trade restriction should increase unemployment in high capital-per-head countries, and we should have $\beta_1 > 0$ and $\beta_2 < 0$ such that $\beta_1 + \beta_2 (K_i/L_i) < 0$ if $K_i/L_i > (K/L)^*$ where $(K/L)^*$ is the turning point capital–labor ratio, equal to $-\beta_1/\beta_2$, at which trade starts to have a positive impact on unemployment. This turning point is determined endogenously from the data, given the estimates of β_1 and β_2. The results of the OLS regressions of Dutt et al. (2009) using various indicators for trade policies are reported in table 11.5. There is little support for the Stolper-Samuelson theorem: coefficients are rather insignificant and/or of the wrong sign. Higher tariffs do not lower unemployment in high capital-to-labor countries, nor does more openness increase it. Only higher import duties seem to be associated to lower unemployment, the higher the capital-to-labor ratio, but the authors show that this result does not hold when the measure of trade is instrumented.

2.1.2 THE ADVANTAGES AND DRAWBACKS OF PANEL DATA ANALYSIS

Cross-section analyses have several limitations. First, they do not make it possible to identify how shifts in trade policies impact macroeconomic outcomes within countries over time. Arguably, the short-term impact on unemployment, for instance, might differ from the long-run (steady-state) impact predicted by the typical models reviewed previously. Second, when the data available are in panel form, equation (11.25) can be augmented with country effects so as to account for time-invariant, unobserved heterogeneity. This can be extremely useful because country specificities such as geography, stable institutional setting, or political regime can influence unemployment. Adding a longitudinal dimension to the analysis, and taking into account the persistence of

some macroeconomic outcomes like unemployment and inequality, equation (11.25) becomes:

$$y_{it} = \sum_{s=1}^{S} \rho_s y_{i,t-s} + \beta T_{it} + \mathbf{X}_{it}\boldsymbol{\gamma} + \mu_i + \varepsilon_{it} \tag{11.27}$$

In this equation, all the previous variables now have a time dimension, so that i is the index for the country and t is the index for time. The dependent lagged variable $y_{i,t-s}$ has been added to account for the persistence of the dependent variable over time (with S denoting the total number of lags) and a country-specific effect is denoted by μ_i. In this setting, the problems reviewed previously for equation (11.25) are still potentially present, but they take different forms and can be dealt with in slightly different ways:

- First, the business cycle fluctuations present in time series heighten the difficulty of interpreting correlations between trade and unemployment or wages. For instance, any positive shock to domestic spending is likely to increase domestic as well as import demand, and thus both lower unemployment and increase openness. It is therefore essential to control for the business cycle by including the GDP gap in the set of control variables. Also, averaging the data over periods of several years can help smooth out the variations due to the business cycle; but this has the disadvantage of reducing the size of the sample.
- Second, omitted variables that do not vary over the sample can be controlled for by the country-fixed effects.
- Third, the reverse causality problem, still present in this setup, is addressed by using the time dimension of the data. In equation (11.27), the measure of trade can be instrumented by past values, which cannot possibly be influenced by the current level of the dependent variable.

This equation cannot be estimated reliably with a simple OLS because the regressors $y_{i,t-s}$ are the lagged dependent variable, which gives rise to autocorrelation of residuals. Finally, and typical of macroeconomic panels, the time dimension is generally smaller than the number of countries. This makes the autocorrelation problem more complex to deal with. In panels with large time dimensions the correlation of the lagged dependent variable with the error term would become insignificant over time.

The Arellano-Bond (GMM) Estimator

The method comes down to differencing both sides of equation (11.27), then looking for all available instrumental variables (IV) for the endogenous variables, and using the general method of moments (GMM) to estimate coefficients. Let us consider equation (11.27) with only one lagged dependent variable ($S = 1$), and let us temporarily drop the set of controls $\mathbf{X}$ to simplify the presentation. Its first difference reads:

$$\Delta y_{it} = \rho \Delta y_{i,t-1} + \beta \Delta T_{it} + \Delta \varepsilon_{it} \tag{11.28}$$

whereby definition $\Delta x_{it} = x_{it} - x_{i,t-1}$. Note that in equation (11.28) the time-invariant, fixed, country-specific effect μ_i is removed. This allows us to get rid of the correlation

problem between $y_{i,t-1}$ and μ_i. But $y_{i,t-1}$ is still a function of $\varepsilon_{i,t-1}$, and so $\Delta y_{i,t-1}$ is unfortunately correlated with $\Delta\varepsilon_{it} = \varepsilon_{it} - \varepsilon_{i,t-1}$. This requires IV methods for consistent estimation. Arellano and Bond (1991) set out to use two or more lags of y_{it} as IVs for $\Delta y_{i,t-1}$ (see also Wooldridge, 2010, chapter 11). Let us denote by $t = 1$ the first date for which data are available. The first period where we can use an instrumental variable is $t = 3$. At this date, equation (11.28) reads:

$$\Delta y_{i3} = \rho \Delta y_{i2} + \beta \Delta T_{i3} + \Delta \varepsilon_{i3}$$

We can use y_{i1} as an instrument for Δy_{i2} since it is uncorrelated with $\Delta\varepsilon_{i3}$ (but not y_{i2}, which is correlated with ε_{i2} and hence $\Delta\varepsilon_{i3}$). Similarly in $t = 4$, we can use y_{i1} and y_{i2} as instruments for Δy_{i3}, and so forth. This is possible under the crucial identifying assumption stating that $Cov(y_{is}, \Delta\varepsilon_{it}) = 0$ for any $s \in [1, t-2]$, $t \geq 3$, that is, that the instrument is exogenous, which amounts to assuming that the residual ε_{it} is not autocorrelated. Hence, the set of all possible instruments for $\Delta y_{i,t-1}$ in any period t is $(y_{i,t-2}, y_{i,s-3}, \ldots y_{i1})$. This means that the set of instrumental variables is larger than the number of coefficients to estimate in the model and that it is not balanced since the number of instruments can vary over time. Thus we must resort to the general method of moments (GMM) as a method of estimation. This method is presented in the appendix at the end of this chapter along with some specification tests. Most statistical software contains modules that perform the GMM estimations, along with specification tests such as the autocorrelation of residuals.

Results with Panel Data

For the panel data analysis, Dutt et al. (2009) use time dummies as a measure of the trade variable T_{it}. These time dummies identify permanent trade liberalization periods. As such $T_{it} = 1$ after trade liberalization and 0 before. They also include the lagged trade liberalization dummies, $T_{i,t-1}$ to $T_{i,t-4}$ in order to allow the response of unemployment to trade policies to vary over time. Using this measure, the authors again find support for the view that unemployment falls in the wake of trade liberalization, as shown in table 11.6. However, unemployment falls after an initial phase of increase the year immediately after liberalization occurs: the coefficient of the T_{it} is positive, and the coefficients of the first two lags of trade liberalization are negative and significant. The OLS results are stable to the introduction of fixed effects (column 2), and similar to the estimation with the Arellano-Bond Method (column 3), with the trade measure considered as endogenous (in which case the trade dummy is instrumented by its lagged values using the same approach as the lagged unemployment), even when labor market participation is introduced (and instrumented as well). Overall, the results of Dutt et al. (2009) show that over the 1985–2004 period, unemployment is correlated negatively not only to international trade across countries (cross-section analysis) but also within countries (panel analysis): a trade liberalization episode is associated, although with some delay, to a decline in unemployment over time.

2.1.3 Trade and Unemployment

As Davidson and Matusz (2004) state, whether trade affects the level of equilibrium unemployment is "primarily an empirical issue." Yet the number of studies on this

TABLE 11.6
The effect of trade policies on the unemployment rate within countries.

	OLS	OLS, FE	GMM	GMM
$y_{i,t-1}$ = lagged unemployment	.963***	.773***	.616***	.597***
T_{it} = liberalization dummy	.814**	.701*	.925***	.818***
$T_{i,t-1}$ = lagged liberalization dummy	−.841*	−.664*	−1.549**	−1.346***
$T_{i,t-2}$ = lagged liberalization dummy	−.756*	−.653*	−.481*	−.838**
Controls (output, demography, labor market)	Yes	Yes	Yes	Yes
Labor market participation	No	No	No	Yes
Observations	1096	1096	1011	1011
Number of countries	73	73	72	72

Note: In the GMM estimates, trade liberalization and labor force participation are treated as endogenous. OLS, FE is the OLS method with country-fixed effects (*, **, ***: significant at the 10%, 5%, and 1% level respectively).

Source: Dutt et al. (2009, table 7).

topic remains small, and the same is true for the effect of trade on inequality. These studies tend to confirm the conclusions of Dutt et al. (2009).

For instance, Felbermayr et al. (2011b) perform panel data regressions for a set of 20 OECD countries and cross-section regressions for a larger set of 60 countries over the period 1990–2007, using openness as a trade measure. After controlling for the endogeneity of the trade measures (using instrumental variables), business cycle effects, and a host of institutional and geographical settings, they find in most of their regressions that unemployment decreases with trade openness—never that it increases. And this result is not biased upward by endogeneity in the trade measures. The decline in unemployment is mostly driven by lower unemployment among skilled workers. Using a panel regression for 20 OECD countries over the period 1982–2003, with 5-year averages to mitigate business cycle concerns, and a real measure of openness,[3] they find that a 10 percentage point increase in trade openness reduces unemployment by about three quarters of one percentage point. With a larger set of cross-sectional data in 62 countries (using averages of variables over 1990–2006 to control for the business cycle and for lower quality of data) they find that a 10 percentage point increase in trade openness reduces unemployment by one percentage point. The results from OLS regression are very close to those obtained with IV methods (where openness is instrumented by demographic and geographical variables). These results are robust to various measures of trade and unemployment.

Moreover, as predicted by Helpman and Itskhoki (2010) and Felbermayr et al. (2013), the impact of international trade on unemployment might be influenced by labor market regulations. Trade acts as a vehicle through which the labor market in

[3]The measure of trade openness most often used in empirical work is nominal imports plus exports relative to nominal GDP. Alcalá and Ciccone (2004) argue that the Balassa–Samuelson effect distorts nominal openness measures since countries with low labor productivity, and hence a high price of traded relative to nontraded goods, have artificially high degrees of openness. They propose to use real openness defined as imports plus exports in US$ relative to GDP in purchasing-power-parity US$ (PPP).

one country can affect unemployment in its trading partners. When unemployment increases in one country, due for instance to a higher tax wedge on labor or other detrimental institutional settings, domestic income falls, and this also hurts trading partners. For instance, based on a panel of 20 OECD countries, and controlling for business cycle co-movements in trade and unemployment, Felbermayr et al. (2013) show that more severe labor market search frictions (such as those summarized by the parameters of the matching function in chapter 9) in trading partners increase domestic unemployment. Larger trading partners spill more of the effect of their labor market search frictions over onto their trading partners, and more open economies are more sensitive to their partners' unemployment. This tends to invalidate the relevance of the "beggar-thy-neighbor" assumption, by which one country may attempt to remedy its own problems (e.g., by reducing labor market frictions) in ways that tend to worsen the problems of its partners, an assumption that some models highlight (e.g., Helpman and Itskhoki, 2010).

2.1.4 Trade and Productivity

Trade is positively correlated with average per capita income. In a seminal paper, Frankel and Romer (1999) used a sample of 150 countries to analyze the influence of trade on per capita income in 1985. They first instrumented trade in equation (11.25) by geographical determinants such as the distance between countries' principal cities, common borders, the distance to the equator, and the identification of landlocked countries. More precisely, they predicted bilateral trade shares (as a % of GDP) between countries based on geographical determinants and then aggregate predicted trade shares to obtain a geography-based instrument for trade. They also controlled for within-country trade, proxied by the size of the domestic market. In the controls they included only country size and population, arguing that even though many other factors might influence income, they might also be influenced by trade and thus interfere with the impact of trade on income (see the discussion on bad controls in chapter 8). They find that the effect of trade on per capita income is positive and significant, and rises when trade is instrumented by geographical variables compared with OLS estimates, suggesting that OLS understate rather than overstate the effect of trade, as is the case in Dutt et al. (2009) for unemployment. Their estimates imply that a one-percentage-point increase in the trade share raises income per capita by 2 percent. Based on the same instrumental approach, Alcalá and Ciccone (2004) find a consistent impact of trade on productivity, measured as GDP per worker, and use real openness (imports plus exports relative to *purchasing power parity* GDP) as a measure of trade. They find that the elasticity of productivity to real openness is around 1.2, and taking a country from the 30th percentile to the median value of openness raises productivity by 80%.

2.1.5 Trade and Inequality

The impact of trade on wage inequality is less clear-cut. As shown in figure 11.8, there is a moderate cross-country positive correlation between changes in trade exposure (a weighted average of import penetration and export intensity, a measure close to trade openness) and change in wage dispersion (defined here as the decile 9/decile 1 ratio of weekly earnings among full-time workers). Moreover, this correlation is driven by a few countries. OECD (2012a) analyzed the impact of trade on wage dispersion, based on annual time-series data covering 22 countries from the early 1980s to 2008.

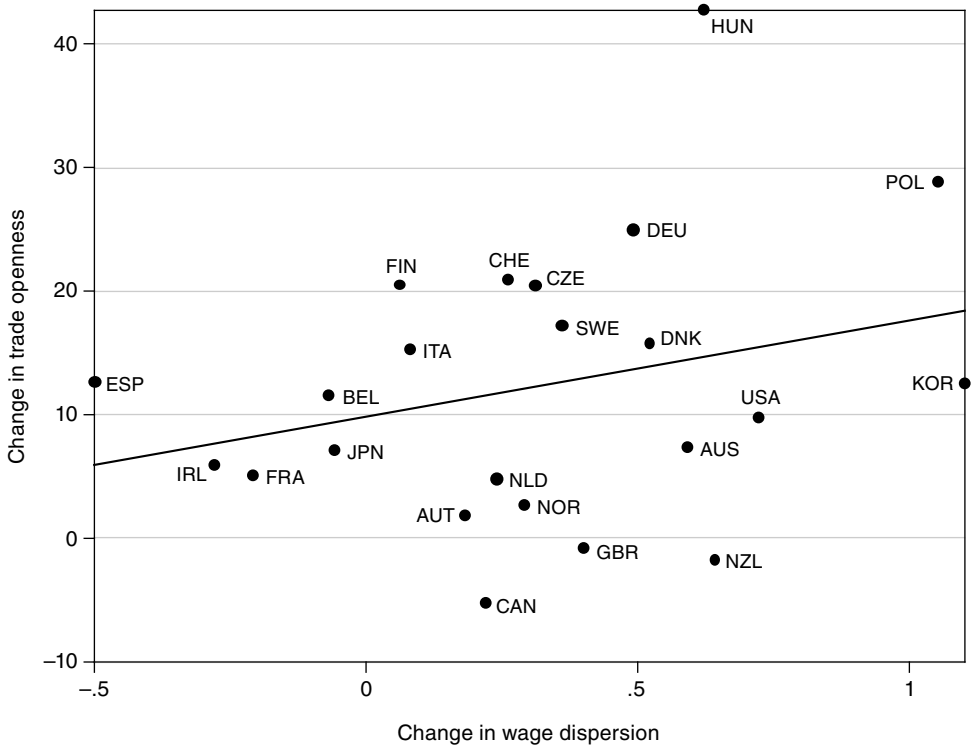


FIGURE 11.8
Trends in wage dispersion and trade openness (1985–2007, 23 countries). (Percentage points.)

Note: Trade exposure is a weighted average of export intensity (exports as a % of GDP) and import penetration (imports as a % of domestic demand); wage dispersion is the D9/D1 ratio for full-time weekly earnings. Data start in the mid-1990s for the Czech Republic, Ireland, Hungary, Norway, Poland, Spain, and Switzerland.

Source: OECD (2012, figure 1.5).

This panel allows the researchers to control for unobserved time-invariant heterogeneity across countries. The analysis focuses on within-country variation in inequality, relating changes in wage dispersion to various channels through which globalization might operate, and to technology and policy factors that are considered crucial drivers of inequality trends in countries over recent decades. OECD (2012a) regresses an equation of type (11.27) with country-fixed effects but with no lag on the wage inequality (dependent variable y_{it}). The trade variable T is measured by trade exposure (but trade openness was also tested with similar results). The set of control variables, $\mathbf{X}$, includes expenditure on the business sector, research and development (R&D) as a share of GDP to account for the impact of technological progress on wage dispersion, as well as product and labor market institutional variables, the sectoral composition of employment, the share of female employment, the percentage of the population that has attained postsecondary education, and the output gap. All these variables can have an impact on wage dispersion.

With or without these controls, table 11.7 shows that there is no clear evidence of a correlation between trade and wage dispersion. Wage dispersion is positively associated

TABLE 11.7
The effect of trade policies on wage inequality in panel of 22 OECD countries.

Dependent var.: *ln* D9/D1 ratio	(1)	(2)	(3)	(4)	(5)	(6)
ln Overall trade exposure	.049	.035				
ln Exports/GDP			.038			
ln Imports/Domestic demand				−.052		
ln Imports from low and med.-income countries/GDP					−.017	−.037***
ln Imports from low and med.-income countries/GDP × dummy for low EPL						.073***
dummy for low EPL						.001
ln Union coverage rate		−.039*	−.040*	−.033*	−.039*	−.004
EPL		−.052***	−.052***	−.058***	−.053***	−.066***
ln Tax wedge		−.112***	−.110***	−.106***	−.102***	−.110***
ln Product market regulation		−.040**	−.039**	−.041**	−.036**	−.048***
ln Technological change		.097**	.098**	.103**	.093**	.090*
ln % Postsecondary education	−.119**	−.116***	−.120***	−.102***	−.115***	−.089***
Observations	333	333	333	333	333	333
R^2	.45	.55	.55	.55	.55	.57

Note: Controls include country and year fixed effects, output gap, and sectoral share of unemployment. Technological change is proxied by the ratio of business R&D spending to GDP. EPL stands for the OECD employment protection index (*,**,***: significant at the 10%, 5%, and 1% level respectively).

Source: OECD (2012, tables 2.1 and 2.2).

with technological progress, as well as the share of the population that has attained secondary education. Adding product and labor market regulation does not alter these results (column 2). When introduced into the regression, labor union coverage, employment protection legislation (EPL), the tax wedge on labor, as well as product market regulations are all significant with a negative sign on wage dispersion. Disaggregating the overall trade exposure variable into subcomponents such as exports and imports does not change these results (columns 3 and 4; see also IMF, 2007, for similar conclusions). Further disaggregating the trade indicator by region of origin and destination, with a focus on imports from low/medium-income developing/emerging countries (like China and India), which could decrease the wages of the low-skilled and increase the wages of the high-skilled in advanced economies (according to the prediction of the Stolper-Samuelson theorem), indicates no apparent relation between wage dispersion and imports from emerging economies (column 5).

The coefficient of trade can vary depending on the labor market's degree of flexibility. For instance, with respect to employment protection legislation (EPL), βT_{it} in equation (11.27) can be replaced by $\beta_1 T_{it} + \beta_2 T_{it} D_i + \beta_3 D_i$, where $D_i = 1$ for countries with a low degree of employment protection, and $D = 0$ otherwise. In that case, β_1 measures the correlation between trade and wage dispersion in countries with a high degree of EPL. We see that importation from low/medium-income economies is associated with a larger wage dispersion in countries where employment protection is less strict (column 6).

While the effect of trade integration has been estimated to be insignificant for wage dispersion at the aggregate level, it is possible that there are effects at the more disaggregated level. Similarly, trade can have a heterogeneous impact on employment and job separations across firms and industries. The next section, which is devoted to the microeconomic evidence, aims to shed some light on these issues.

2.2 Empirical Evidence at the Micro Level

The macroeconomic studies reviewed previously suffer from a number of drawbacks, including the lack of reliable data for developing/emerging economies and the difficulty of identifying the impact of trade separately from the impact of other factors that can influence or can be influenced by trade. Some empirical studies have relied on data at the firm or the individual level, comparing outcomes such as wages or job separation across different types of firms (exporters/nonexporters) or following identified trade opening episodes.

Yet identifying the impact of trade at the firm level is a further challenge because many competing factors can influence wages, employment, and job turnover. Moreover, firms that export might have unobserved characteristics or might hire workers with special abilities that also influence wages and turnover. This selection effect must be considered carefully when comparing exporters and nonexporters. Doing so usually calls for detailed, firm-level data, possibly matched with individual data on workers. Even with this type of data, as in the case of macroevidence-based studies, causal interpretations should be considered carefully and results should be viewed as mostly descriptive, unless clearly exogenous factors (unrelated to wage and job turnover) can explain why firms export or import.

One mechanism adduced to explain why exporting firms pay higher wages is that search cost in the labor market implies that wages are determined by rent-sharing between firms and workers, which leads exporting firms to pay a higher wage to otherwise identical workers (see chapter 3). Another explanatory phenomenon adduced is that trade induces a "quality" upgrading of the whole chain of production, including the skills of workers in exporting firms, even among firms within the same sector.

2.2.1 Wage Inequality: Technological Change vs. Trade

In a seminal paper, Bernard and Jensen (1997) analyzed the increased demand for skilled labor and rising wage inequality in the 1980s in the U.S. manufacturing sector, using an exhaustive microeconomic data set on individual establishments over the period 1973–1987 at the plant level. More precisely, they examined whether the employment share for nonproduction workers (the ratio of nonproduction workers to total employment) and the wage share for nonproduction workers (the ratio of the wage bill for nonproduction workers to the total wage bill) were increasing as a result of within-plant changes or shifts in employment and wages across plants. For the decomposition analysis within a given sector, the basic formulation is:

$$\Delta P = \sum_i \Delta S_i \bar{P}_i + \sum_i \Delta P_i \bar{S}_i$$

where P is the ratio of interest in the sector at the aggregate level (for instance, if N denotes the employment level for nonproduction workers in the sector and E denotes

total employment in the sector, $P = N/E$). Δ is the difference operator over the period under consideration. S_i is the employment or wage share of plant i in the sector (for instance, if E_i denotes the employment level in firm i, $S_i = E_i/E$). P_i is the ratio of interest in plant i (for instance, $P_i = N_i/E_i$). A bar over a variable stands for time average. The between effect is given by the first term on the right side. It shows the change in the aggregate ratio due to reallocations of workers between plants. The within effect is given by the second term on the right side.

Bernard and Jensen found that while there is evidence that plants were increasing their within-plant share of nonproduction workers, the data suggest that between-plant movements contributed to the rise in relative wages for nonproduction workers. In other words, wage share increases mostly occurred because of shifts across plants. Additionally, Bernard and Jensen found that the increase in the wage gap between high-skilled (or nonproduction) and low-skilled (or production) workers can be attributed substantially to changes at exporting establishments. This suggests that trade could be the cause of increasing wage inequality rather than technological change that would rather have impacted wage inequality within all types of plants. But exporting plants also tend to have larger R&D outlays. To test the competing roles of technological change and trade on wage inequality with greater precision, the authors regress changes in the plant's outcomes on a set of determinants including exports and technological change:

$$\Delta y_i = \alpha \Delta Z_i + \beta \Delta T_i + \mathbf{X}_i \boldsymbol{\gamma} + \varepsilon_i$$

In this equation, Δy_i is plant i's annual percentage contribution to the within or between change in either the employment share for nonproduction workers or the wage share for nonproduction workers observed at the aggregate level over the period 1980–1987. ΔZ_i is the change in technology proxied by changes in the ratio of R&D to sales and computer investment, ΔT_i is the change in total value of shipments abroad and domestically, $\mathbf{X}_i$ includes a set of characteristics of plants (age and size, capital–labor ratio, industry, and export status). The authors cannot control for imports with their data. Their results, summarized in table 11.8, suggest that the between-plant movements of workers and wages, which are especially important in the increases in the aggregate wage gap, are largely determined by export-related demand movements across plants. Technology plays an ambiguous role. Increases in the R&D/sales ratio at the plant level are related positively to between-plant increases in the share of white-collar workers and in the wage share for white collars. But changes in capital intensity, computer investment per employee, and the capital–labor ratio are significantly negatively related to the between-plant movements in the share of white-collar workers and in the wage share for white collars.

Now if trade is the most relevant characteristic, ahead of technological progress, for explaining the observed shift in the share of white-collar workers and in their wage share, what makes firms export?

2.2.2 Are Exporting Firms Different?

The model of Melitz (2003) suggests that due to entry costs, only highly productive firms self-select into the group of exporters. There is some evidence to support this view.

TABLE 11.8
Determinants of wage and employment changes in the United States, 1980–1987.

	Dependent variable			
	Δ Employment share		Δ Wage share	
	Between	Within	Between	Within
Δ R&D/sales	1.00 (2.46)	0.75 (7.38)	1.56 (2.37)	1.73 (6.44)
Δ Computer investment	−5.65 (2.82)	3.79 (3.57)	−4.76 (1.47)	3.38 (2.57)
Δ Domestic shipments	5.66 (31.53)	−0.70 (7.38)	9.27 (31.90)	−0.80 (6.65)
Δ Foreign shipments	15.94 (18.51)	2.83 (6.22)	25.16 (18.07)	1.70 (3.02)
R^2	0.38	0.17	0.38	0.19

Note: T-statistics are in parentheses. All specifications include age dummies, change in capital-to-labor ratios, change in specialization, and employment in 1980, an export dummy, and five-digit industry dummies. There are 8,981 observations in each specification.

Source: Bernard and Jensen (1997, table 8).

Using Firm or Individual Panel Data

Bernard and Jensen (2004) examined the factors that increase the probability of exporting in the U.S. manufacturing sector, using panel data. Following plants over time allows them to control for unobserved plant characteristics and to identify entries and exits on the export market. To accomplish this, they estimate an equation with plant-fixed effects[4] very close to equation (11.27) where y_{it} is a dummy that has value 1 if the plant exports and 0 otherwise:

$$y_{it} = \rho y_{i,t-1} + \mathbf{Z}_{i,t-1}\boldsymbol{\beta} + \mathbf{X}_{i,t-1}\boldsymbol{\gamma} + \mu_i + \varepsilon_{i,t}$$

In this equation, the probability of exporting in period t, y_{it}, depends on the probability of having exported in the past period, $y_{i,t-1}$, and other factors $\mathbf{Z}_{i,t-1}$ that could have influenced the decision, such as terms-of-trade shocks in $t-1$ (based on exchange rates, which may vary depending on the export destination), industry demand shocks, state industry spillovers (export activity in the plant's state and/or industry), and government subsidies (which may vary at the plant level); $\mathbf{X}_{i,t-1}$ includes lagged time-varying plant characteristics, such as size, productivity, labor quality, type of ownership (to identify multinationals), and a dummy to identify any change of product. The plant-fixed effect is denoted by μ_i. Bernard and Jensen (2004) estimate this equation with data based on 13,550 plants and 95,902 observations. Given the presence of the endogenous lagged variable, this equation is regressed both as such in levels or in first difference using the Arellano-Bond GMM method presented in the previous section and in the appendix to this chapter. Bernard and Jensen find that productivity, wages, and the share of

[4]Individual heterogeneity can be modeled by either fixed or random effects. Random effects require that the individual effect be uncorrelated with the other regressors. This is probably not the case in this model, since the plant's observed characteristics included in the regression, such as plant size, wage levels, and ownership status, are probably correlated with the product attributes, the technology, and the managerial ability contained in the unobserved plant effect.

nonproduction workers are positively related to the probability of exporting. Moreover, exporting the previous year raises the probability of exporting today between 39% and 66%, but the effect diminishes over time when an additional lag is introduced, suggesting that sunk costs act as a slowly depreciating investment. Subsidies seem to play no role. These results tend to confirm the importance of entry costs in exporting and the importance of plant characteristics which explain past success, either observed (size, new product) or not.

Based on a virtually exhaustive panel of French firms, Biscourp and Kramarz (2007) use an approach similar to that of Bernard and Jensen (1997), which provides further evidence not only on exporting firms but also on importing firms. Biscourp and Kramarz (2007) investigate the relationship between job flows (job creation and destruction, notably for production workers) and changes in trade variables, namely imports and exports of firms across the period 1986 to 1992. They find that exports are positively associated with job creation, but also find a strong correlation between increasing imports, in particular imports of finished goods, and job destruction, most notably destruction of production jobs. The effect on job destruction is stronger for large firms. Finished goods imports may reflect outsourcing strategies. Egger et al. (2007) find similar results for Austrian workers between 1988 and 2001 based on a detailed and high-frequency panel from individual social security records, which allows them to control for observed and unobservable characteristics of persons. They find that trade affects labor turnover in various industries, notably those featuring net imports. Increased outsourcing reduces the probability of remaining or being hired in manufacturing sectors with comparative disadvantage, but it does not alter the transitions to those with a comparative advantage.

Using Matched Employer–Employee Data

One of the problems of research that relies only on data concerning firms is that it fails to control for the quality of human capital. Exporting firms might also be those that manage to employ the most productive workers, independently of trade. In that case the correlation between trade and wages or productivity would be spurious. Hence, following Bernard and Jensen (1997), additional research based on plant or firm data has analyzed the differences in wages and employment between exporters and nonexporters using matched employer–employee data sets. A typical regression of these studies is:

$$\ln w_{ijt} = \mathbf{Z}_{jt}\boldsymbol{\beta} + \mathbf{X}_{it}\boldsymbol{\gamma} + \mu_{ij} + \mu_t + \varepsilon_{ijt} \tag{11.29}$$

where w_{ijt} is the wage of individual i working in firm j at date t. $\mathbf{Z}_{jt}$ includes firm j's characteristics, such as its ratio of exports to sales and/or a dummy for exporting firms, as well as the size, the capital, and the workforce composition (skill composition) of the firm; $\mathbf{X}_{it}$ includes some of worker i's characteristics (labor market experience, job tenure, age, family status, education level). The fixed effect for the match of person i to plant j (since workers can move across plants) is denoted by μ_{ij}, while μ_t is a year-fixed effect. This specification is often called the job spell specification (and is sometimes tested against the alternative specification of plant- and person-fixed effects considered separately—μ_i and μ_j instead of μ_{ij}—to check the importance of controlling for a job spell). It can be run for all job spells or separately for different types or workers (e.g., by occupation, education level).

For instance, Schank et al. (2007) use a large longitudinal set of employer–employee data for Germany between 1995 and 1997 and show that the wage premium in exporting firms does not vanish even when observed and unobserved heterogeneity of workers and workplaces are controlled for with either plant-fixed effects, persons-fixed effects, or alternatively "spell"-fixed effects. They find that blue-collar workers in plants with an export sales ratio of 60% earn 2% more than similar employees in otherwise similar plants that do not export. This persists even when controlling for the working time of employees.

Similarly, Munch and Skaksen (2008) base their analysis of the wage premium of exporters on Danish data for the period 1995 to 2002 with a detailed record of firms' international trade, notably trade with the European Union and Norway, and socioeconomic information on workers, particularly their level of education. Indeed, even if the most productive firms self-select into being exporters, the size of the wage premium could be influenced by the skill intensity of their employees. Munch and Skaksen introduce an interaction term between the exports to sales variable and the skill intensity in equation (11.29). They find that the sign of this interaction on wages is positive and significant, suggesting that the export premium is larger in skill-intensive firms. Interestingly, running the regression for unskilled workers alone, they find that the sign of the export intensity variable is negative, while the interaction variable with the skill intensity is positive, suggesting that for these workers, trade can lead to lower wages, unless they work in firms with high skill intensity.

Using Natural Experiments

The evidence described above should be considered merely as descriptive because even after controlling for external economic factors, as well as for firm and worker individual effects, there is no certainty that the measured effect is causal. A better strategy to identify a causal impact would be to measure the effect of trade following some liberalization "shocks" affecting firms differently. A few papers follow this strategy.

Verhoogen (2008) studies the development of wage inequalities in Mexico, where the D9/D1 decile gap increased over the 1990s. This change in wages coincided with an increase in the share of exports of manufactured goods, following the peso devaluation of December 1994. The Krugman-Melitz class of models would suggest that the change stems from the fact that only the most productive high-paying firms should be able to seize the opportunity presented and expand to export markets, which can lead to higher wage inequality, notably if they improve the quality of their products to do so. But the Stolper-Samuelson theorem would predict on the contrary that wage inequality should fall in a country such as Mexico, intensive in low-skilled labor, when trade expands. Controlling for the productivity of firms, Verhoogen compares wages and other outcomes in initially less and initially more productive firms during the crisis or "treatment" period (1993–1997) to outcomes in the same firms during "control" periods before (1989–1993) and after (1997–2001) the crisis period. Productivity is proxied by log domestic sales deviated from industry means as observed in the initial year of each period (but using alternative proxies such as total factor productivity, domestic sales per worker, or employment would give the same results qualitatively). For instance, considering only two groups of plants, with either high or low productivity, his strategy amounts to comparing the difference in the change in wages between the two groups

during the crisis period (i.e., over the period spanning the crisis, 1993–1997) and after the crisis (the 1997–2001 period). This amounts to achieving a *triple difference* (across plants, before–after the crisis, and before–after a period without crisis). Doing a triple difference has the advantage of controlling for potential unobserved differences between the two groups of plants which would be unrelated to the crisis. The equation is:

$$\Delta y_{ijr} = \alpha + \beta z_{ijr} + \gamma_j + \gamma_r + \varepsilon_{ijr}$$

where Δy_{ijr} is the change in the outcome over the considered period (1993–1997 or 1997–2001) in firm i in industry j and in Mexican state r, z_{ijr} is the proxied productivity level in the initial year of the period, γ_j is an industry-fixed effect, and γ_r is a state-fixed effect. Table 11.9 displays Verhoogen's main results. Based on a panel of about 1,000 plants, he finds greater differential changes in the export share of sales for higher-productivity firms, as well as higher white-collar wage growth, blue-collar wage growth, and higher relative wage of white-collar workers in the peso crisis period than in the placebo period. He also finds that the capital-to-labor ratio increased more in higher-productivity firms. But the results show no significant impact on the white-collar employment share. Overall this paper tends to confirm the predictions of the Krugman-Melitz class of models, as well as the potential role of quality upgrading in assessing the impact of trade on emerging economies.

The selection effect of trade can also be identified through the analysis of firms' exit probability. Indeed, according to the Melitz model, trade not only favors high-productivity firms, but also hinders lower-productivity firms, forcing some of them to exit their market. For instance, Eslava et al. (2013) study the impact of the major trade liberalization that occurred in Colombia between 1990 and 1992. Over that period, effective tariffs dropped on average from 62.5% to 26.6% as compared to 86% in 1984.[5] At the same time, the dispersion of tariffs across sectors fell substantially. The movement

TABLE 11.9
The effect of the peso devaluation in Mexico on trade and wage inequality.

Dependent var.	Δ (export share)	Δ ln (white-collar wage)	Δ ln (blue-collar wage)	Δ ln (wage ratio)	Δ (K/L ratio)
OLS regressions					
1993–1997	.020***	.072***	.036***	.036***	.083***
1997–2001	.007***	.016**	.008	.008	.026***
Diff. (1993–1997 vs. 1997–2001)	.014***	.056***	.028***	.028**	.057***

Note: All regressions include 205 industry dummies and 32 state dummies. Number of observations is 3,263 for all regressions. **, *** significant at the 5% and 1% level respectively.

Source: Verhoogen (2008, table II).

[5] Effective tariffs in one sector under a given protection system (made up of nominal tariffs) may be defined as the difference between the industry's value added under this system and under free market conditions expressed as a percentage of free market value added.

of trade liberalization was stopped (but not reversed) in 1992 after the change of government. The authors use this variation to identify the impact of trade on firm exits. They use a detailed panel of firms, which includes the prices of sales and that of intermediate goods bought, which allows them to carefully control for real productivity differences across firms. Comparing firms from sectors treated differently by the trade liberalization before 1990 and after 1992, they find that greater international competition magnifies the contribution of productivity to the probability of exiting the market: trade reform helps to weed out the least productive plants. Trade also contributed to the increase in average industry productivity, by 8.2 log points. Both the selection effect and the improvement in productivity among surviving firms explain this rise.

Overall, the evidence reviewed in this section tends to support the predictions of the Krugman-Melitz model over the Stolper-Samuelson theorem. At the aggregate level, there is no evidence that trade increases unemployment or contributes significantly to the rise in wage inequality, even among the most advanced economies. On the contrary, the development of trade appears to be negatively correlated with unemployment both across and within countries. At the firm level, there is converging evidence that the selection effect set out by the Krugman-Melitz model is at play: exporting firms tend to be larger, create more jobs, and pay higher wages than nonexporting firms. However, trade also seems to be associated with a change in the distribution of wages in favor of nonproduction workers. The study of exceptional trade liberalization events suggests that these features are indeed the result of the selection effect of trade and are not mere correlations.

3 MIGRATIONS

The immigration of low-skilled workers is sometimes denounced as a factor in both the decline of wages and the rise of unemployment for this category of worker. The putative consequence is a diminution in the well-being of native workers with few skills and an increase in inequality. Scrutiny of migratory flows does reveal that the rich countries do have immigrant populations less skilled, on average, than natives. We shall see nevertheless that the immigration of low-skilled workers has, in theory, an ambiguous impact on inequality. Empirical research confirms this outlook, suggesting that the immigration of low-skilled workers has little effect on earnings and employment among the least skilled native workers.

3.1 THE CHARACTERISTICS OF MIGRATIONS

As figure 11.9 shows, the foreign-born represent widely varying percentages of the populations of the different OECD countries. Among the 24 countries present in this figure in 2010, Australia is at the top with 27%, while Japan, Korea, and Mexico are at the bottom with a foreign-born part of their population of about 1%. The United States and Germany occupy a middle position, with about 13% of their population foreign-born. These differences reflect different degrees of country attractiveness, as well as differences in immigration policy, which itself varies over time in each country. For instance, Australia, Switzerland, New Zealand, and Canada, which have the highest shares of

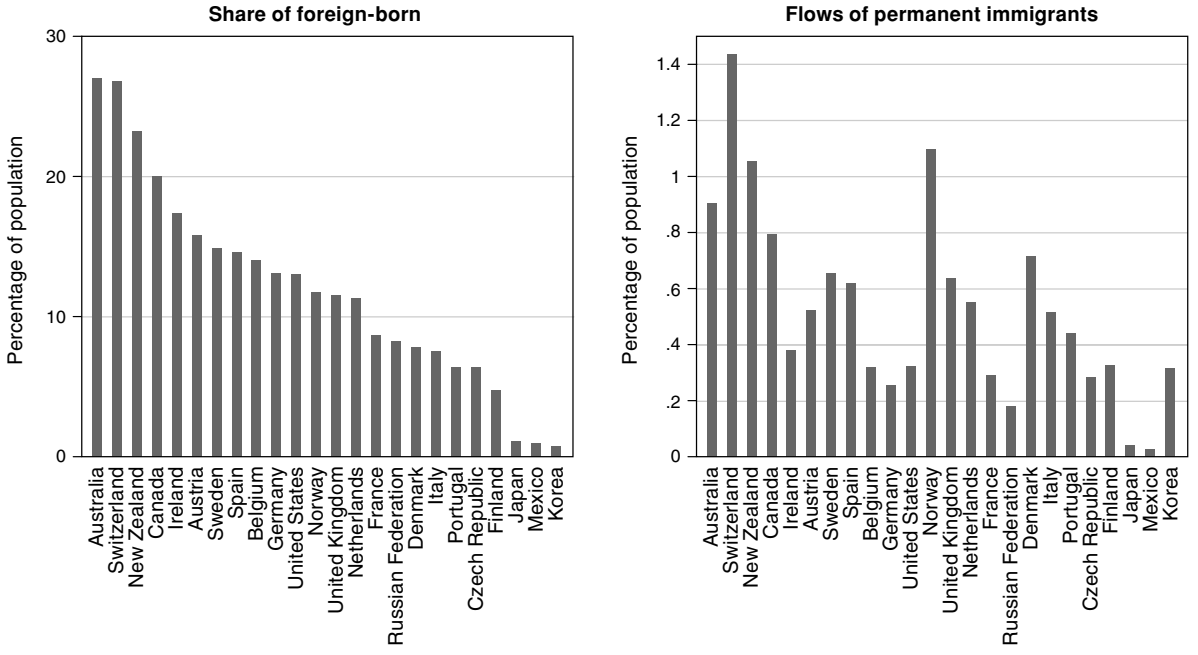


Figure 11.9

Number of foreign-born in 2010 (left panel) and annual flows of permanent immigrants in 2010 (right panel), as a percentage of the population of the host country.

Note: Permanent immigration reflects movements that the receiving country considers are for the long term (i.e., the persons considered are on a "migration track" that normally leads to permanent residence in the host country).

Source: OECD International Migration Database.

immigrants in their populations, also feature some of the highest annual rates of permanent migrant inflows. The characteristics of migration have evolved markedly over the last several decades in the OECD countries. Historically the United States is an important destination and receives the largest number of immigrants of all the OECD countries. It took in 1,040,000 persons in 2010, but the rate of immigration there at present is two or three times lower than it was in the middle of the nineteenth century and the early part of the twentieth century. Thus, in 2010 there were three arrivals for every thousand inhabitants (see the right panel of figure 11.9).

Migration flows are also influenced by the economic cycle. During the 2008–2009 economic crisis, the decline in immigration was significant in Ireland, which was hard hit by the crisis: permanent inflows declined by 80% between 2007 and 2010. But the decline was also significant in Southern Europe (Portugal, Italy, and Spain) as well as Japan. Migration flows remained strong in countries with an immigration tradition and where growth was stronger such as Canada and Germany, or even Norway, where immigration reached a new record high in 2010.

On the other hand, many European countries have gone from being sending countries to being receiving countries. This emerges from figure 11.10, which shows that the net inward flow of migrants (immigrants minus emigrants) became largely positive

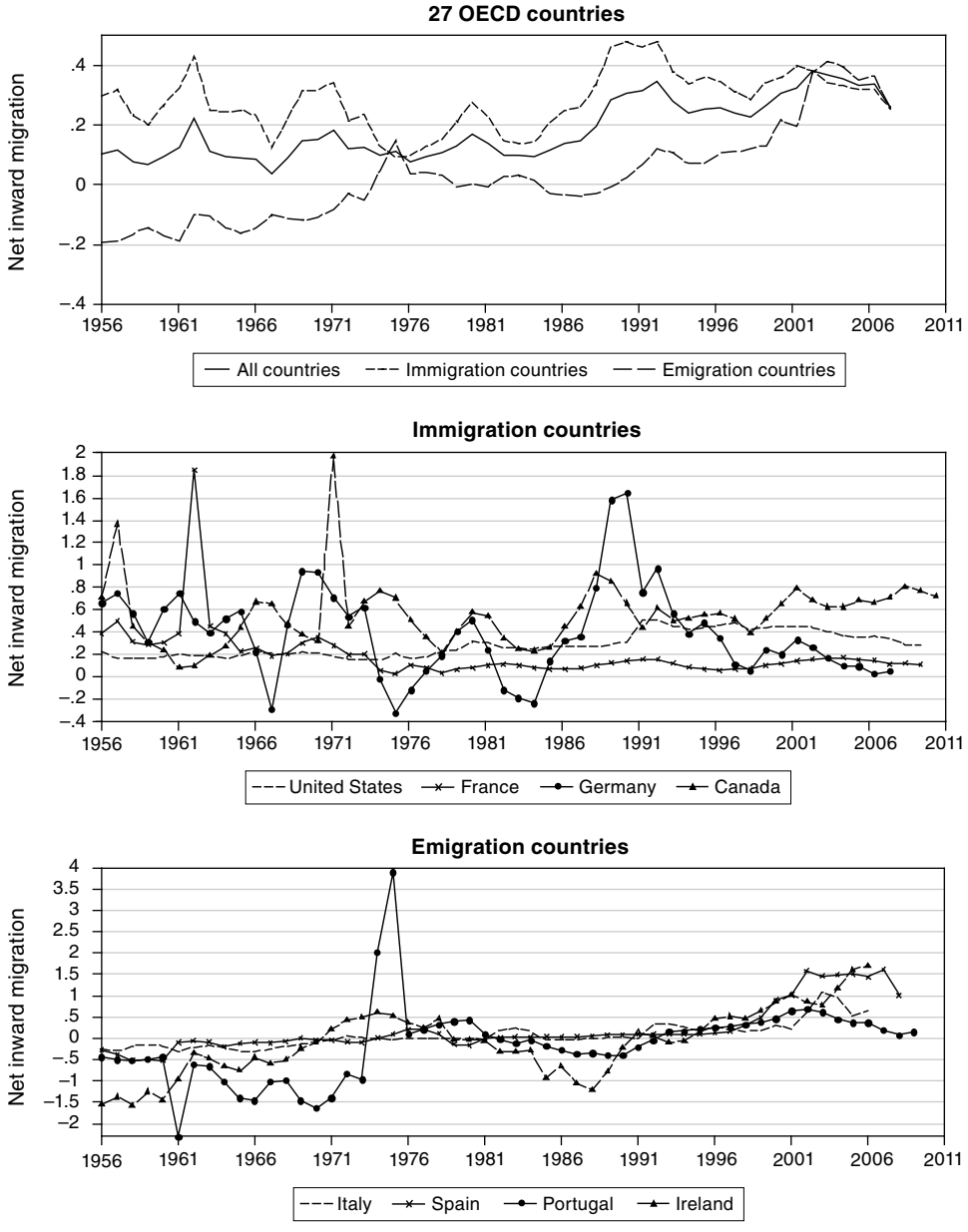


FIGURE 11.10
Net inward migration flows in 27 OECD countries since 1956 (percentage of population).

Note: Net inward migration is the ratio of immigration minus emigration flows over the period to the total population. Note: Emigration countries are those where net inward migration flows tended to be negative until the 1970s: the Czech Republic, Denmark, Finland, Greece, Hungary, Iceland, Ireland, Italy, Japan, Norway, Poland, Portugal, the Slovak Republic, and Spain. Immigration countries are Australia, Austria, Belgium, Canada, France, Germany, Luxembourg, the Netherlands, New Zealand, Sweden, Switzerland, the United Kingdom, and the United States.

Source: OECD International Migration Database.

in the OECD countries, even among European countries with an emigration tradition such as Ireland, Italy, Portugal, and Spain after the 1980s, and notably from the 1990s following the collapse of the Soviet bloc. This event generated a particularly large flow of immigrants into Germany—a typical "shock" of the kind that can help identify the effects of immigration on the labor market (see section 3.3.2).

Global orders of magnitude aside, it is important to emphasize that the migrants arriving in the rich countries of the OECD have socioeconomic characteristics that generally differ from those of natives. The migrants are younger, the proportion of men is larger, they are concentrated in the major cities, their educational level is lower, they hold fewer skilled jobs for comparable levels of education and experience, and they are more frequently unemployed. These average differences may conceal differences among nationalities, inasmuch as socioeconomic characteristics are strongly influenced by the country of origin.

However, most immigration stems from countries in the same region as the host country, notably due to free movement agreements in Europe and facilitated procedures (based on NAFTA) in North America, but this is also true in Japan, where most immigrants come from Asia (see panel B of figure 11.11). These differences also reflect the selection criteria of immigration authorities, which vary significantly across countries. For instance, panel A of figure 11.11 shows that work-related immigration (work visa for workers and their families) represents the largest share of permanent immigrants in Australia, Canada, and the United Kingdom, whereas in France, Germany, and Sweden the weight of free movement within the European area is at the origin of a larger fraction of intakes. In the United States, family immigration (including families accompanying workers) remains the main type of immigration. Consistently, the share of foreign-born with tertiary education is largest in Canada, the United Kingdom, or Australia (around 40%), and close to or above the share of native-borns having university diplomas, because their systems of selection of migrants (notably, points-based systems) favor highly educated candidates seeking work (see figure 11.12). This is in contrast with continental and Southern European countries where the share of highly educated immigrants is lower than that of native-borns.

What is more, differences between the performance of migrants and that of natives appear to dwindle, the longer immigrants are present in the receiving country. Chiswick (1978) initially identified this phenomenon in the United States from U.S. census data for 1970. He shows that immigrants arriving in the United States earn, on average, an income 17% lower than that of natives with comparable characteristics (educational level, experience, sex, region). This difference dwindles at around 1% per year. The earnings of migrants who arrived more than 15 years ago even overtake those of natives. This phenomenon, which also seems to be discernible in other OECD countries, has drawn much attention. It might result from the progressive integration of immigrants into the receiving economy, which would explain the shrinkage of the gap in relative earnings between migrants and natives, but not the fact that the migrants end up with higher earnings than natives. Selection biases might be at the origin of this finding: migrants whose unobservable characteristics (appetite for work, efficiency…) are above average should end up with higher average earnings once the integration phase is over. Finally, it is not out of the question that the cross-section estimate of Chiswick (1978) is sensitive to a cohort effect if the average quality of migrants falls

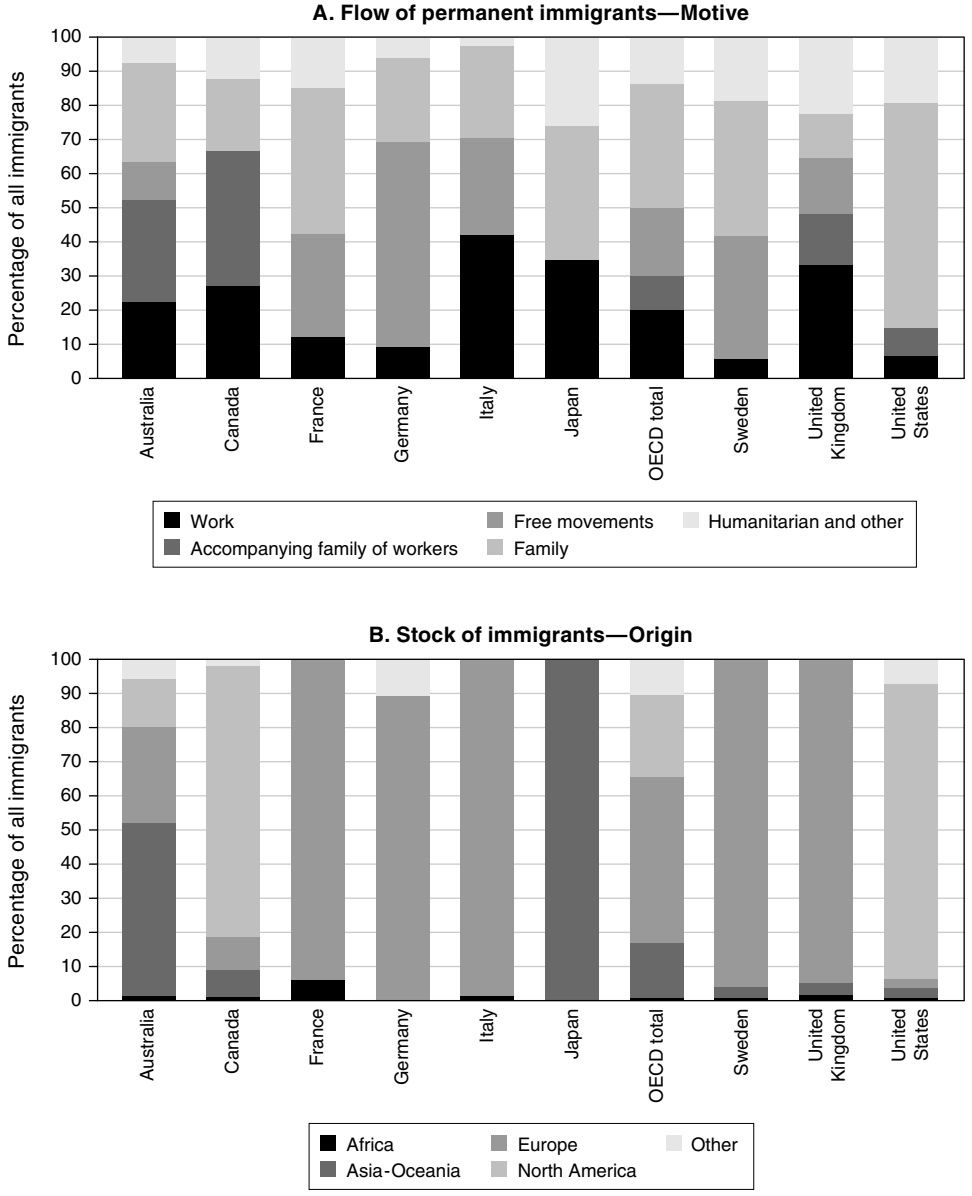


FIGURE 11.11

Motivations and origins of migrants in selected countries. Graph A = Category of entry of permanent immigrants (flow), in 2010. Graph B = Geographical origin of immigrants (stock), by host country, around 2000.

Note: The aggregate estimates of permanent immigration are decomposed among the following categories: work-related migration, accompanying family of worker migrants, family reunification and formation (e.g., marriage), humanitarian migration (including protection and accompanying family of humanitarian migrants).

Source: For the motives, OECD International Migration Outlook (2012, figure I.4), and for the origins, Database on Immigrants in OECD Countries (DIOC).

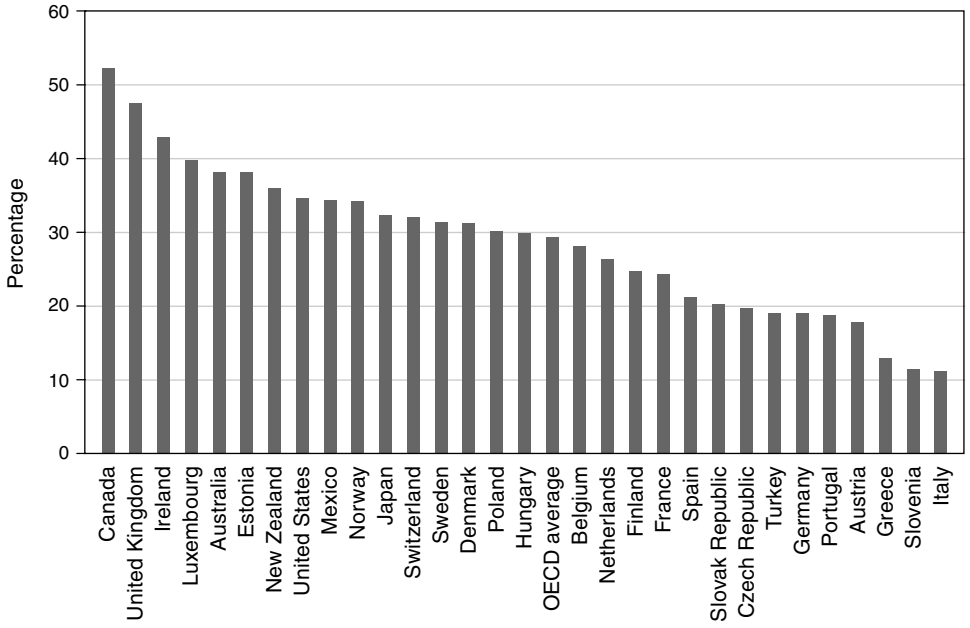


Figure 11.12
Percentage of highly educated in the foreign-born population, 2009–2010.

Source: OECD International Migration Outlook (2012, figure I.15).

off over time (only individuals present in the sample at one point in time are considered). If this is the case, the observation of an improvement in the relative earnings of immigrants with time passed in the United States may simply result from the fact that the migrants who have been there longest belong to cohorts the average quality of which was higher. More recent work by Abramitzky et al. (2012) suggests that this phenomenon does indeed play a role. In newly assembled panel data following immigrants over time in the United States, Abramitzky et al. find that the initial immigrant earnings penalty disappears almost entirely, and immigrants experience occupational upgrading at the same rate as natives. The cross-sectional patterns are driven by declines over time in arrival cohort quality and the departure of negatively selected return migrants. However, these findings vary substantially across sending countries.

Immigration also concerns a growing number of students in tertiary education. As shown in figure 11.13, in 2010 a large share of these students came from Asia, and notably China, India, and Korea, which cumulated 30% of all international students present in the OECD area in 2010. They only represented a modest share of their native countries' youth population, except in Korea, where they rise to 4% of youth aged 20–24, as well as in Greece and the Slovak Republic in the context of the Great Recession. On the other side, the international market in tertiary education is largely dominated by English-speaking countries, and particularly by the United States, which is the destination of a quarter of all international students in the OECD, followed by the United Kingdom and Australia. Altogether these three countries welcome 50% of all international students. In Australia foreign students represent more that 20% of all tertiary students.

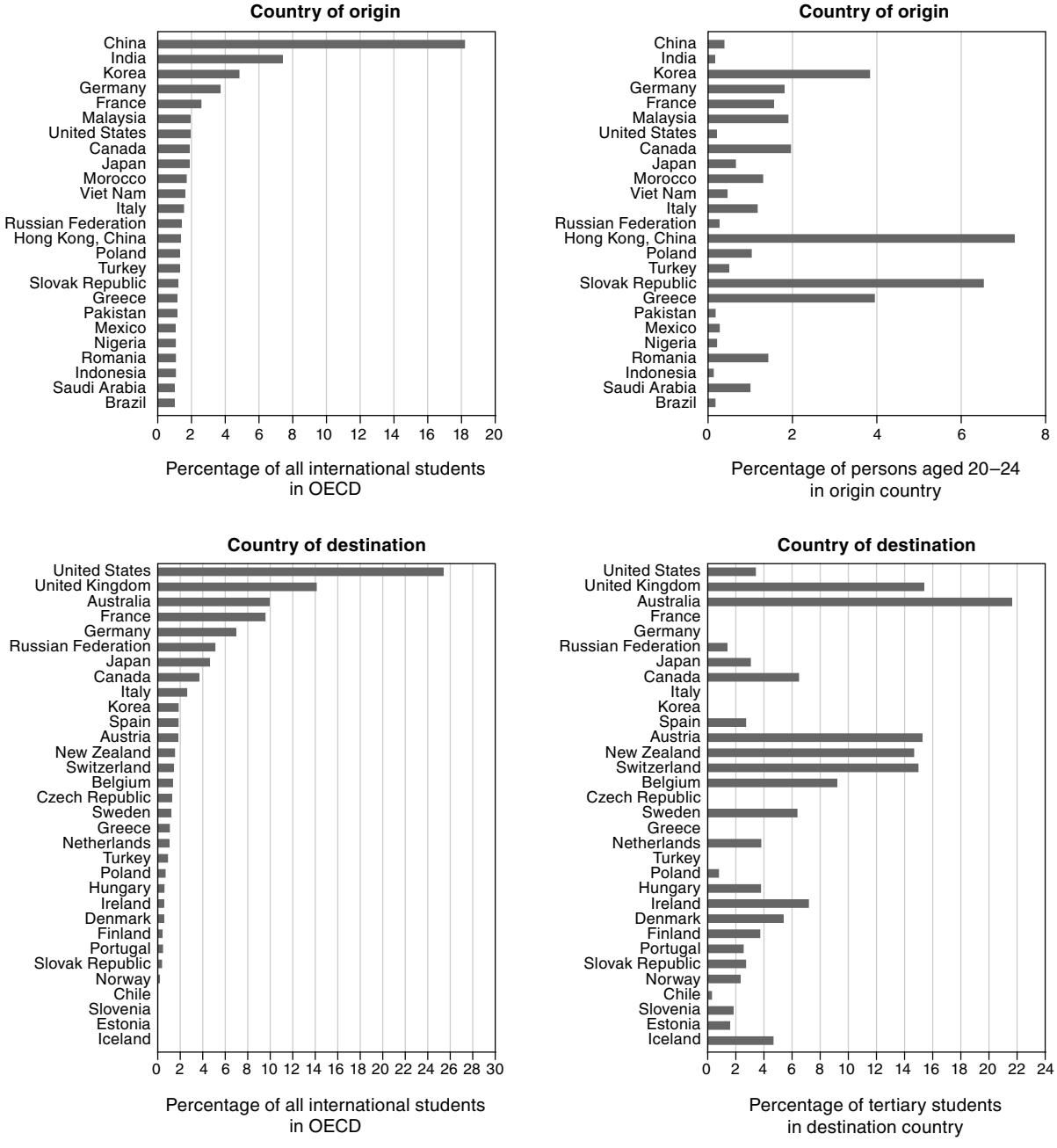


FIGURE 11.13

Main origins and destinations of international students in the OECD countries, 2009.

Source: OECD International Migration Outlook (2012, tables I.3 and I.4).

Granting student visas to young foreigners is an alternative strategy for these countries to attract qualified migrants who will probably look for a job locally after earning their degree.

This rapid review of the facts suggests that immigration may potentially increase inequality in the rich countries of the OECD, since these take in workers whose performance in the labor market is on average less good than that of natives, at least in some countries. The same line of reasoning could lead us to think that net migration might induce an increase in unemployment, at least in the short run. However, immigrant-receiving countries such as the United States, Canada, and Australia have also featured stronger growth and lower unemployment rates on average over the long run than emigrant-sending countries. Indeed, at first glance there is no significant correlation between net inward migration flows over the period 1980 to 2010 and change in unemployment, as shown in figure 11.14. The same figure shows a weak negative cross-section correlation between net inward migration and inequality over the period 1985–2007, which is only driven by two extreme countries (Ireland and Spain, where net migration became positive). Again, as in the case of trade, these correlations are

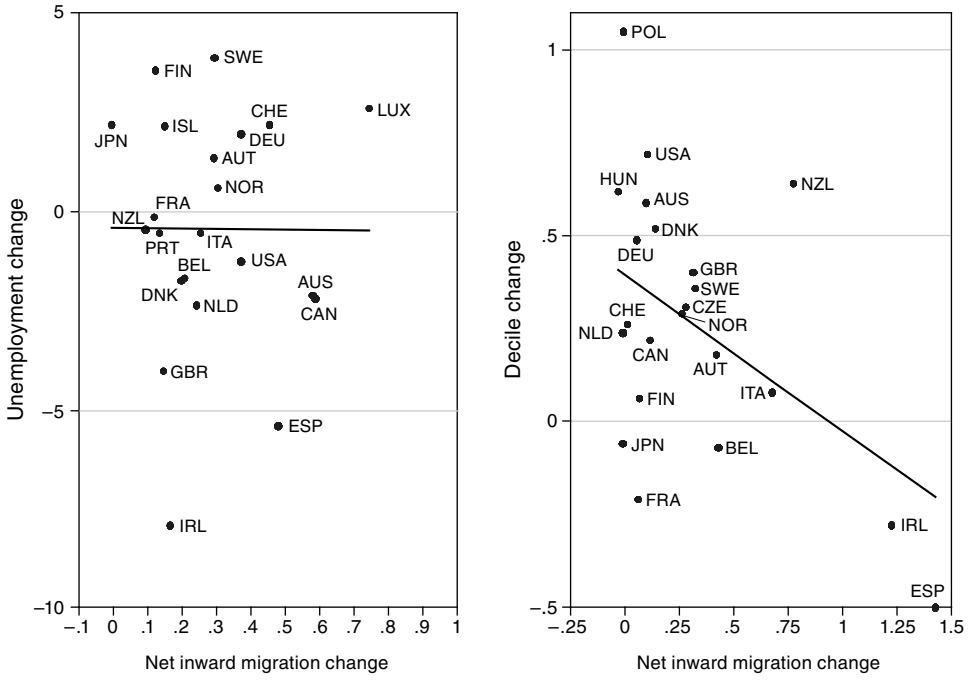


Figure 11.14

Change in unemployment or inequality, and change in net inward migration flows in the OECD countries over the period 1980–2010. Note: For net inward migration flows: average annual percentage change over the period 1980–2010 (left panel) or the period 1985–2007 (right panel); for the unemployment: difference between the averages of unemployment over the periods 2000–2010 and 1980–1990; for the decile change: change in wage dispersion D9/D1 between 1985 and 2007 (23 countries).

Source: OECD International Migration Database, OECD (2012) and IMF World Economic Outlook Database.

difficult to interpret in isolation. For instance, migration flows could be too small to significantly influence the aggregate rate of unemployment or the wage distribution in the long run, but that does not mean that exceptional migration events would not cause a temporary rise in unemployment or a decrease in low wages. Moreover, migrants could be attracted by low-unemployment countries or more unequal countries, if they think these countries have the potential to offer better employment opportunities, which would then lead to a reverse relationship between immigration and labor market performances. For this reason, we now examine what theory has to tell us on this point before turning to evaluation studies relying on a clear identification strategy.

3.2 THEORY

The impact of migrations on the labor market is usually studied using an elementary model of labor demand. The procedure is to analyze the consequences of migration for wages, which are assumed to be determined in perfectly competitive markets. Labor supply is equal to the size of the labor force, including natives and immigrants, and it is the properties of labor demand that play a determining role (see Borjas, 1999).

3.2.1 WHAT THE ELEMENTARY MODEL OF SHORT-RUN LABOR DEMAND TELLS US

Let us begin by considering an economy in which labor is a homogeneous factor. Production is described using a function with constant returns $F(K,L)$, of which the two arguments are the quantity of labor L and the quantity of capital K. Let us assume that the labor market is competitive, and let N be the size of the labor force. The wage w is then given by the marginal productivity of labor at full employment, $w = F_L(K,N)$. In the short run, the stock of capital does not vary, and an increase in the labor force (through a wave of immigration, for example) necessarily leads to a wage reduction due to the decrease in the marginal productivity of labor. This reasoning shows that the immigration of a population whose productive characteristics are identical to those of the residents entails a reduction in all wages in the short run, and an increase in the remuneration of capital, $r = F_K(K,N)$, inasmuch as capital is less quickly adjustable than employment. It is possible to assess the wage reduction if we know the wage elasticity with respect to employment, η_L^w, which is equal to the inverse of the wage elasticity of labor demand, $\eta_w^L = F_L/LF_{LL}$. For a given stock of capital,[6] we can estimate that η_w^L takes the approximate value -3. An immigration corresponding to 1% of the labor force then reduces wages by $(1/3)\% \simeq 0.3\%$. So the short-run effects are potentially slight.

Despite the wage reduction, immigration entails an overall gain for the natives as a whole if they are owners of capital. This we can show by calculating the variations in their wages and the variations in the remuneration of capital due to immigration. Figure 11.15 represents the impact of immigration when the labor force comprises N

[6] Readers will recall that the elasticity of substitution between capital and labor can be written $\sigma = F_K F_L / YF_{KL}$ when the production function is homogeneous of degree 1 (see chapter 2). Moreover, homogeneity of degree 1 of F entails $LF_{LL} = -KF_{KL}$. Since $w = F_L$, the wage elasticity of labor demand is $\eta_w^L = F_L/LF_{LL} = -\sigma/(1-s^L)$ with $s^L = wL/Y = 1 - (KF_K/Y)$. Assuming that $\sigma = 1$ and that $s^L = 0.7$, we get $\eta_w^L = -1/0.3 \simeq -3$. Note that η_w^L stands here for the elasticity of the unconditional demand for a given capital stock, which is different from the elasticity of the conditional demand, denoted $\bar{\eta}_w^L = (1-s^L)\sigma$ (see chapter 2, sections 1.2.2 and 1.3.1).

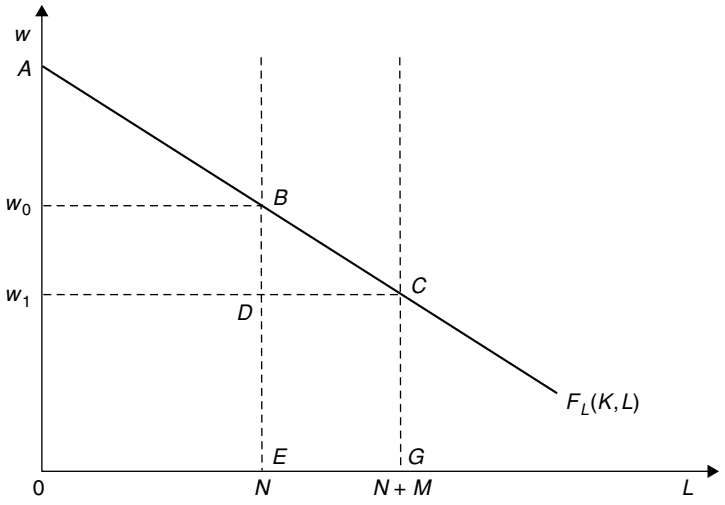


Figure 11.15
The consequences of immigration in a model with homogeneous labor and fixed capital.

natives and M migrants, and the labor market is assumed to be perfectly competitive. Let w_0 be the wage in the absence of immigration; in this hypothesis, we have $w_0 = F_L(K,N)$ and the GDP, equal to $F(K,N)$, is represented by the surface of the quadrilateral $OABE$.[7] With the presence of immigrants, the GDP is higher, since it corresponds to the surface of the quadrilateral $OACG$, of which an amount Mw_1 is obtained by the immigrants in the form of labor remuneration. Immigration thus produces a surplus to the profit of natives equal to the surface of the triangle BCD. This surplus represents the sum of the variations in the earnings of labor and capital. We can approximate it by the term $(M/2)(w_0 - w_1)$. Since $w_1 - w_0 = F_L(K,N+M) - F_L(K,N)$, assuming that M is small with respect to N, a first-order expansion gives $w_1 - w_0 = MF_{LL}(K,N)$, and the surplus S is equal to $-(M^2/2)F_{LL}(K,N)$. In practice, it is more instructive to focus on the relationship between the surplus and production Y. Since the wage elasticity of labor demand is $\eta_w^L = F_L/LF_{LL}$, we get:

$$\frac{S}{Y} = -\frac{1}{2}\left(\frac{M}{N}\right)^2 \frac{NF_{LL}}{F_L}\frac{F_L N}{Y} = -\frac{m^2 s^L}{2\eta_w^L}$$

In this expression, $s^L = wN/Y$ designates the share of labor earnings in the GDP and $m = M/N$ represents the ratio of the number of migrants to the number of natives. This expression of the surplus allows us to make quantitative evaluations, inasmuch as the labor share in the GDP is of the order of $2/3$, and the wage elasticity of labor demand takes the value, in the framework chosen with fixed capital, of around -3. With these orders of magnitude, for $m = 10\%$, we get $S/Y = 0.1\%$. A population of immigrants

[7] Since $F(N) = \int_0^N F_L(\xi)\mathrm{d}\xi$.

representing 10% of the native population thus gives the natives a surplus of around 0.1% of GDP (evaluated before immigration).

This line of reasoning, pursued with the hypothesis of homogeneous labor, clarifies only one part of the impact of immigration on the remuneration of labor and capital earnings. If we distinguish between skilled and unskilled labor, using a function of the type $F(K, L_h, L_\ell)$, it turns out that the immigration of a population less qualified on average than the native population leads to a reduction in wages for the unskilled—since $w_\ell = F_\ell(K, L_h, L_\ell)$ and $F_{\ell\ell} < 0$—and to an increase in the remuneration of capital. The impact on the wages of skilled workers is a priori ambiguous, for skilled labor is complementary to capital, itself substitutable for low-skilled labor (see chapter 2 on this point). Simulations carried out for reasonable values of the elasticities of the factor demands show that the wages of skilled workers are reduced by the immigration of workers with few skills, but in a smaller proportion than is the case with the unskilled. In the short run, the immigration of workers with few skills thus entails an increase in inequality, since it increases the remuneration of capital and reduces wages, with the latter effect being more pronounced for those earning low wages.

3.2.2 What the Elementary Model of Labor Demand Tells Us in the Long Run

Let us come back to the case of homogeneous labor. In the long run, the marginal productivity of capital equals the interest rate, that is, $r = F_K(K, N)$. This condition determines the capital–labor ratio, $k = K/N$, which satisfies $r = F_K(k, 1)$, and entails, with labor demand, that wages are finally independent of the size of the labor force, since $w = F_L(k, 1)$. Variations in the stock of capital, financed by domestic or foreign savings, ensure that in the long run wages and population size are independent of each other. In figure 11.15, the graph of the labor demand function becomes a horizontal line $w = F_L(k, 1)$.

Obviously, if labor is heterogeneous, the composition of the population affects the relative incomes from different types of labor. To illustrate this phenomenon, let us return to the labor demand model used in the previous section, leaving capital aside (the mechanisms are generalizable to the case with capital; see Borjas, 1999). In a closed economy, the wage level of each category of labor is given by relation (11.3), that is, $w_i = A_i F_i(\alpha\nu, 1)$, $i = h, \ell$, with $\alpha = A_h/A_\ell$, and $\nu = N_h/N_\ell$. It turns out that immigration has an impact on the structure of wages if and only if it alters the proportion of skilled workers. On the contrary, if the immigrants have, on average, levels of skill identical to those of the natives, immigration has the effect of increasing production while leaving wage inequality untouched. When the immigrants are less skilled than the natives, immigration helps to reduce the relative number of skilled workers, ν, which increases their wage and reduces that of the unskilled. So the immigration of low-skilled workers does have the effect of deepening the inequality between the skilled and the unskilled.

Overall, the picture painted by the labor demand model indicates that the immigration of low-skilled workers increases inequality. This prediction is not, however, ironclad. Simulations of this model carried out by Borjas (1999) show that the impact of immigration on inequalities is small in extent. The elementary model of labor demand allows us to calculate the impact of variations in the quantities of the different inputs on their prices from our knowledge of the elasticities of substitution and of the shares of the factor remunerations in the total cost (see chapter 2). Borjas (1999) presents the

Table 11.10
The impact of an inflow of immigrants equal to 10% of the labor force.

Variation (%)	Capital fixed	Price of capital fixed
Earnings of capital	6.43	—
Earnings of skilled workers	−2.29	0.46
Earnings of unskilled workers	−3.72	−4.27
Dollar gain to natives over GDP	0.27	0.14

Note: The boundary between the unskilled and the skilled corresponds to a high school diploma.

Source: Simulations made by Borjas; see Borjas (1999, table 1).

results of simulations for the U.S. economy, using a production function comprising three arguments: capital K, skilled labor L_h, and unskilled labor L_ℓ. In the United States in 1995, if we take a high school diploma as marking the boundary between the unskilled and the skilled, the skilled represent 91% of the labor force, but only 68% of the migrant population. Assuming that this proportion holds, Borjas studies the impact of a 10% increase in the labor force as a result of immigration. He considers several plausible values of the elasticities of labor demand and capital demand. Table 11.10 presents the results for intermediate values of these elasticities. Overall, the simulations carried out point to the conclusion that immigration has a limited impact on wages. These orders of magnitude imply that immigration explains no more than a very small part of the evolution of wage inequality in the United States.

3.2.3 The Influence of Technological Progress and International Trade

A number of arguments undermine the generality of the notion that the immigration of low-skilled workers increases inequality. These include the endogeneity of the technological bias, the influence of international trade, and access to social assistance.

The very simple model of labor demand used to study the impact of immigration on factor remuneration leaves out the response of technological progress to changes in labor supply. We have pointed out, in chapter 10 when discussing the possibility of endogenous technological progress, that the interactions between technological progress and labor demand might lead to an increasing relation between the relative supply of skilled labor and the relative wage of this type of labor. It is indeed possible for firms to promote innovation using techniques that complement the type of labor that is most abundant. In consequence, an increase in the relative supply of low-skilled labor may bend the technological bias in their favor, and entail, in the end, if strong enough, an increase in their relative wage.

Another limitation of the labor demand model lies in its failure to take international trade into account. Actually, it turns out that in an open economy, immigration may have no impact on inequality whatever its composition (Johnson and Stafford, 1999). If we go back to the model from section 1.2, the wage level of each category of labor is given by relation (11.4), or $\bar{w}_i = A_i F_i(\alpha\bar{\nu}, 1)$, $i = h, \ell$, with $\bar{\nu} = \left(N_h + \tilde{N}_h\right) / \left(N_\ell + \tilde{N}_\ell\right)$. In an economy facing international competition in the goods market, wages depend on the *global* structure of labor supply, independently of where it is located. By equalizing the prices of inputs, international trade has the effect of

neutralizing the impact of migrations on wages. Here again, this textbook case illustrates a very stylized situation, in which the only source of heterogeneity among countries lies in their factor endowments. If we take a situation in which countries utilize different technologies, equation (11.5) shows that equilibrium wages depend on the ratio $\left(A_h N_h + \tilde{A}_h \tilde{N}_h\right) / \left(A_\ell N_\ell + \tilde{A}_\ell \tilde{N}_\ell\right)$, and are thus influenced by where the inputs are located. For example, if low-skilled migrants are less productive than in their country of origin, immigration leads to a reduction in the global productivity of low-skilled labor—represented by quantity $\left(A_\ell N_\ell + \tilde{A}_\ell \tilde{N}_\ell\right)$—and thus an increase in ratio $\left(A_h N_h + \tilde{A}_h \tilde{N}_h\right) / \left(A_\ell N_\ell + \tilde{A}_\ell \tilde{N}_\ell\right)$, which entails a wage reduction for all low-skilled workers—see equation (11.5). It should be noted that immigrants may be attracted to a country where they are less productive than they are in their countries of origin because of differences between, for example, collective goods or amenities.

Finally, immigrants, because they are generally unskilled, resort more frequently to social assistance and unemployment insurance than natives (Borjas and Hilton, 1996, for the United States; Brücker et al., 2001, and Zorlu, 2011, for Europe). From this perspective, if the fiscal system is progressive, immigration, by increasing the amount of payroll deductions, may compress the magnitude of take-home pay and so reduce inequality. The corollary of this reduction in inequality is evidently a transfer from natives to immigrants, which reduces the surplus the natives derive from immigration. If these transfers are large, this surplus can even become negative.

These different lines of reasoning show that the immigration of low-skilled workers has, in theory, ambiguous effects on inequality. Empirical research has much to tell us about this matter.

3.3 Empirical Results

In essence, two methods are used to study the impact of migration on the labor market. The first analyzes correlations between spatial movements of workers and earnings at the aggregate level. We illustrate this method using the framework of Boustan et al. (2010) for the United States during the Great Depression of the 1930s. Data and programs are available at www.labor-economics.org. The second method relies on natural experiments. The results of empirical research converge to suggest that migrations have a very feeble impact on inequality.

3.3.1 Spatial Correlations

The elementary model of labor demand concludes that wages, or the probability of employment, for workers who are highly substitutable by immigrants ought to be reduced by immigration. The method of spatial correlations aims to test this type of prediction and to assess the influence of immigration on the opportunities of natives.

The Basic Regression

The method of spatial correlation consists of estimating the correlation between the variation in the number of migrants Δm_{ijt} of skill level i, in region j between dates $t-1$ and t, on the variations in the employment opportunities (wages or probability of employment), Δy_{ijt}, of similarly skilled native workers present in region j at dates t and $t-1$. Let $\mathbf{x}_{it}$ be a vector of characteristics of natives and of a labor market of skill

level i at date t (age, sex, size of the market…) and ε_{ijt} a disturbance term. We then estimate the equation:

$$\Delta y_{ijt} = a_t \Delta m_{ijt} + \mathbf{x}_{it}\mathbf{b}_t + \varepsilon_{ijt} \quad \textbf{(11.30)}$$

Estimation of parameter a_t by ordinary least squares generally leads to results not significantly different from zero, with average values that change erratically according to periods (see Friedberg and Hunt, 1995; Borjas et al., 1997; Borjas, 1999; Hartog and Zorlu, 2005; and Longhi et al., 2005, for surveys of the literature). This is at odds with the short-run prediction of the basic model presented in the previous sections whereby higher levels of immigration should lower the wage of competing workers.

This approach raises delicate problems however. The first arises from the endogeneity of the number of new migrants, inasmuch as the latter are attracted by regions where wages are rising. That being so, the observation of a positive correlation between employment opportunities and variations in the number of migrants may simply reflect migrants' choice of where to settle. It is possible to solve this problem by using the instrumental variables method: attempts to do so assume that the immigrants are attracted by the presence of compatriots, and take the foreign-born proportion of the labor force at $t-1$ as an instrument for the variation in the number of migrants between dates t and $t-1$. The results obtained using these methods still pose the same problems as those obtained by ordinary least squares, inasmuch as they are not generally significantly different from zero, with average values that change erratically according to periods.

The second problem arises from the mobility of natives, who may themselves leave regions that receive an inflow of immigrants. Quite clearly, if every immigrant drives out a native, it is not surprising to find that immigration has no impact on wages, in the model of spatial correlation represented by equation (11.30). Card and DiNardo (2000) suggest however that this problem is not statistically significant in the United States.

The third problem is what Aydemir and Borjas (2011) call the "attenuation bias." Indeed, the estimated wage impact of immigration could be attenuated by measurement error of the key independent variable in the analysis, that is, the variation in the number (or fraction) of migrants in the local labor market. This number is typically estimated from a sample of workers, in which immigrants might be underestimated because they represent a small fraction of the total population. In that case, after controlling for permanent factors that determine wages, there is little identifying variation left in the variable that captures the immigrant supply shift, permitting any sampling error in the immigrant share to play a disproportionately large role. Correcting for the resulting attenuation bias can substantially increase estimates of the wage impact of immigration, notably for the United States and Canada.

An Example: Migration During the Great Depression

To study the impact of migration on labor market outcomes, substantial migration flows are needed. Such flows can be triggered by exceptional events. For instance, the Great Depression in the 1930s in the United States triggered massive movements of population across the country. Boustan et al. (2010) use this unique episode to investigate the causal

impact of migration on local labor markets. They regress labor market outcomes in 1940 on the migration flows between 1935 and 1940 and a set of control variables:

$$y_{j_{40}} = \alpha + \beta(m_{j,40-35} - o_{j,40-35}) + \mathbf{x}_{j40}\mathbf{b} + \phi y_{j_{35}} + \varepsilon_{j40} \tag{11.31}$$

where $m_{j,40-35}$ is the inward flow of migrants over the period 1935–1940 in city j, while $o_{j,40-35}$ represents the outward flow of migrants from city j, and $\mathbf{x}_{j40}$ is a vector of control variables, such as the share of blacks or foreign-born in the population, or the unemployment rate in 1940, and $y_{j_{35}}$ is the lagged outcome at the beginning of the period. Since migrant location choices can be determined by the local labor market conditions (such as higher wages or higher employment rates) in the area to which people decide to move, this might generate an upward bias in the OLS estimation.

To solve this problem, one strategy is to instrument these flows with variables that are strongly correlated with them but are not correlated with the local labor market condition of the cities or region where people decide to move (see chapter 4 for a detailed presentation of the instrumental variables method). To this end, the authors use the variation in the generosity of New Deal programs across the country. Indeed, while positive economic shocks pulling migrants from area k to area j are likely correlated to labor market outcomes in area j, poor economic conditions pushing migrants to leave area k are arguably exogenous to labor market conditions in j. Therefore, local economic conditions in areas that typically send migrants to destination j are natural instruments for in-migration to that destination. As such, areas where the New Deal programs were more generous, including work relief and public works projects, were less likely to generate large migration outflows. To these federal programs, state or local officials added a series of work relief projects, causing the level of funding both between and within states to vary widely. Additionally, the authors used the variation in local weather conditions (differences in temperatures and precipitation) as another instrument for migrant flows to and from U.S. cities, as well as the square of the distance between areas. So the instrument for in-migration to a given city j is the sum over all areas of the predicted out-migration from other cities departing toward city j, using this set of variables. More precisely, in the first stage, Boustan et al. (2010) estimate the following equations (dropping period indices for clarity):

$$o_k = \alpha + \mathbf{Z}_t\delta + \varepsilon_k \tag{11.32}$$

$$p_{kj} = \alpha_k + \theta_k(\text{distance}_{kj}) + \gamma_k(\text{distance}_{kj})^2 + \mu_k \tag{11.33}$$

$$m_j = \sum\nolimits_k o_k.p_{kj} \tag{11.34}$$

where o_k is the outflow of migrants from region k over the considered period, and $\mathbf{Z}_t$ is the set of instruments including New Deal programs and a set of extreme weather conditions, while p_{kj} is the share of migrants leaving k and reaching j which is influenced by the proximity of regions (distance between k and j). Then the flow of migrants reaching destination j is the sum of the predicted total outflows from all other regions times the share of migrants heading to j.

Based on the 1940 census, which gathers systematic information on internal mobility in the United States, and focusing on the 86 cities with more than 100,000 residents at that time, the IV estimation of the coefficient beta in equation (11.31)

Table 11.11
Effect of net migration on earnings, wage, and working time in 1940.

Dependent variable	OLS	IV
ln(annual earnings)	−.218 (.578)	−.948 (.610)
ln(weekly wage)	.561 (.379)	.006 (.536)
ln(hourly wage)	.471 (.545)	−.521 (.730)
ln(weeks worked)	−.779 (.325)	−.954 (.304)
Work less than 26 weeks	.402 (.200)	.528 (.196)
ln(hours worked)	.089 (.267)	.527 (.295)

Note: Number of observations: 96,070. Data are coefficients on net number of migrants between 1935 and 1940 as a percentage of the 1935 population. Regressions estimated at the individual level. The sample includes only men employed during the census week who report positive earnings. Controls included: city-level variables for the share of blacks, foreign-born, and illiterate in 1930, as well as age distribution, unemployment rate, and lagged annual earnings. Standard errors in parentheses.

Source: Boustan et al. (2010, table 3).

displayed in table 11.11 shows that in-migration had little effect on the hourly earnings of existing residents (whereas in the OLS specification, net migration has a positive effect on wages, although not significant). To show this, the authors regressed weekly and hourly earnings in 1940 on the predicted in-migration rates (using the set of instruments) and the probability of out-migration among existing residents by metropolitan area, over the period 1935–1940. However, in-migration prompted some residents to move away and others to lose weeks of work or access to relief jobs. In table 11.11 the coefficient on the number of weeks worked is negative and significant, although this resulted in a small and only marginally statistically significant decrease in annual earnings. To investigate the impact on employment more clearly, the authors analyzed the effect of migration on the probability that an individual was employed, on work relief, or idle during the year. They find that in-migration to a metropolitan area reduced work opportunities somewhat for existing residents, conditional on being out of work, and increased the probability of leaving the area altogether. For every 10 arrivals into an area, they estimate that 1.9 existing residents left the area, 2.1 were prevented from finding a relief job, and 1.9 shifted from full-time to part-time work. So overall, the adjustment to migration supply shocks occurred more through reductions in employment opportunities than through lower wages, which is consistent with the presence of sticky wages and high unemployment during the Depression.

The Importance of Capital Mobility

The impact of migration flows on labor market outcomes might also depend on the ability to adjust other factors, such as capital, to absorb the increase in the labor force. For instance, Strobl and Valfort (2013) study the impact of weather-induced net internal migration rates on the employment probability of nonmigrants within regions in Uganda. More precisely, they identify the impact of the net migration rates in a given region on the employment probability of the nonmigrants in that region, using individual census data. The problem is that migration inflows might be influenced by

the employment rate of nonmigrants or by some other missing variable influencing both employment and migration inflows, such as work opportunities at the regional level (simultaneity and omitted variable problems). To solve this difficulty, they instrument migration inflows in a given region using (1) weather shocks affecting the other regions and (2) the geographic distance between these other regions and the region under scrutiny. They also control for the level of economic development in each region proxied by the average nightlights intensity. They find a larger negative impact of migration on local labor outcomes than the one documented for developed countries: a 10 percentage point increase in the net in-migration rate decreases the employment probability of nonmigrants in the destination region by 7.8 percentage points. This effect is twice as large as that obtained when migration rates are not instrumented. Moreover, the effect is not constant across regions. Using data on road density in Uganda, the authors show that the negative impact is significantly stronger in regions with below-median road density, that is, less conducive to capital mobility: in these regions a 10 percentage point increase in the net in-migration rate decreases the probability of being employed of nonmigrants by more than 10 percentage points.

Overall, the approach based on spatial variations of the share of migrants on the labor market outcomes of native workers yields no evidence of any substantial impact on wages. This absence of effect is robust to the use of instrumental variables to deal with the endogeneity of migration choices to local labor market conditions. However, the impact on employment opportunities could be somewhat larger than that on wages, at least in the short run.

3.3.2 Natural Experiments

To solve the difficulties encountered by research based on spatial correlations, other studies have looked at certain exceptional flows of migration—most often due to political events, like the Cuban immigration to Miami in May 1980 (Card, 1990) and immigration to France in the wake of Algerian independence in 1962 (Hunt, 1992)—as "natural experiments."

The Cuban Immigration to Miami in May 1980

Card's (1990) study deals with the Cuban immigration, which swelled the labor force of Miami by around 7% between May and September 1980, following the opening of Cuba's borders. Card's strategy was to compare the evolution of unemployment rates and wages in Miami with those of cities presenting characteristics taken to be similar for this purpose but which did not undergo the same inflow of migrants. Examination of the evolution of these variables before 1980 led Card to select Atlanta, Los Angeles, Houston, and Tampa–St. Petersburg, cities that, like Miami, have large black and Hispanic populations. The impact of the immigration was assessed with the help of a difference-in-differences estimator, which consists of comparing the changes in the variables pertaining to the group studied in Miami and those pertaining to the control group in the other cities between 1979 and subsequent years (see chapter 14 for a more detailed presentation of this approach). More precisely, let Δu_m be the variation in Miami's unemployment rate between 1979 and a subsequent year (1981 for example), and let Δu_c be the average variation in the unemployment rate in the other cities

Table 11.12
Difference-in-differences estimates of the impact of immigration on the unemployment rate of the black population in Miami in 1980.

Unemployment rate (%)	1979	1981	1981–1979
Miami	8.3 (1.7)	9.6 (1.8)	$\Delta u_m = 1.3$ (2.5)
Other cities	10.3 (0.8)	12.6 (0.9)	$\Delta u_c = 2.3$ (1.2)
Miami − other cities	−2 (1.9)	−3 (2.0)	$\Delta u_m - \Delta u_c = -1.0$ (2.8)

Note: The figures between parentheses are standard deviations.

Source: Angrist and Krueger (1999, table 4).

over the same span of time. The estimated impact of the immigration on the unemployment rate is simply equal to $\Delta u_m - \Delta u_c$. Table 11.12 shows that the immigration had no significant impact on the differences in the evolution of unemployment rates of black workers (those most exposed to competition from the new refugees) between 1979 and 1981, since the difference-in-differences estimator takes a value of −1 (meaning that the unemployment rate rose less in Miami than in the other cities during this period) with a standard error of 2.8. The results for wages are of the same order, and would be similar if we had considered the white population (see the introduction to this book).

The Consequences of Algerian Independence for the French Labor Market

The study by Hunt (1992), which deals with a massive flow of migration that swelled the labor force in France by 1.6% in 1962 in the wake of Algerian independence, also finds that migration had only a small impact on unemployment and wages. Following the war in Algeria, the Accords d'Evian were signed in March 1962 and granted independence to a region of North Africa which was previously a French administrative district. After the agreement, a real exodus started. In total, 900,000 people, mostly of European origin and holding French citizenship, repatriated to France. Over just a few months, between May and August 1962, about 500,000 people arrived, mostly in southern regions of France, where the climate is the closest to that of the Algerian coast. They were usually skilled but came to France with few assets and needed to work. Overall, the repatriated participating in the labor market represented 1.6% of the total French labor force. Hunt (1992) shows that a few years later, in 1967, annual wages were 1.3% lower due to the exodus (i.e., a 1 point rise in the share of the repatriated in the total French labor force led to a drop of 0.8% in the average wage), while in 1968 unemployment among the nonrepatriated rose by 0.3 percentage points (a 1 point rise in the share of the repatriated in the total French labor force entailed unemployment higher by 0.2 percentage points among the nonrepatriated). There was no effect of immigration on labor force participation of the nonrepatriated. Elasticities are therefore of rather small magnitude.

The Collapse of the USSR and the Fall of the Berlin Wall

The collapse of the USSR is another example of a "natural experiment" that triggered large migratory flows in a number of countries, and notably to Israel and Germany. For

instance, between 1989 and 1995, 610,100 immigrants arrived in Israel from the former Soviet Union, increasing the size of the Israeli population by 13.6%. This exodus was triggered by the lifting of emigration restrictions in an unstable USSR and by the open immigration policy of Israel toward Soviet Jews, who faced more restrictive entry policies elsewhere. Using panel surveys of new immigrants, such as the Immigrant Employment Survey (IES) which covers a large sample of new immigrants who arrived in Israel in 1990, Friedberg (2001) finds no adverse impact on native Israeli labor market outcomes. She regressed native Israeli wages at the individual level on the share of Russians in individual occupations. Actually, as shown in table 11.13 the OLS estimates yield significant reductions in hourly wages and small reductions in employment. But as explained above, OLS are likely to be biased if the distribution of immigrants across occupations in Israel is not exogenous to relative wage and employment conditions. This is probably the case, since individuals are likely to choose occupations where wages are higher and labor demand is strong. For this reason, Friedberg instruments the entry of Russians into a given occupation using their former occupations in the USSR, which cannot be correlated with labor market conditions in Israel subsequent to their migration. The correlation coefficient between the number of Russians employed in one occupation in Israel in 1994 and the number of Russians formerly employed in the same occupation is equal to 0.37. When using this instrument, estimates indicate that immigration did not have an adverse impact on native Israeli labor market outcomes, as shown by the second column of table 11.13. On the contrary the increase in employment due to the inflow of Russians led to a significant rise in the hourly earnings of Israelis.

Similarly, with the fall of the Berlin Wall, ethnic Germans living in eastern Europe and the former Soviet Union were given the opportunity to migrate to Germany. In 1988, with the end of the Cold War looming, travel restrictions in central and eastern Europe were lifted. In 1990 alone, some 397,000 individuals, mainly from the former Soviet Union (37%), Poland (34%), and Romania (28%), arrived in Germany. Faced with these enormous movements, the government limited their inflow in subsequent years to around 225,000 per year. Within 15 years, 2.8 million individuals had migrated. From 1993 onward, more than 90% of the ethnic German immigrants originated from territories of the former Soviet Union. Glitz (2012) analyzes the impact of this massive shock on the German local labor market. More precisely, he analyzes the relative supplies of different skill groups in a locality, and their impact on wages and employment among

TABLE 11.13
Effect of immigration (share of Russians in Israel by occupation) on native Israeli wages.

Dependent variable	OLS	IV
ln(hourly wage)	−.324 (.086)	.718 (.343)
R^2	.53	

Note: Number of observations: 8,353. Individual level regressions, controlling for sex, ethnicity, nativity, education, potential market experience, and including occupation and industry dummies. Robust standard errors, clustered by occupation and by year, in parentheses.

Source: Friedberg (2001, table 3).

native workers. He uses the exogenous immigrant inflows to instrument for the potentially endogenous changes in relative skill shares in a locality. Upon arrival, immigrants were exogenously allocated to different regions to ensure an even distribution across the country. Hence immigration in this setting can be viewed as a quasi-experiment. Similarly to studies reviewed above on the United States and Israel, he finds a displacement effect of 3.1 unemployed workers for every 10 immigrants who find a job, but no effect on relative wages.

Overall, research suggests that immigration often has little impact on inequality as regards wages and access to employment. The share of migrants can exert only a modest impact, if any, on job opportunities and wages of native workers, even following large and exceptional flows of migration. However, some studies find some significant impact of immigration on labor market outcomes, which seems to be related to the mobility of capital and to the flexibility of labor markets. From this perspective, more research is needed to understand more precisely the impact of immigration on labor market outcomes.

4 SUMMARY AND CONCLUSION

- Over the last 40 years, trade as a percentage of GDP has been multiplied by three in the United States, Germany, and Japan, and by two in France, Sweden, and many other OECD countries. This increase has been driven in part by imports from the developing economies, and notably China, which is now the main source of imports into these countries and also a growing destination of exports.
- In the meantime migration flows were also multiplied by two in the advanced economies, notably due to growing inflows into European countries that used to be countries of emigration until the 1960s or the 1970s.
- These changes are at the origin of the growing awareness of what is called globalization, with its potential to exert pressure on the job opportunities and wages of low-skilled workers in the advanced economies.
- The classical models of trade predict that countries intensive in capital should suffer from more low-skilled unemployment and/or lower wages for the low skilled. The new trade theories insist on the selection effect, by which trade favors more-productive firms, which helps decrease unemployment and increase wages as well as aggregate productivity. The empirical evidence, at both the macro and micro levels, tends to favor the latter theories.
- Examination of migratory flows shows that the rich countries tend to have an immigrant population less well qualified, on average, than natives. From a theoretical standpoint, the immigration of low-skilled workers has an ambiguous effect on inequality. Empirical work, based notably on exceptional migration events, confirms this conclusion, suggesting that the immigration of low-skilled workers often has little effect on wages and employment among workers with the fewest skills.

5 RELATED TOPICS IN THE BOOK

- Chapter 2, section 1: The static theory of labor demand
- Chapter 3, section 1.3: The effect of a shock on labor supply
- Chapter 8, section 3.3: Direct assessment of discrimination
- Chapter 8, section 4.1: Race- and ethnicity-related discrimination
- Chapter 9, section 3: The matching model
- Chapter 10, section 2.3.1: Wage inequality between high- and low-skilled workers
- Chapter 13, section 2: Employment protection

6 FURTHER READINGS

Borjas, G. (1999). The economic analysis of immigration. In O. Ashenfelter & D Card (Eds.), *Handbook of labor economics* (vol. 3A, chap. 28, pp. 1697–1760). Amsterdam: Elsevier Science.

Constant, A., & Zimmermann, K. (Eds.). (2013). *International handbook on the economics of migration*. Cheltenham, U.K.: Edward Elgar.

Felbermayr, G., Prat, J., & Schmerer, H.-J. (2011). Trade and unemployment: What do the data say? *European Economic Review*, *55*(6), 741–758.

Helpman, E. (2010). Labor market frictions as a source of comparative advantage, with implications for unemployment and inequality (NBER Working Paper No. 15764). National Bureau of Economic Research, Inc.

Longhi, S., Nijkamp, P., & Poot, J. (2005). A meta-analytic assessment of the effect of immigration on wages. *Journal of Economic Surveys*, *19*(3), 451–477.

OECD. (2012a). *Divided we stand*. Paris: OECD Publishing.

7 APPENDIX

In this appendix we briefly present the Arellano-Bond GMM estimator. See Cameron and Triverdi (2010) and Wooldridge (2010) for further details.

Let us rewrite equation (11.28) using simplified notations:

$$\widetilde{y}_{it} = \widetilde{\mathbf{X}}_{it}\boldsymbol{\phi} + \widetilde{\varepsilon}_{it} \tag{11.35}$$

where the whole set of explanatory variables, including the lagged dependent variable, is now denoted $\widetilde{\mathbf{X}}_{it} = (\Delta y_{it-1}, \Delta T_{it})$, $\widetilde{y}_{it} = \Delta y_{it}$, $\widetilde{\varepsilon}_{it} = \Delta\varepsilon_{it}$, with $t = 3 \ldots \bar{t}$. The vector $\boldsymbol{\phi}$ stands for the set of parameters to be estimated. Then the GMM estimator of Arellano

and Bond is similar to an instrumental variable estimator using weights for each observation to take into account the change in the instrument set over time:

$$\widehat{\boldsymbol{\phi}}_{GMM} = \left[\left(\sum_{i=1}^{N} \widetilde{\mathbf{X}}_i' \mathbf{Z}_i\right) \mathbf{W} \left(\sum_{i=1}^{N} \mathbf{Z}_i' \widetilde{\mathbf{X}}_i\right)\right]^{-1} \left(\sum_{i=1}^{N} \widetilde{\mathbf{X}}_i' \mathbf{Z}_i\right) \mathbf{W} \left(\sum_{i=1}^{N} \mathbf{Z}_i' \widetilde{\mathbf{Y}}_i\right)$$

If K is the total number of explanatory variables in equation (11.28), $\widetilde{\mathbf{X}}_i$ is a $(\bar{t}-2, K)$ matrix (made of the $\widetilde{\mathbf{X}}_{it}$), $\widetilde{\mathbf{Y}}_i$ is a $(\bar{t}-2, 1)$ vector (made of the $\widetilde{y}_{it}$), and $\mathbf{Z}_i$ is a $(\bar{t}-2, v)$ matrix of v instruments defined by:

$$\mathbf{Z}_i = \begin{pmatrix} z_{i_3} & 0 & \dots & 0 \\ 0 & z_{i_4} & & \\ & & \dots & 0 \\ 0 & \dots & 0 & z_{i\bar{t}} \end{pmatrix}$$

with $z_{i_t} = (y_{i,t-2}, y_{i,t-3}, \dots, y_0, \Delta T_i)$. $\mathbf{W}$ is a weighting matrix that takes into account the heterogeneity of instruments over time and minimizes the mean of errors. An optimal choice is $\mathbf{W} = \left[\frac{1}{N}\sum_{i=1}^{N} \mathbf{Z}_i' \widetilde{\boldsymbol{\varepsilon}}_i \widetilde{\boldsymbol{\varepsilon}}_i' \mathbf{Z}_i\right]$ where N is the number of countries in the panel, and where $\widetilde{\boldsymbol{\varepsilon}}_i$ is a vector made of $\widetilde{\varepsilon}_{it}$.

Consider the following matrix expression of (11.35) where both sides have been multiplied by $\mathbf{Z}$ the instrument matrix: $\mathbf{Z}'\widetilde{\mathbf{Y}} = \mathbf{Z}'\widetilde{\mathbf{X}}\boldsymbol{\phi} + \mathbf{Z}'\widetilde{\boldsymbol{\varepsilon}}$, which can be rewritten to simplify notations:

$$\widehat{\mathbf{Y}} = \widehat{\mathbf{X}}\boldsymbol{\phi} + \widehat{\boldsymbol{\varepsilon}} \tag{11.36}$$

where $\widehat{\mathbf{y}} = \mathbf{Z}'\widetilde{\mathbf{y}}, \widehat{\mathbf{x}} = \mathbf{Z}'\widetilde{\mathbf{x}}$, and $\widehat{\boldsymbol{\varepsilon}} = \mathbf{Z}'\widetilde{\boldsymbol{\varepsilon}}$.

If $\widehat{\boldsymbol{\varepsilon}}$ does not have a variance-covariance matrix proportional to the identity matrix, then the OLS are not efficient. In that case we must give different weights to the different equations. Assume that we use a weighting matrix $\mathbf{W}$. Then minimizing the weighted square of errors $\widehat{\boldsymbol{\varepsilon}}'\mathbf{W}\widehat{\boldsymbol{\varepsilon}} = (\widehat{\mathbf{Y}} - \widehat{\mathbf{X}}\boldsymbol{\phi})'\mathbf{W}(\widehat{\mathbf{Y}} - \widehat{\mathbf{X}}\boldsymbol{\phi})$ gives:

$$\widehat{\boldsymbol{\phi}}_{GMM} = (\widehat{\mathbf{X}}'\mathbf{W}\widehat{\mathbf{X}})^{-1}\widehat{\mathbf{X}}'\mathbf{W}\widehat{\mathbf{Y}} \tag{11.37}$$

Now we want to choose the weighting matrix so as to achieve the lowest variance for this estimator. We get the most efficient estimator by weighting each equation by the inverse of the standard deviation of its error term, which suggests choosing the weighting matrix as the inverse of the variance matrix of the error term $\widehat{\boldsymbol{\varepsilon}}$: $\mathbf{var}(\widehat{\boldsymbol{\varepsilon}}) = \mathbf{var}(\mathbf{Z}'\widetilde{\boldsymbol{\varepsilon}}) = \mathbf{Z}'\mathbf{var}(\widetilde{\boldsymbol{\varepsilon}})\mathbf{Z} = (\mathbf{Z}'\widetilde{\boldsymbol{\varepsilon}}\widetilde{\boldsymbol{\varepsilon}}'\mathbf{Z})^{-1}$, which gives:

$$\begin{aligned} \widehat{\boldsymbol{\phi}}_{GMM} &= (\widehat{\mathbf{X}}'(\mathbf{Z}'\widetilde{\boldsymbol{\varepsilon}}\widetilde{\boldsymbol{\varepsilon}}'\mathbf{Z})^{-1}\widehat{\mathbf{X}})^{-1}\widehat{\mathbf{X}}'(\mathbf{Z}'\widetilde{\boldsymbol{\varepsilon}}\widetilde{\boldsymbol{\varepsilon}}'\mathbf{Z})^{-1}\widehat{\mathbf{Y}} \\ &= (\widetilde{\mathbf{X}}'\mathbf{Z}(\mathbf{Z}'\widetilde{\boldsymbol{\varepsilon}}\widetilde{\boldsymbol{\varepsilon}}'\mathbf{Z})\mathbf{Z}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'\mathbf{Z}(\mathbf{Z}'\widetilde{\boldsymbol{\varepsilon}}\widetilde{\boldsymbol{\varepsilon}}'\mathbf{Z})^{-1}\mathbf{Z}'\widetilde{\mathbf{Y}} \end{aligned}$$

and if $\mathbf{var}(\widetilde{\boldsymbol{\varepsilon}}) = \widetilde{\boldsymbol{\varepsilon}}\widetilde{\boldsymbol{\varepsilon}}' = \mathbf{I}$ we can simply choose $\mathbf{W}$ as $(\mathbf{Z}'\mathbf{Z})^{-1}$, which gives $\widehat{\boldsymbol{\phi}}_{GMM} = (\widehat{\mathbf{X}}'(\mathbf{Z}'\mathbf{Z})^{-1}\widehat{\mathbf{X}})^{-1}\widehat{\mathbf{X}}'(\mathbf{Z}'\mathbf{Z})^{-1}\widehat{\mathbf{Y}} = \widehat{\boldsymbol{\phi}}_{IV}$. The GMM estimator then corresponds to the 2SLS

estimator. If $\mathbf{var}(\widetilde{\varepsilon}) = \widetilde{\varepsilon}\widetilde{\varepsilon}' \neq \mathbf{I}$ then $W = \mathbf{Z}'\widetilde{\varepsilon}\widetilde{\varepsilon}'\mathbf{Z} = \left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{Z}_i'\widetilde{\varepsilon}_i\widetilde{\varepsilon}_i'\mathbf{Z}_i\right]$, where N is the number of countries in the panel.

Arellano and Bond (1991) have proposed several tests for the crucial identifying assumption stating that $Cov\,(y_{is}, \Delta\varepsilon_{it}) = 0$ for any $s \in [1, t-2]$, $t \geq 3$, that is, that the instruments are exogenous, which amounts to assuming that the residuals ε_{it} are not autocorrelated. To show this, let us first recall that if x, y, w, and u are real-valued random variables and a, b, c, d are constant, nonrandom variables, then:

$$Cov(ax + by, cw + du) = ac\ Cov(xw) + ad\ Cov(xu) + bc\ Cov(yw) + bd\ Cov(yu)$$

Thus, if the ε_{it} are serially uncorrelated, then $\Delta\varepsilon_{i,t}$ must be correlated with $\Delta\varepsilon_{i,t-1}$ at the first order, since:

$$Cov(\Delta\varepsilon_{it}, \Delta\varepsilon_{i,t-1}) = Cov(\varepsilon_{it} - \varepsilon_{i,t-1}, \varepsilon_{i,t-1} - \varepsilon_{i,t-2}) = -Cov(\varepsilon_{i,t-1}, \varepsilon_{i,t-1}) = -Var(\varepsilon_{i,t-1}) \neq 0$$

But, by assumption, $\Delta\varepsilon_{it}$ cannot be correlated with $\Delta\varepsilon_{i,t-k}$ if $k \geq 2$. Then a test of specification is to check if the first-differenced errors are autocorrelated at the first order but not at higher orders. See Cameron and Triverdi (2010, p. 300) for an example.

Another way to test the exogeneity of instruments is to run overidentifying restriction tests. These tests can be run when the model is overidentified, that is, where there are more instruments than strictly needed. When we instrument one given variable with two instrumental variables, there is one overidentification restriction, and if we use three IVs then there are two overidentification restrictions. One key assumption in using IVs is that they must be uncorrelated with the residual of the main equation at hand. Assuming that at least one of the IVs used is exogenous, it is possible to check if the other IVs are uncorrelated with the residual. The test amounts to checking if each IV is uncorrelated with the residual when the regression is run without it. If the test is passed, the set of instrumental variables can be deemed exogenous. For the overidentifying restriction test, see Cameron and Triverdi (2010, p. 191 and p. 301); for overidentifying restriction tests (also called Sargan's test), see Wooldridge (2013, p. 535).

REFERENCES

Abramitzky, R., Boustan, L., & Eriksson, K. (2012). A nation of immigrants: Assimilation and economic outcomes in the age of mass migration (NBER Working Paper No. 18011).

Alcalá, F., & Ciccone, A. (2004). Trade and productivity. *Quarterly Journal of Economics*, *119*(2), 612–645.

Angrist, J., & Krueger, A. (1999). Empirical strategies in labor economics. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 3, chap. 23). Amsterdam: Elsevier Science.

Arellano, M., & Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies*, *58*, 277–297.

Aydemir, A., & Borjas, G. (2011). Attenuation bias in measuring the wage impact of immigration. *Journal of Labor Economics*, *29*(1), 69–113.

Bernard, A., & Jensen, J. (1997). Exporters, skill upgrading, and the wage gap. *Journal of International Economics*, *42*(1–2), 3–31.

Bernard, A., & Jensen, J. (2004). Entry, expansion, and intensity in the US export boom, 1987–1992. *Review of International Economics*, *12*(4), 662–675.

Biscourp, P., & Kramarz, F. (2007). Employment, skill structure and international trade: Firm-level evidence for France. *Journal of International Economics*, *72*(1), 22–51.

Borjas, G. (1999). The economic analysis of immigration. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 3A, chap. 28, pp. 1697–1760). Amsterdam: Elsevier Science.

Borjas, G., Freeman, R., & Katz, L. (1997). How much do immigration and trade affect labor market forces? *Brookings Paper on Economic Activity*, *1*, 1–85.

Borjas, G., & Hilton, L. (1996). Immigration and the welfare state: Immigrant participation in means-tested entitlement programs. *Quarterly Journal of Economics*, *111*, 575–604.

Boustan, L., Fishback, P., & Kantor, S. (2010). The effect of internal migration on local labor markets: American cities during the great depression. *Journal of Labor Economics*, *28*(4), 719–746.

Brücker, H., Epstein, G., McCormick, B., Saint-Paul, G., Venturini, A., & Zimmerman, K. (2001). *Managing migration in the European welfare state*. Mimeo, DIW Berlin.

Cameron, A., & Triverdi, P. (2010). *Microeconometrics using Stata*. College Station, TX: Stata Press.

Card, D. (1990). The impact of the Mariel boatlift on the Miami labor market. *Industrial and Labor Relations Review*, *43*, 245–257.

Card, D., & DiNardo, J. (2000). Do immigrant inflows lead to native outflows? *American Economic Review*, *Papers and Proceedings*, *90*, 361–367.

Chiswick, B. (1978). The effect of Americanization on the earnings of foreign-born men. *Journal of Political Economy*, *86*(5), 897–921.

Davidson, C., & Matusz, S. (2004). An overview of the issue. In *International trade and labor markets: Theory, evidence, and policy implications* (pp. 1–15). Kalamazoo, MI: W. E. Upjohn Institute for Employment Research.

Dixit, A., & Stiglitz, J. (1977). Monopolistic competition and optimum product diversity. *American Economic Review*, *67*(3), 297–308.

Dutt, P., Devashish, M., & Priya, R. (2009). International trade and unemployment: Theory and cross-national evidence. *Journal of International Economics*, *78*(1), 32–44.

Egger, P., Pfaffermayr, M., & Weber, A. (2007). Sectoral adjustment of employment to shifts in outsourcing and trade: Evidence from a dynamic fixed effects multinomial logit model. *Journal of Applied Econometrics*, *22*(3), 559–580.

Eslava, M., Haltiwanger, J., Kugler, A., & Kugler, M. (2013). Trade and market selection: Evidence from manufacturing plants in Colombia. *Review of Economic Dynamics*, *16*(1), 135–158.

Eurostat. (2011). *External and intra-EU trade: A statistical yearbook: Data 1958–2010*. Luxembourg: Publications Office of the European Union.

Felbermayr, G., Larch, M., & Lechthaler, W. (2013). Unemployment in an interdependent world. *American Economic Journal: Economic Policy*, *5*(1), 262–301.

Felbermayr, G., Prat, J., & Schmerer, H.-J. (2011a). Globalization and labor market outcomes: Wage bargaining, search frictions, and firm heterogeneity. *Journal of Economic Theory*, *146*(1), 39–73.

Felbermayr, G., Prat, J., & Schmerer, H.-J. (2011b). Trade and unemployment: What do the data say? *European Economic Review*, *55*(6), 741–758.

Frankel, J., & Romer, D. (1999). Does trade cause growth? *American Economic Review*, *89*(3), 379–399.

Friedberg, R. (2001). The impact of mass migration on the Israeli labor market. *Quarterly Journal of Economics*, *116*(4), 1373–1408.

Friedberg, R., & Hunt, J. (1995). The impact of immigrants on host country wages, employment and growth. *Journal of Economic Perspectives*, *9*, 23–34.

Glitz, A. (2012). The labor market impact of immigration: A quasi-experiment exploiting immigrant location rules in Germany. *Journal of Labor Economics*, *30*(1), 175–213.

Hartog, J., & Zorlu, A. (2005). The effect of immigration on wages in three European countries. *Journal of Population Economics*, *18*(1), 113–151.

Helpman, E., & Itskhoki, O. (2010). Labor market rigidities, trade and unemployment. *Review of Economic Studies, 77*(3), 1100–1137.

Helpman, E., Itskhoki, O., & Redding, S. (2010a). Inequality and unemployment in a global economy. *Econometrica*, *78*(4), 1239–1283.

Helpman, E., Itskhoki, O., & Redding, S. (2010b). Unequal effects of trade on workers with different abilities. *Journal of the European Economic Association*, *8*(2–3), 421–433.

Hunt, J. (1992). The impact of the 1962 repatriates from Algeria on the French labor market. *Industrial and Labor Relations Review*, *43*, 245–257.

International Monetary Fund. (2007). *World economic outlook: Globalization and inequality*. Washington, DC: IMF Publication Services.

Johnson, G., & Stafford, F. (1999). The labor market implications of international trade. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 3, chap. 34, pp. 2215–2288). Amsterdam: Elsevier.

Krugman, P. (1979). Increasing returns, monopolistic competition, and international trade. *Journal of International Economics*, *9*(4), 469–479.

Krugman, P. (1980). Scale economies, product differentiation, and the pattern of trade. *American Economic Review*, *70*(5), 950–959.

Krugman, P., & Obstfeld, M. (2009). *International economics: Theory and policy*. Boston, MA: Pearson Addison-Wesley.

Longhi, S., Nijkamp, P., & Poot, J. (2005). A meta-analytic assessment of the effect of immigration on wages. *Journal of Economic Surveys*, *19*(3), 451–477.

Melitz, M. (2003). The impact of trade on intraindustry reallocations and aggregate industry productivity. *Econometrica*, *71*, 1695–1725.

Munch, J., & Skaksen, J. (2008). Human capital and wages in exporting firms. *Journal of International Economics*, *75*(2), 363–372.

OECD. (2012a). *Divided we stand*. Paris: OECD Publishing.

OECD. (2012b). *Education at a glance*. Paris: OECD Publishing.

OECD. (2012c). *Migration at a glance*. Paris: OECD Publishing.

Schank, T., Schnabel, C., & Wagner, J. (2007). Do exporters really pay higher wages? First evidence from German linked employer-employee data. *Journal of International Economics*, *72*(1), 52–74.

Stolper, W., & Samuelson, P. (1947). Protection and real wages. *Review of Economic Studies*, *9*, 58–73.

Strobl, E., & Valfort, M.-A. (2013). The effect of weather-induced internal migration on local labor markets: Evidence from Uganda. *World Bank Economic Review,* forthcoming.

Temin, P. (1999). Globalization. *Oxford Review of Economic Policy*, *15*(4), 76–89.

Verhoogen, E. (2008). Trade, quality upgrading, and wage inequality in the Mexican manufacturing sector. *Quarterly Journal of Economics*, *123*(2), 489–530.

Wooldridge, J. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). Cambridge, MA: MIT Press.

Wooldridge, J. (2013). *Introductory econometrics: A modern approach* (5th ed.). Mason, OH: South-Western College Publishing.

World Trade Organization. (2012). *International trade statistics*. Geneva: World Trade Organization.

Zorlu, A. (2011). Immigrant participation in welfare benefits in the Netherlands (IZA Discussion Paper No. 6128). Institute for the Study of Labor.

PART FOUR

Public Policies

CHAPTER 12

Income Redistribution

In this chapter we will:

- Learn about the differences among the fiscal regimes of the industrialized countries
- Understand the impact of taxes on employment, unemployment, labor market participation, and hours of work
- Estimate the impact of the Earned Income Tax Credit in the United States, using data and programs that replicate the main results of the paper of Eissa and Leibman (1996) (These data and programs are available at www.labor-economics.org)
- See to what extent differences among taxation regimes can explain differences in hours worked
- Review the various ways of regulating the minimum wage in various countries
- Understand the impact of the minimum wage on employment, unemployment, and labor market participation
- Estimate the impact of minimum wage hikes in the United States, using data and programs that replicate the main results of the paper of Card and Krueger (1994) on the fast-food industry in New Jersey and Pennsylvania (These data and programs are available at www.labor-economics.org)
- Learn what empirical studies show about the impact of the minimum wage on employment and inequality
- Observe to what extent the minimum wage is useful when the government can also use taxes

INTRODUCTION

In most countries, the stated intention of the government is to ensure the highest possible standard of living for citizens while compressing the spread of income inequality. To meet this twofold objective, the state levies taxes, a portion of which are redistributed; and in certain cases, it sets a minimum wage. In pursuing this double goal, the

government is forced to engage in arbitrage between equity and efficiency, since limiting the spread between the highest and the lowest incomes can in certain cases lead to a diminution in the volume of hours worked. That may come about when fiscal pressure reduces labor supply or when jobs are destroyed by hikes in the minimum wage. If we are to clarify the choices facing governments and taxpayers, it is imperative to know precisely the impact of taxation, social transfers, and the minimum wage on labor force participation.

We will see in this chapter that, through theoretical analysis and empirical research, economic analysis is able to document in detail, and shed light on, the arbitrage between equity and efficiency.

Following a review of the main characteristics of the tax regimes in the OECD countries, we start by analyzing, from a theoretical perspective, the impact of taxation and transfers on wages, employment, hours worked, and unemployment. The analysis will reveal that it is not always taxpayers who bear the cost of taxation. It will also allow us to show that variations in the marginal rates and the average rates at which tax is levied can have opposing effects on the volume of hours worked.

We then examine how empirical research evaluates the impact of taxes and transfers. Such research generally finds that fiscal pressure has a negative, and significant, impact on the volume of hours worked, thereby confirming that there is a cost—in the form of reduced labor force participation—incurred by compressing the spread of inequality through income redistribution linked to increased fiscal pressure. Hence the information yielded by this research has value when it comes to "optimizing" fiscal regimes, in other words, minimizing their impact on the volume of labor force participation while ensuring a given flow of revenue into the fisc.

Since heightened fiscal pressure reduces labor force participation, one might suppose that a legal minimum wage, in addition to taxes, would efficiently redistribute income toward unskilled workers. And a legal minimum wage is indeed set by the government in a great many countries. But the topic still generates much debate.

The main point at issue is the impact of minimum wage on employment. Employers generally maintain that it destroys jobs by pushing up the cost of labor. Trade unions assert on the contrary that a minimum wage increases employment by incentivizing workers to start participating again in the labor market. As we will see, economic theory suggests that each of these viewpoints holds a degree of truth. Empirical research confirms the dichotomy, showing that minimum wage can have a positive impact on employment for certain categories of worker when it is not "too" high. But this impact can turn negative if it is. The upshot is that minimum wage has an ambiguous effect on income inequality and poverty: boosting it increases the income of those earning wages at this level who keep their jobs, but doing so lowers the income of persons who lose their jobs as a result of the boost.

Another important issue is the usefulness of minimum wage as a policy lever for achieving income redistribution. The fact that it may have a positive impact on employment among certain categories of worker does not automatically make it an adequate instrument for income redistribution. In fact, economic analysis suggests that it is generally more efficient to use the tax regime rather than minimum wage to redistribute income.

Section 1 of this chapter deals with taxes and transfers. Section 2 is dedicated to the minimum wage.

1 TAXATION AND TRANSFERS

This section begins with a presentation of the main features of taxation in OECD countries. As we will see, fiscal regimes vary greatly from one country to another. Not only is the degree of fiscal pressure variable, but the progressivity of taxation and the structure of mandatory tax contributions are also very heterogeneous. We then proceed to study the impact of taxes and transfers on hours worked, employment, and unemployment, using the models of perfect and imperfect competition presented in previous chapters. The last part of this section reviews the empirical research dedicated to the impact of taxation on the labor market. This research shows that taxes do have a significant influence on labor force participation and that they likely explain an important part of the difference in hours worked per person across countries.

1.1 The Main Features of Taxes in OECD Countries

The structure of mandatory contributions and the extent of redistribution differ considerably from country to country. The "tax wedge" is a synthetic indicator which proves useful in assessing the degree of fiscal pressure in many circumstances. It needs to be complemented by measurements of the degree to which taxes are progressive, if we are to have an adequate overview of the characteristics of the fiscal system.

1.1.1 Mandatory Contributions

Mandatory contributions are all payments made by all actors to public authorities with no direct compensation in return. They comprise taxes in the strict sense, as well as social security contributions. Taxes are collected by the government and by local public authorities. Social security contributions are collected by the government, or by dedicated organizations, for the purpose of insuring persons against certain contingencies like illness, disability, old age, childbirth, and unemployment, which temporarily or permanently prevent them from working. Among mandatory contributions, a distinction is normally made between contributions paid by the employer and ones paid by the employee. In reality this distinction has little meaning because in either case mandatory contributions are entirely deducted from the value added which production creates. For employees and employers, the relevant magnitude is the difference between the value added and the *total* amount of mandatory contributions. Out of this difference they must compensate themselves and pay their remaining taxes. Figure 12.1 gives an idea of the system of mandatory contributions in several OECD countries.

The bottom section of bars in this figure shows the value of revenue from personal income tax in 2010 as a percentage of GDP, assessed on income from labor and capital. Personal income tax is high in Denmark and other Northern European countries but lower in France, Japan, and Eastern European countries, while in Germany, the United Kingdom, and the United States it is close to the OECD average. Social security contributions (middle section of bars in gray), on the other hand, constitute a fault line between what we may schematically see as two blocs. In the first, comprising Western Europe and Sweden, social security contributions come to around

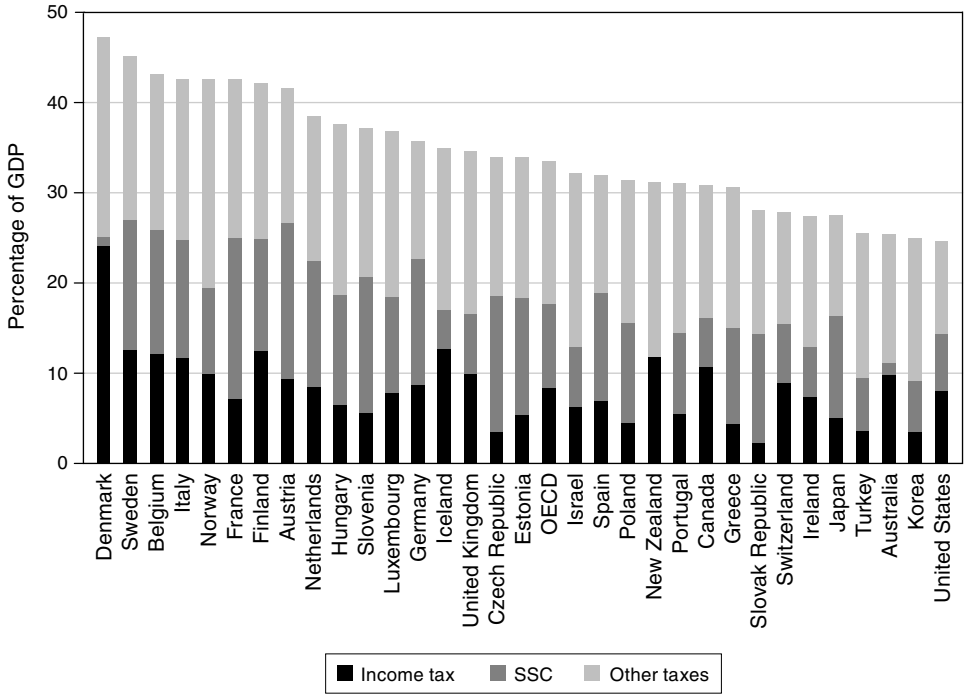


FIGURE 12.1
Tax revenues expressed as a percentage of GDP at market prices in 2010.

Note: SSC = Social security contributions; comprises also taxes on payroll and workforce when they apply. OECD refers to the nonweighted average of percentages among OECD countries.

Source: OECD Revenue Statistics.

15% of GDP, while in the second, comprising Australia, New Zealand, the United Kingdom, the United States, but also Korea and Denmark, social security contributions are less than 6% of GDP. Other taxes (this means principally indirect taxes, represented by the top section of each bar) are not insignificant either, running from 10% in the United States to more than 20% in France and the United Kingdom. The total height of the bars in figure 12.1 gives total tax revenue. This total, expressed as a percentage of GDP, is also called the "rate of mandatory contributions." By this criterion, we see that European countries have high tax pressure.

1.1.2 Social Benefits

The distinction suggested in figure 12.1 between an Anglophone model and a European one with respect to tax pressure is often mentioned in the literature. But it has to be set in perspective by taking into account the extent of social security benefits. Social benefits are all transfers received by households and intended to provide for needs arising from certain events or circumstances related to health (sickness, disability); the situation of the labor market (unemployment); demography (retirement, education,

or family circumstances); or other risks in general (living is a risky business). These benefits for the most part assume the profile of an insurance system financed by mandatory social contributions. To get them, one has to have paid in for a defined period. Unemployment insurance, retirement and disability pensions, as well as allowances for days lost to illness enter into this category. Most health care, especially in Europe, is often financed by social security contributions. These benefits represent the bulk of social spending. Social benefits also comprise allowances providing social assistance on a means-tested basis, which do not require prior payments into a specific fund and are generally financed by taxes other than social security contributions. Since they target individuals most in need, these allowances often play a key role in reducing poverty and inequality, along with the direct tax system. In France family allowance, housing allowance, and minimum guaranteed income fall into this category. In the United States the Medicare and Medicaid programs, which cover the health care costs of the elderly and the most disadvantaged, are examples (see Immervoll, 2010, for a description of the minimal levels of social assistance in the industrialized countries). Social benefits or tax credits can also top up earned income for low-wage workers: examples include the earned income tax credit (EITC) in the United States or the working family tax credit (WFTC) in the United Kingdom. In many countries there are also unemployment and disability assistance schemes to help those who have not contributed enough to be eligible for insurance-based benefits.

Spending

Overall, the European countries deliver measurably higher social benefits than other OECD countries. Figure 12.2 shows how much countries spent on social programs in 2009. While the OECD average is about 23% of GDP, eight European countries are close or above 30%, and Anglo phone countries usually lie below the average around 20% (except the United Kingdom), which means that the *net* rates of mandatory contributions to social security present less divergence between European countries and Anglophone countries than the gross rates (see Bourguignon, 2001). In other words, a large part of the gap in the rates of mandatory contributions in the two models is explained by the *different coverage* provided by the various social insurance systems. For instance, as shown in figures 12.1 and 12.2, France and Sweden feature both high social contribution rates and large social spending (representing about 30% of GDP), while the United States and Canada show an opposite pattern. In Denmark social policies are financed mainly through general taxation rather than social contributions, but this is an exception. The respective roles of the *public* sector and the *private* sector (through either mandatory or voluntary schemes) are not constant from one country to another.

Altogether, income support for the working-age population, based either on insurance or on assistance schemes, represents only about a quarter of total spending on public and mandatory private social programs. Figure 12.2 shows that old-age pensions and health care make up about two thirds of social expenditure in the OECD countries and have been the main drivers of social expenditure over the last 30 years. Still, benefits play a significant role in the redistributive system. Means-tested minimum-income schemes are one of the main tools to alleviate poverty, even if they only represent a small fraction (about 6% including housing support and various supplements) of total social spending.

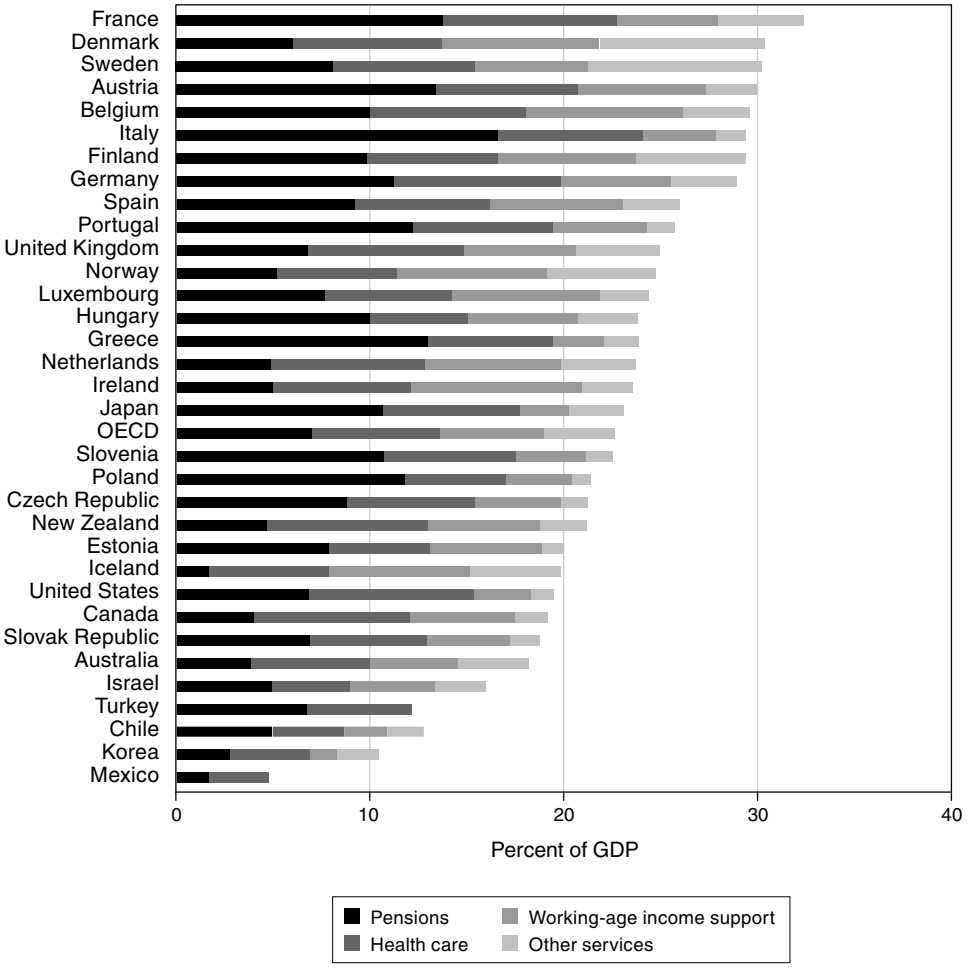


Figure 12.2
Public and mandatory private social spending as a percentage of GDP in the OECD, 2009.

Note: OECD refers to the nonweighted average of percentages among OECD countries.

Source: OECD Social Expenditure database.

Minimum Income Schemes and Assistance for the Poorest Persons

The benefits described above are often not sufficient to bring income above the poverty threshold. When comparing benefit generosity across countries, a useful approach is to look at benefit levels relative to commonly used poverty thresholds (50% or 60% of the median income). Figure 12.3 shows that in a large majority of OECD countries, benefits of last resort can be significantly lower than the relative poverty lines, and other income sources are needed everywhere to obviate substantial poverty risks. Minimum income benefits are usually larger for households with children, and the net minimum income of single persons often remains below the poverty thresholds. Assuming that beneficiaries also claim housing support on top of basic social assistance, the Nordic countries as well as Ireland, the Netherlands, the United Kingdom, and, outside Europe, Japan, reach a

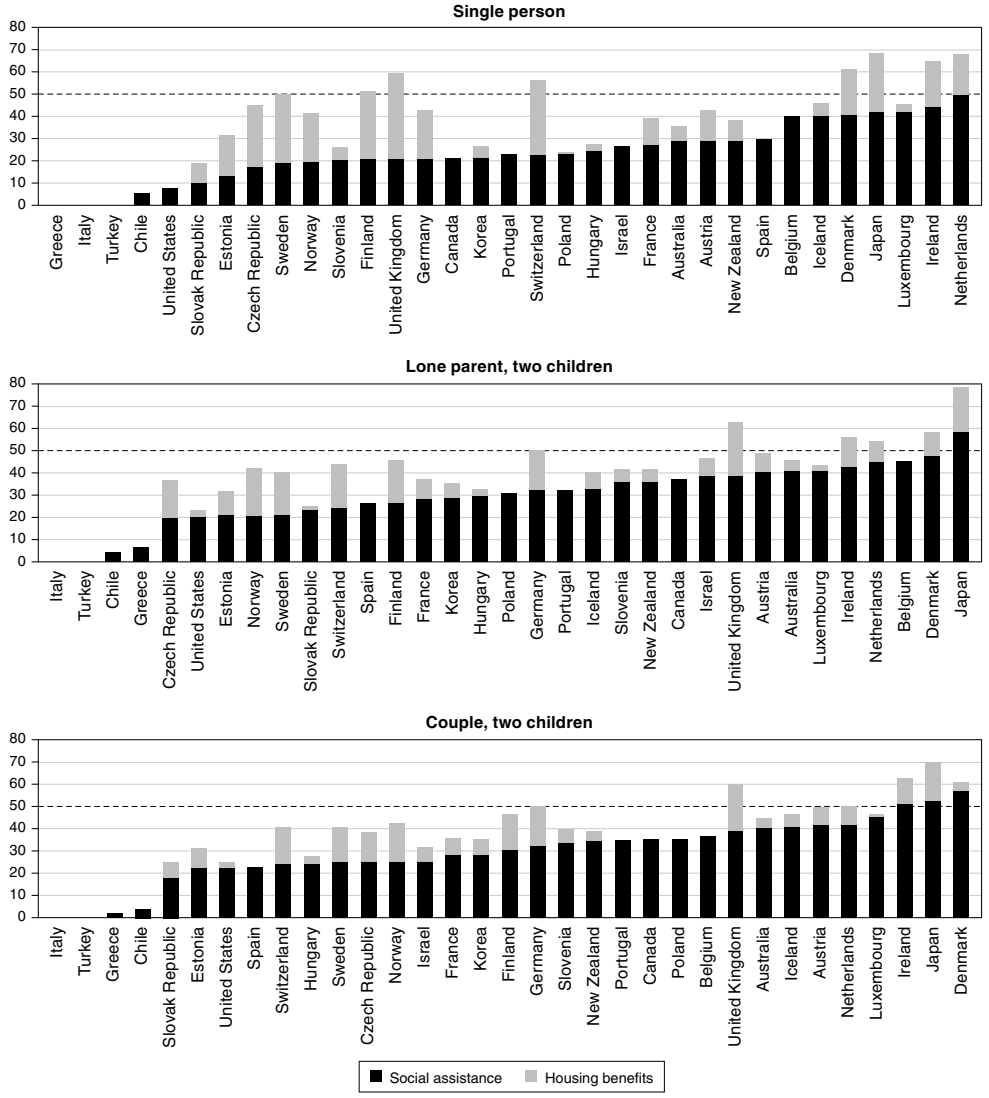


FIGURE 12.3

Net income levels provided by cash minimum-income benefits, as a percentage of median household incomes, 2011.

Notes: Median incomes are for a year around 2011 and are equivalized, i.e., adjusted for the size of households (using the "square root of household size" scale). Results take into account all relevant cash benefits (social assistance, family benefits, housing-related cash support as indicated). U.S. results also include the value of food stamps. Income levels take into account all cash benefit entitlements of a family with a working-age head, no other income sources, and no entitlements to primary benefits such as unemployment insurance. They are net of any income taxes and social contributions. Where benefit rules are not determined on a national level but vary by region or municipality, results refer to a "typical" case (e.g., Michigan in the United States, the capital city in some other countries). Calculations for families with children assume two children aged 4 and 6 and neither child care benefits nor child care costs are considered. The "housing benefits" indicates the range of benefit levels in countries where they depend on actual housing expenditure. Housing benefits represent cash benefits for someone in privately rented accommodation with rent plus other charges amounting to 20% of average gross full-time wages. There are no general social assistance schemes in Greece, Italy, and Turkey.

Source: OECD, Tax-Benefit Models (www.oecd.org/els/social/workincentives) Income Adequacy—Reliant on minimum income benefits, 2011.

level of support close to or above the poverty line. In comparison, the level of support in the United States, Southern European, and Eastern European countries is well below the poverty lines. Of course, the main limit to the generosity of social assistance is that the higher the net income flowing from such support, the greater the risk of creating disincentives to work (by inducing high effective taxation on earned income; see below) in the absence of a tax credit scheme or in-work benefits.

Minimum income benefits paid to inactive persons (that is, nonparticipants in the labor force) are not the only form of redistribution towards the poorest groups. In most countries, social assistance (e.g., in the form of means-tested in-work benefits), family benefits, or housing allowances are also paid to poor working families. Again this support tends to be larger for households with children. Figure 12.4 represents a simulation of what a household earning 50% of the average wage (an amount considered low earnings) would get in terms of benefits and would pay in taxes, as a percentage of the national equivalized (i.e., adjusted for the size of households) median disposable income among households. The dotted line, set at 50% of the national equivalized median disposable income, represents the relative poverty line. Since single persons live alone and have no children, in most countries they turn out to have income above the poverty line, even though they earn low wages in this simulation. They also pay low taxes (net of any benefits received) but still pay more taxes than they receive in terms of benefits (no net benefits). In comparison, lone working parents and couples with one earner and two children derive a substantial share of their disposable income from benefits (net of taxes). Anglophone countries, European Nordic countries (notably for lone parents), Japan, and Ireland provide substantial help to low-earning families, as opposed to Southern and some Eastern European countries.

Participation in Minimum-Income and Unemployment Assistance Schemes

About 4% to 5% of the working-age population get assistance benefits in one form or another in the OECD countries. But participation in last-resort benefits varies greatly across countries. Figure 12.5 shows the share of the working-age population receiving either unemployment assistance (when it is not the primary form of unemployment compensation) or minimum-income support (general social assistance schemes and schemes targeted at single parents). If we include food stamps, which represent the main form of support for poor households, the United States appears to support the largest share of the working-age population in 2010, just ahead of Ireland. In Italy and Chile, in the absence of national minimum-income schemes, there is no support of last resort at all. The interpretation of differences in participation at one point in time is difficult, though. For one thing, economic conditions might explain part of the variation: even though assistance benefits are far less sensitive to the business cycle than primary unemployment insurance benefits, they still react with some delay to changes in employment. Second, participation in one specific type of benefit is influenced by the overall architecture of social protection systems in each country. For instance, in some countries such as the Netherlands, but also the United Kingdom, participation in disability benefits bulks large (above 8% of the working-age population), and these benefits have become in some cases substitute sources of long-term support, replacing general social assistance (actually in the United States, participation in disability is also high; see Autor, 2011). In some countries, the duration of unemployment benefits is short, or this type of insurance benefit may not even exist (Mexico), and social assistance is the only remaining out-of-work support for a large share of the unemployed.

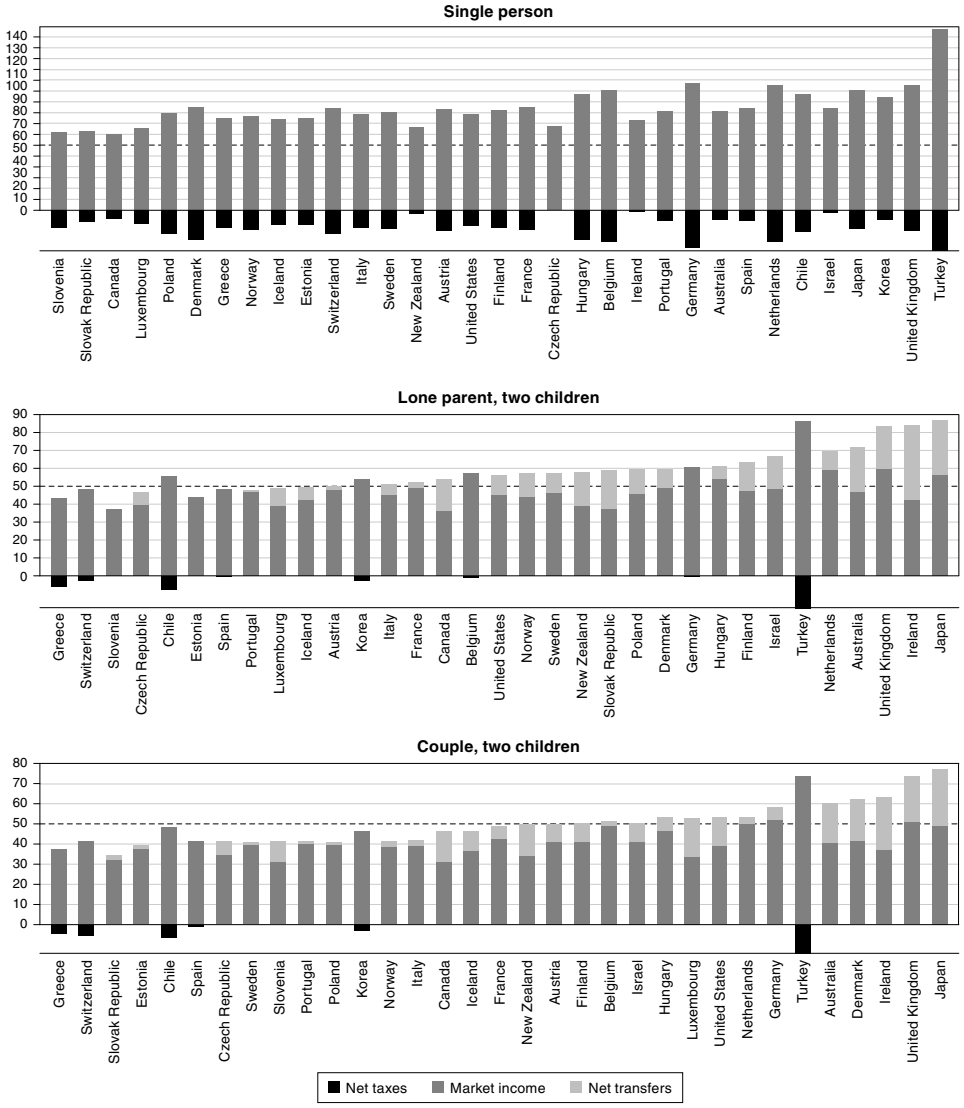


FIGURE 12.4

Income levels provided by full-time, low-wage employment as a percentage of median household incomes.

Note: Low-wage employment is taken at 50% of the *average* wage. Median net household incomes are from a survey in or close to 2011 (based on an equivalized basis with the equivalence scale being the square root of the household size) and take into account all relevant cash benefits (social assistance, family benefits, housing-related cash support). Income levels are net of any income taxes and social contributions and account for all cash benefit entitlements of a family with a working-age head employed full-time earning 50% of the average wage. Net taxes are income taxes and social security contributions net of cash benefits. Net transfers correspond to negative net taxes. Where benefit rules are not determined on a national level but vary by region or municipality, results refer to a "typical" case (e.g., Michigan in the United States, the capital in some other countries). Calculations for families with children assume two children aged 4 and 6 and neither child care benefits nor child care costs are considered. Cash housing assistance represents cash benefits for someone in privately rented accommodation with rent plus other charges amounting to 20% of average gross full-time wage.

Source: OECD, Tax-Benefit Models (www.oecd.org/els/social/workincentives), Income Adequacy—Reliant on minimum income benefits, 2011.

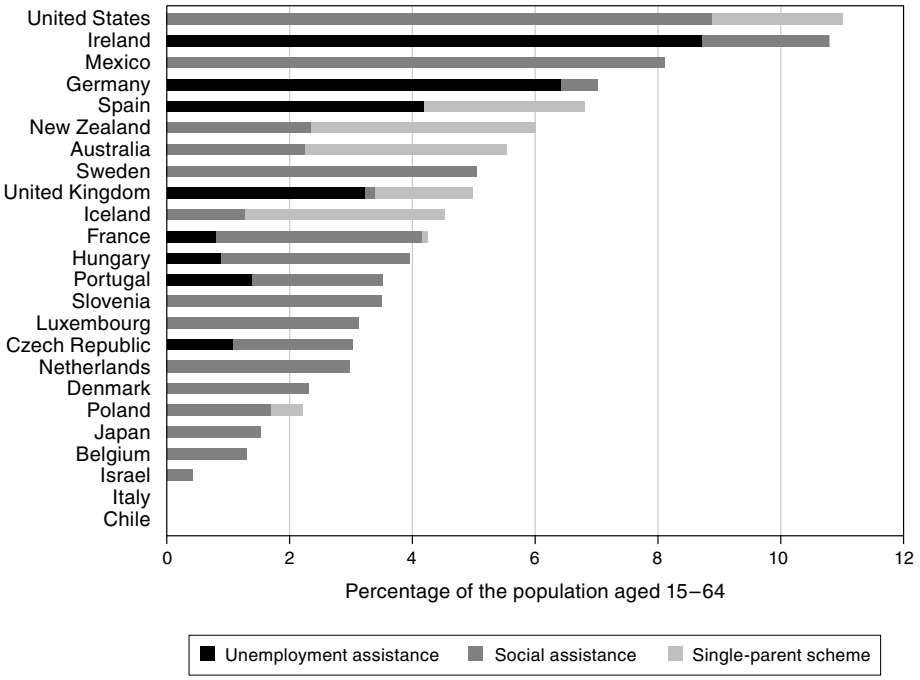


FIGURE 12.5
Participation in out-of-work assistance benefits, excluding primary unemployment benefits, in 2010.

Note: This figure excludes unemployment insurance beneficiaries and unemployment assistance benefits for Australia and New Zealand, where they are the primary form of out-of-work assistance. In the United States, the food stamps are included. In the United Kingdom, the beneficiaries of the Job Seeker Allowance can be means-tested or not. However, the administration does not provide this breakdown and all beneficiaries are included in the chart. In Italy and Spain, social assistance is provided by local administrations (provinces or municipalities, not included).

Source: OECD (2013).

1.1.3 THE TAX WEDGE

The gap between the cost of labor and the purchasing power of wages is usually gauged by the *tax wedge*. Let W and P_f respectively be the nominal wage received by an employee and the producer price index. If we denote by t_f the average rate of mandatory deductions from wages borne by firms, the real labor cost for the employer is written:

$$w_f = \frac{W(1 + t_f)}{P_f}$$

Let us again denote by t_c and t_e respectively the average rate of indirect taxes on consumption and the average rate at which earned income is taxed, net of benefits received—approximate indicators of these two magnitudes appear in the third and first lines respectively of figure 12.1—and let P_c represent the consumer price index exclusive of consumption taxes: the *purchasing power* of an employee takes the form:

$$w_e = \frac{W(1 - t_e)}{P_c(1 + t_c)}$$

Eliminating the nominal wage W between the expressions of w_e and w_f, we get:

$$w_f = \tau w_e \quad \text{with} \quad \tau = \frac{(1+t_c)(1+t_f)}{(1-t_e)}\left(\frac{P_c}{P_f}\right) \tag{12.1}$$

The term τ defines the *wedge*; it measures the ratio between the cost of labor borne by the employer and the purchasing power of wages. The wedge has two components. First, the ratio (P_c/P_f), which is influenced by the price of imports, because P_c comprises import prices, whereas the producer price index only comprises prices of domestic goods (which can however be indirectly influenced by import prices). The ratio (P_c/P_f) is a relatively volatile component of the wedge, especially because of exchange rate variations. The second component of the wedge is the tax wedge, which hinges on the tax rates t_c, t_e, and t_f. Henceforth, we will focus only on the tax wedge by setting the ratio (P_c/P_f) equal to 1.

Figure 12.6 displays the tax wedge in the mid-2000s in some OECD countries. Since personal income tax, benefits, and social security contributions often vary with the level of earned income, but also with the family situation, it is necessary to consider typical cases to compare wedges across countries. In this figure, rates apply to singles and couples, with and without children, and paid 67%, 100%, or 167% of the average wage. Including taxes on consumption is typically difficult in this type of comparison because these taxes depend not on income levels and family composition but on the level and type of consumption instead. Figure 12.6 uses estimates of these taxes based on household budget surveys (see also McDaniel, 2011, for alternative estimates, but without a breakdown by typical family cases). We see that wedges are large, especially in some European countries such as France, Belgium, and Austria, in comparison to Australia, the United States, and Mexico. For instance, in Belgium the cost of labor represents 2.8 times the purchasing power of wages for a single person without children paid at 167% of the average wage, compared to 1.6 times in the United States. In all countries, direct contributions represent the largest share of the tax wedge.

Direct taxes make up most of the tax wedge for a majority of households, except possibly for the poorest ones. Figure 12.7 decomposes further the taxes on earned incomes in the case of singles without children for a more recent year, focusing this time on these direct taxes only (thus excluding taxes on consumption and cash benefits). Within Europe striking differences emerge. Social security contributions are highest in France, while in Belgium and Germany income tax plays a somewhat larger role. Income tax plays a predominant role in a number of Northern Europe countries, notably Denmark and Norway. Table 12.1 shows how direct taxes evolved between 1979 and 2009 for workers in the manufacturing sector, that is to say, during a period of mounting unemployment in Europe. This indicator followed diverging paths. It shrank in the United Kingdom, Sweden, and the Netherlands, remained stable in the United States and Spain, and grew in Germany and Canada until the late 1990s and in Japan until the late 2000s.

1.1.4 The Progressivity of Taxes

When dealing with taxation, it is important to distinguish the *average* tax rate from the *marginal* tax rate. The average rate is an indicator of the global volume of taxation, while the marginal rate—which measures the increase in taxation on each extra unit of income

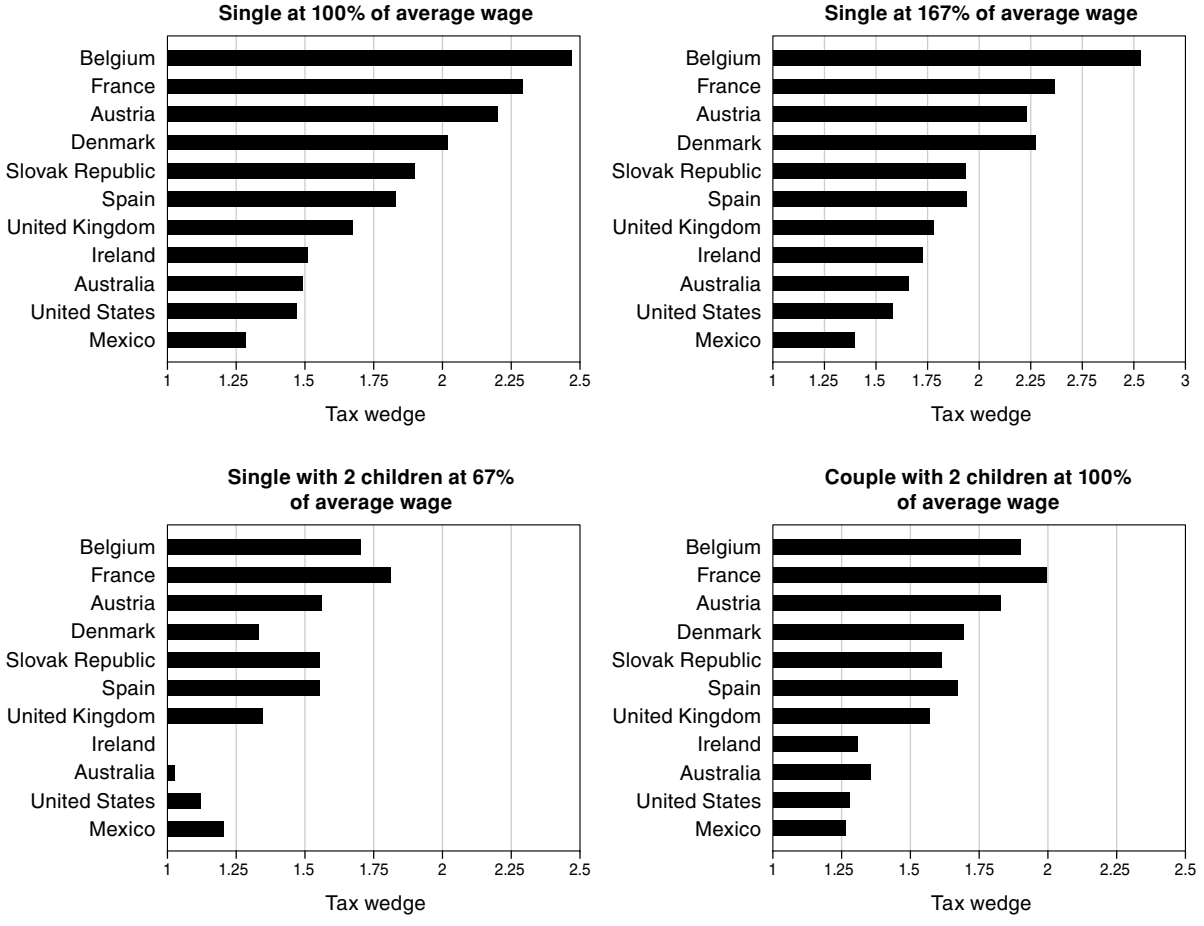


FIGURE 12.6

Tax wedge between labor cost and take-home pay for several family types around 2005. The tax wedge is defined by equation (12.1).

Note: SSC = Social security contributions include other forms of payroll taxes where applicable. Income tax is net of tax credits and cash benefits to which each specific family-type may be entitled.

Source: OECD (2008, tables S.7 and S.8, p. 32).

or expenditure—is an indicator of the *progressivity* of taxes. Most systems of mandatory contribution show a certain progressivity, in which case the marginal rate exceeds the average rate.

Marginal Rates and Average Rates

To study the consequences of progressivity, we must first define a system of mandatory contributions that will allow us to distinguish marginal rates from average ones. We will designate by w the real gross wage received by the worker and will assume, in order to simplify the exposition, that contributions are indexed to it. The purchasing power w_e of

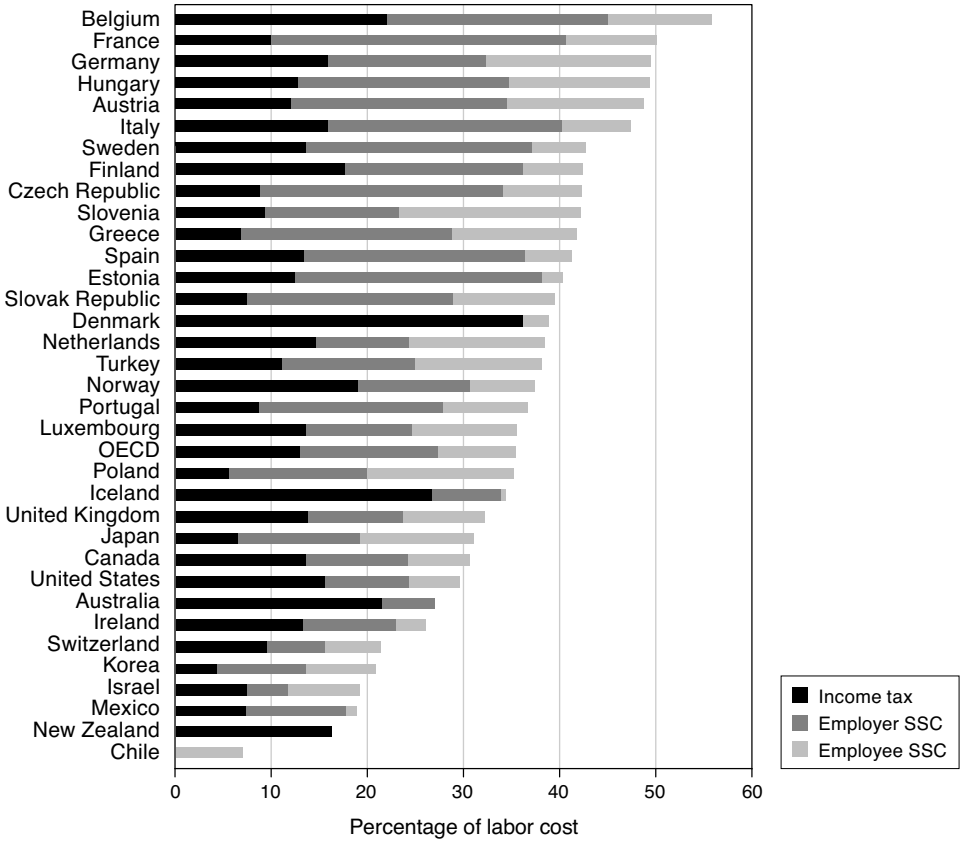


FIGURE 12.7

Direct taxes on earned income (income tax plus employees' and employers' contributions) for a single person with no children paid at 100% of the average wage.

Note: SSC = Social security contributions include other forms of payroll taxes where applicable; cash benefits are excluded. OECD refers to the nonweighted average of percentages among OECD countries.

Source: OECD Taxing Wages database.

wages and the labor cost w_f for the firm can then be written in the following manner:

$$w_e = w - T_e(w) \quad \text{and} \quad w_f = w + T_f(w) \tag{12.2}$$

Function T_e represents the sum of the direct and indirect taxes on earned income paid by the worker, less any cash benefits received, and function T_f stands for all the payroll taxes paid by the employer. In reality, these two functions depend on many parameters characterizing taxation in each country, including different tax brackets and the marginal tax rates that apply to each of them, thresholds that trigger tax relief, and ceilings on certain contributions (see Malcomson and Sator, 1987). To simplify the notation, we have not included these parameters in writing the functions T_e and T_f. It is the extent of the variation in the contributions T_e and T_f when income rises that

TABLE 12.1
Direct taxes on earned income (income tax plus employees' and employers' contributions) over time in the manufacturing sector in some OECD countries.

Country	1979	1989	1999	2009
Canada	23.2	27.2	31.1	30.2
Germany	40.8	45.5	51.9	49.7
Japan	16.7	20.4	24.0	29.2
Netherlands	48.0	47.0	44.3	39.8
Spain	36.4	35.9	37.5	37.4
Sweden	50.7	52.7	50.5	42.8
United Kingdom	36.1	34.2	30.8	29.8
United States	31.9	31.5	31.1	29.8

Note: Percentages of labor cost for single persons without children, with earnings of the average production worker. Year 2009 estimate based on the tax wedge for the average worker in the whole economy.

Source: OECD (2001, table 1.4, p. 341) and OECD Taxing Wages database for 2009.

allows us to pinpoint how progressive a system of mandatory contributions is. This is why the respective elasticities η_e and η_f of w_e and w_f with respect to w play an essential part in measuring this progressivity. Differentiating relations (12.2), we find that they can be written:

$$\eta_e = \frac{1 - T'_e}{1 - (T_e/w)} \quad \text{and} \quad \eta_f = \frac{1 + T'_f}{1 + (T_f/w)} \tag{12.3}$$

In these relations, T'_e and T'_f designate respectively the derivatives of functions T_e and T_f with respect to w. These quantities represent the *marginal rates* of taxation of the employee and the firm, while the quantities (T_e/w) and (T_f/w) represent the *average rates*. The gap between the average rates and the marginal rates characterizes the degree to which taxation is progressive or regressive. These notions can be understood clearly by focusing on the elasticities η_e and η_f (for more detail on this subject, see Musgrave and Musgrave, 1989):

- If $\eta_e < 1$, a rise of 1% in the wage corresponds to a rise of *less* than 1% in the purchasing power of this wage. This property tells us that the income tax (or the consumption tax) is progressive. When this is the case, the marginal rate T'_e is higher than the average rate (T_e/w). Elasticity η_e is often called the "coefficient of residual income progression."
- If $\eta_f > 1$, a rise of 1% in the real wage leads to a rise of *more* than 1% in the cost of labor for the firm. This property tells us that the payroll tax borne by firms is progressive. When this is the case, the marginal rate T'_f is higher than the average rate (T_f/w). When η_f is less than unity, this system is *regressive*.

- If $\eta_e = 1$, the income tax system is said to be *proportional*. The marginal rate T'_e is then equal to the average rate (T_e/w). Likewise, if $\eta_f = 1$, the payroll tax borne by firms is said to be *proportional*. The marginal rate T'_f is then equal to the average rate (T_f/w).

Progressivity in Some OECD Countries

Table 12.2 gives the values of the average rate, the marginal rate, and the coefficient η_e of residual income progression as they apply to taxation on the income of a single person with an income equivalent to 167% of that of an average worker in 2012 in some OECD countries. We see that tax progressivity is prevalent in all these countries. The countries of Northern Europe are distinguished by high marginal rates and high progressivity. France and the United Kingdom have marginal rates close to the average of the OECD countries. Anglophone countries, along with Japan, Korea, and Mexico, typically feature marginal rates lower than the average, and the gap between the average rate and the marginal rate is also relatively narrow there, which is a sign that they are less progressive.

As shown in figure 12.8, working-age households are net taxpayers on average, and households in the top 20% in the income distribution pay more taxes (including social contributions) and receive many fewer benefits than the average. This is particularly true

TABLE 12.2
Average rates and marginal rates for a single person with an income equivalent to 167% of that of an average worker in 2012.

Country	Average rate	Marginal rate	η_e
Germany	43.8	44.3	0.99
Poland	25.4	26.7	0.98
Korea	15.7	18.7	0.96
Japan	25.0	30.8	0.92
Mexico	14.0	22.9	0.90
Canada	26.7	35.4	0.88
United States	28.6	37.4	0.88
France	33.9	42.4	0.87
OECD average	30.5	39.9	0.86
Australia	29.1	39.5	0.85
United Kingdom	30.7	42.0	0.84
Spain	28.1	40.0	0.83
Netherlands	38.1	49.3	0.82
Italy	37.9	49.9	0.81
Denmark	45.1	56.1	0.80
Belgium	49.5	59.8	0.80
Sweden	35.2	56.6	0.67

Note: These rates include income tax and the social security contributions deducted from wages, less cash benefits.

Source: OECD Taxing Wages database and OECD (2013, table I.8, p. 90).

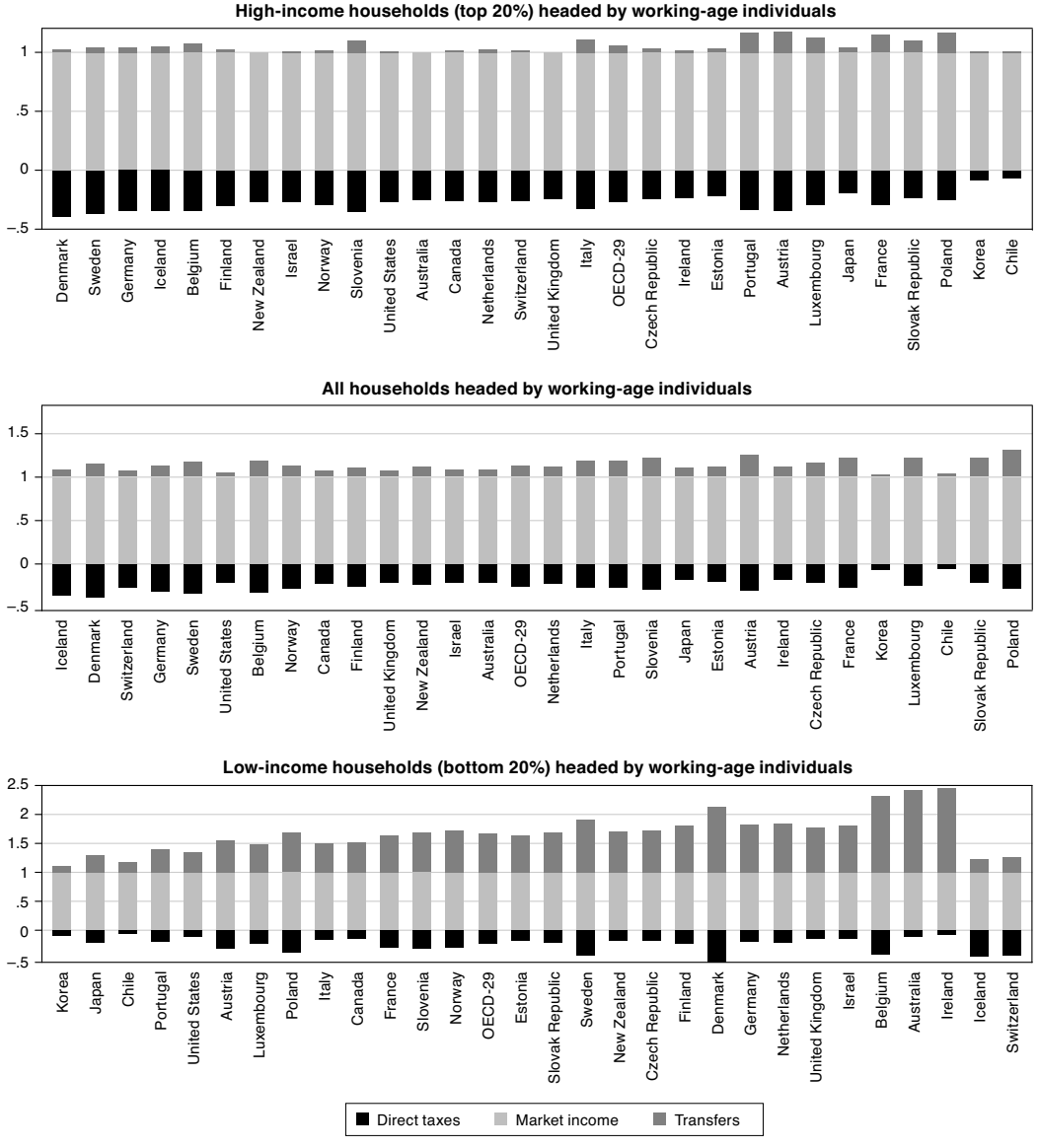


FIGURE 12.8

Overall amounts of direct taxes paid and benefits received in the mid-2000s by level of income as a proportion of market income (all types of household).

Notes: Countries are ranked by the impact of the redistribution system on household income, that is, by net taxes (taxes minus benefits). Direct taxes include personal income tax and social security contributions. Transfers include cash benefits only.

Source: OECD (2011, figure 7.1, p. 265).

in the Nordic countries. These taxes go towards financing other public expenditures, such as publicly provided services, social assistance, current transfers to the elderly and the unemployed, and one's own future pension entitlements. The poorest 20% of working-age households are net benefit recipients in almost all countries, with cash transfers adding up to around two thirds of market income on average. They reach much higher proportions in the Northern European countries but also in Australia and Ireland, where they almost double market income and are modest in comparison in the United States and Japan.

Marginal taxes can also be high at the very low end of the income distribution, if we consider not only income taxes and social security contributions but also the loss of benefits which are "taxed away" when people resume work or increase their working hours. For instance, in most countries, people lose minimum-income and unemployment payments when they start working. Similarly, when family or in-work benefits are paid to households where at least one adult is working, these benefits might be decreased or even suspended when earned income increases, as it may if the second adult in the family starts working or if the main wage earner works more hours. Remember that T_e is the sum of the taxes on earned income less any cash benefits received. Hence when benefits are reduced or suspended, the tax rate increases. Figure 12.9 simulates the marginal tax rate effectively supported by persons returning to work in a job paid 50% or 100% of the average wage in 2011. A high marginal tax rate indicates that transitions into work result in little or no gain in net incomes, and at 100% there is no financial advantage to resuming work. In half of the OECD countries, this rate is at or above 80% for couples with one earner and two children resuming work at 50% of the average wage, which is far more than the marginal tax rates paid by those earning 167% of the average wage displayed in table 12.2. In the Northern European countries the marginal tax rate effectively supported by persons returning to work in a low-paid job is close to 100%, while it is only 25% in the United States, due notably to the earned income tax credit (EITC; see section 1.3.1). Marginal tax rates at lower levels of earned income are usually higher for families with children because they typically receive more benefits when inactive than singles without children.

1.2 The Effect of Taxes on the Labor Market

In this section, we analyze the impact of taxes on labor market equilibrium. When taxes are changed, they can modify the behavior of individuals directly affected by the changes and also the equilibrium wage and the behavior of other agents not directly affected by them. To analyze this issue, we begin by reviewing some results concerning the incidence of taxes in a labor market with perfect competition, where individuals are either employed or idle. This framework is useful for understanding the impact of taxes on employment and hours of work. However it omits unemployment entirely. Accordingly, in a second stage, we use a model of imperfect competition, with search and matching, to analyze the impact of taxes on unemployment.

1.2.1 Perfect Competition

Let us consider a competitive economy where individuals produce a quantity y of output when they work. Firms pay a payroll tax denoted by T_f so that their profit per

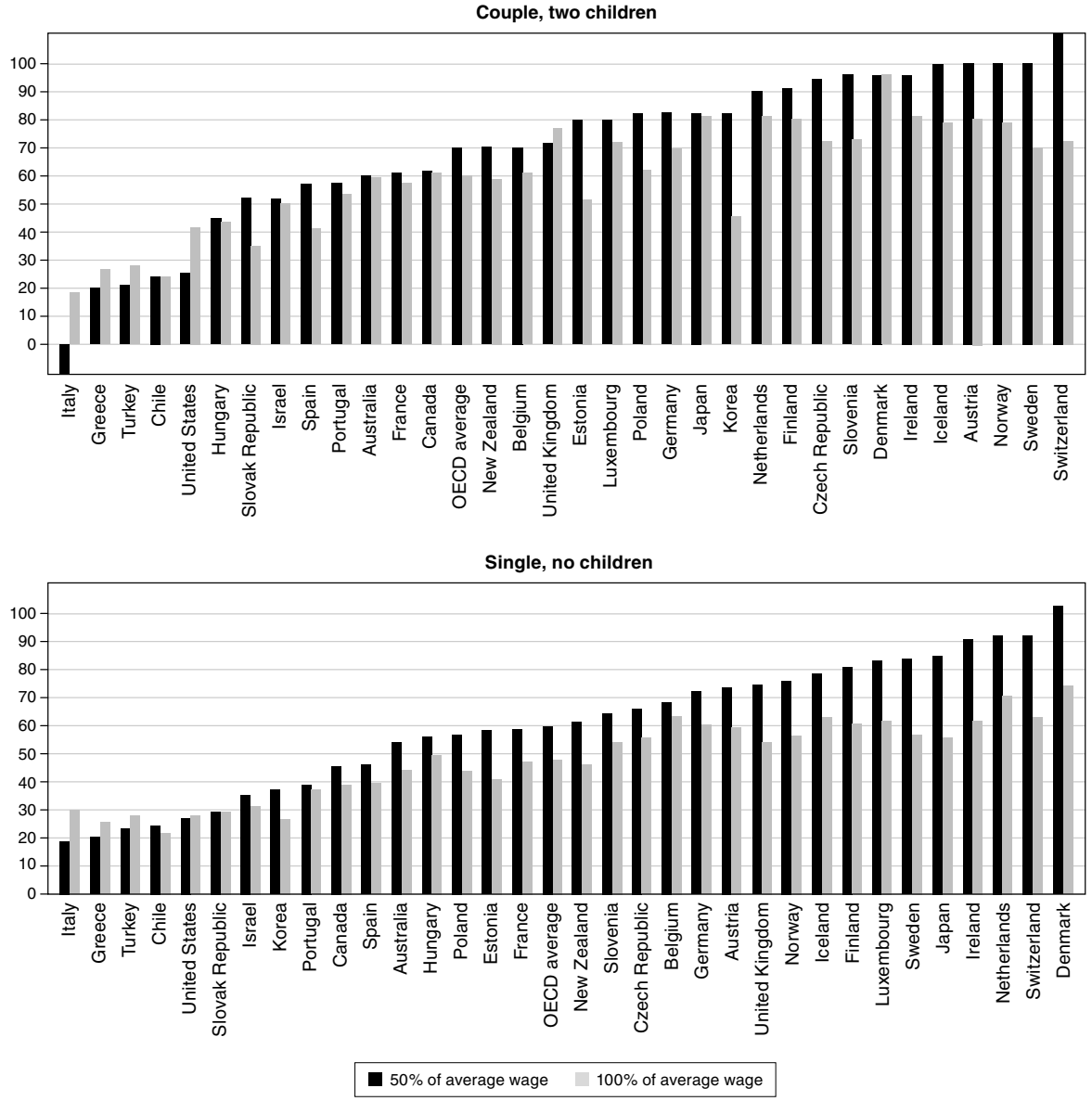


FIGURE 12.9

Marginal tax rate imposed following a shift from nonemployment to full-time employment at 50% and 100% of the average wage in 2011.

Note: It is assumed that no unemployment insurance benefits are available while out of work. In-work earnings are equal to 50%, or 100% of average wage (AW), and any benefits payable on moving into employment are assumed to be paid. Universally available unemployment or social assistance, housing, and family benefits are available if eligibility criteria are met. In the case of couples, the other spouse is assumed to be "inactive" with no earnings and no recent employment history. Where receipt of social assistance or other minimum-income benefits is subject to activity tests (such as active job search or being available for work), these requirements are assumed to be met. Children are aged 4 and 6 and neither child care benefits nor child care costs are considered.

Source: OECD, Tax-Benefit Models (www.oecd.org/els/social/workincentives) Income Adequacy—Reliant on minimum income benefits, 2013.

employee is equal to $y - w - T_f$ where w denotes the real gross wage. Workers pay taxes T_e so that their net wage, w_e, is equal to $w - T_e$.

At competitive equilibrium, the zero profit condition implies that $w = y - T_f$. Accordingly, the net wage is:

$$w_e = y - T_f - T_e$$

This equation shows two important properties.

First, the impact of taxes on the net wage is identical, whether it is the employer or the employee who is paying taxes to the fisc. This property, which was stressed in chapter 3, section 1.2, is very general.

Second, this equation shows that a $1 tax on labor induces a $1 decrease in the net wage. This result holds good in a framework where the marginal productivity of labor is constant, so that labor demand is infinitely elastic with respect to the wage.[1] It is worth noting that this framework is relevant when it comes to analyzing the long-run impact of taxes on earnings of homogeneous labor,[2] where the production function has constant returns to scale with respect to labor and capital, and where the interest rate is equal to the discount rate (see, for instance, Acemoglu, 2009, chapter 8). To bring this property out, let us denote the production function by $F(K, L)$, where K stands for capital and L for labor. Profit maximization implies that the marginal productivity of labor, y, is equal to $F_L(K, L)$. Constant returns allow us to write the marginal productivity as a function of the capital/labor ratio $k = K/L$, that is, $F_L(k, 1)$. The capital/labor ratio is itself determined by the equality between the marginal productivity of capital, $F_K(k, 1)$, and the cost of capital, $r + \delta$, where r stands for the interest rate and δ for the depreciation rate of capital. In the long run, the interest rate is equal to the discount rate ρ, which is an exogenous variable, reflecting the preferences of individuals. Thus, k is determined by the relation $F_K(k, 1) = \rho + \delta$, which shows that k depends only on the discount rate, on the depreciation rate of capital, and on the technology. Therefore, the marginal productivity of labor is independent of the quantity of labor L because firms adjust the stock of capital to keep the capital/labor ratio k constant when the quantity of labor changes.

It should be noted that empirical evaluations of the incidence of taxes presented in chapter 3, section 1.2, generally find that changes in taxes have a strong impact on net wages, which move in the direction opposite to that of the taxes. These findings suggest that using a model where a $1 tax on labor induces a $1 decrease in net wage can be an acceptable approximation. This is certainly one of the reasons why the theory of optimal taxation, stemming from the seminal paper of Mirrlees (1971), makes, in most cases, such an assumption (see Piketty and Saez, 2013).

In the context where net labor income is merely equal to the (constant) marginal productivity of labor minus taxes, the impact of taxes can be analyzed with the labor

[1] It has been shown, in chapter 3, section 1.2, that the impact of tax increases on the net wage depends on the elasticities of labor supply and labor demand.

[2] Kubick (2004) and Rothstein (2010) analyze the impact of taxes in a context where different groups of individuals, whose labor services are imperfectly substitutable, are taxed differently. In this context, the labor demand is imperfectly elastic and the tax incidence is different, as shown in chapter 3, section 1.2.

supply model. Therefore, as shown in chapter 1, increases in average taxes while keeping marginal tax constant increase labor supply at the intensive margin (hours of work of those who were already working at the time of the tax change) to the extent that leisure is a normal good, and decrease labor supply at the extensive margin (the choice to work or not to work). Increases in marginal tax, keeping average tax constant, decrease labor supply at both the intensive and the extensive margins. All in all, in empirically relevant cases, where the substitution effect is larger than the income effect, increases in proportional taxes induce drops in hours of work at the intensive margin (see, for instance, Mirrlees et al., 2010, chapter 3). Higher proportional taxes are also detrimental to labor supply at the extensive margin because they reduce the relative gains of market hours.

1.2.2 Imperfect Competition

Let us now analyze the impact of taxes in a situation where individuals who want to work can be unemployed because they fail to find a job. This will allow us to understand the impact of taxes on unemployment, and more broadly on employment and labor market participation when unemployment is taken into account. To do this, we introduce taxes into the search and matching model of chapter 9, section 3. We begin by writing the expected utilities of workers and the expected profits of firms. Then we analyze the influence of taxes on bargained wages and on equilibrium unemployment.

Value Functions

The expected utilities V_e and V_u of a person respectively employed and looking for work satisfy:

$$rV_e = w - T_e(w) + q(V_u - V_e) \tag{12.4}$$

$$rV_u = z + \theta m(\theta)(V_e - V_u) \tag{12.5}$$

In these equations, w stands for the gross wage, z is the flow of income of unemployed workers, θ is the labor market tightness, and q and $\theta m(\theta)$ designate respectively the (exogenous) job destruction rate and the exit rate from unemployment. $T_e(w)$ is the tax on labor income.

The expected profit Π_e from a filled job is written:

$$r\Pi_e = y - w - T_f(w) + q(\Pi_v - \Pi_e) \tag{12.6}$$

where $T_f(w)$ is the payroll tax and y is the flow of production of the job, while the expected profit from a vacant job always satisfies the relation:

$$r\Pi_v = -h + m(\theta)(\Pi_e - \Pi_v) \tag{12.7}$$

where h is the instantaneous cost of a job vacancy and $m(\theta)$ is the filling rate of vacant jobs.

Bargaining

The outcome of the bargaining corresponds to the solution of the generalized Nash problem described in chapter 9, section 3.3.1. It is written:

$$\max_{w} \gamma \ln(V_e - V_u) + (1-\gamma)\ln(\Pi_e - \Pi_v) \tag{12.8}$$

Let us recall that $\gamma \in [0,1]$ is a parameter representing the bargaining power of the worker. Relations (12.6) and (12.4) allow us to find the contributions of the players to the Nash criterion. They are written:

$$\Pi_e - \Pi_v = \frac{y - w - T_f(w) - r\Pi_v}{r+q} \quad \text{and} \quad V_e - V_u = \frac{w - T_e(w) - rV_u}{r+q} \tag{12.9}$$

These relations imply that the surplus, $S = V_e - V_u + \Pi_e - \Pi_v$, can be written:

$$S = \frac{y - T_e(w) - T_f(w) - r(V_u + \Pi_v)}{r+q} \tag{12.10}$$

The first-order condition of the problem (12.8) is found by setting to zero the derivative with respect to w. After some rearrangements of terms, it takes the following form:

$$V_e - V_u = \frac{\gamma}{\gamma + (1-\gamma)\phi} S \tag{12.11}$$

where the coefficient $\phi = \left[1 + T_f'(w)\right] / \left[1 - T_e'(w)\right]$ is an indicator of the global progressivity of taxes. A rise in ϕ corresponds to a system that is becoming globally more progressive, as for example when the progressivity of income tax is made steeper and/or the progressivity of payroll taxes is made steeper.

Henceforth, the consequence of steeper progressivity is analyzed by looking at the impact of changes in ϕ when the tax system is adapted to keep the receipts $T_f(w)$ and $T_e(w)$ constant. In this framework, equation (12.11) shows that more progressive taxes reduce the workers' share of the surplus. As progressivity becomes steeper, any wage rise procures a smaller marginal utility for workers and entails a higher marginal cost for the firm. For this reason, at partial equilibrium, where V_u is fixed, progressivity exerts a downward pressure on the negotiated wage, making any wage rise less attractive to workers and more costly for the firm (Lockwood and Manning, 1993). From this point of view, it is worth noting that the model of imperfect competition predicts, contrary to the model of perfect competition, that a \$1 increase in taxes does not necessarily induce a \$1 decrease in wages, when the marginal productivity of labor is constant (equal to y). The reason is that wages are below the marginal productivity of labor net of taxes (i.e., $y - T_f(w) - T_e(w)$) in the model of imperfect competition.

The Impact of Taxes on Labor Equilibrium

Let us now analyze the impact of taxes on labor market equilibrium. When the free entry condition $\Pi_v = 0$ is satisfied, expression (12.7) of the expected profit from a

vacant job again gives $\Pi_e = h/m(\theta)$. Using the definition of the expected gains of unemployed workers (12.5), the definition of the surplus (12.10), and the surplus sharing rule (12.11), we can compute the equilibrium value of the expected utility of an unemployed person V_u:

$$rV_u = z + \frac{\gamma h\theta}{(1-\gamma)\phi} \tag{12.12}$$

Using the definition of $V_e - V_u$ given by equation (12.9) and the definition of the surplus (12.10), we get:

$$\frac{w - T_e(w) - rV_u}{r+q} = \frac{\gamma}{\gamma + (1-\gamma)\phi} \frac{y - T_e(w) - T_f(w) - rV_u}{r+q}$$

With equation (12.12) we get:

$$w = \frac{\gamma\left[y - T_f(w)\right] + \gamma h\theta + (1-\gamma)\phi\left[z + T_e(w)\right]}{\gamma + (1-\gamma)\phi}$$

which implies, together with the free entry condition $\Pi_v = 0$:

$$\frac{h}{m(\theta)} = \frac{y - w - T_f(w)}{r+q} = \frac{(1-\gamma)\phi\left[y - T_f(w) - z - T_e(w)\right] - \gamma h\theta}{(r+q)\left[\gamma + (1-\gamma)\phi\right]}$$

or:

$$\frac{h}{m(\theta)} = \frac{(1-\gamma)\left[y - z - T_f(w) - T_e(w)\right]}{(r+q)\left(\frac{\gamma}{\phi} + 1 - \gamma\right) + \frac{\gamma}{\phi}\theta m(\theta)} \tag{12.13}$$

This equation defines the equilibrium market tightness. We revert to equation (9.22) from chapter 9, the model with no fiscal system, by setting $T_f(w) = T_e(w) = 0$ and $\phi = 1$. As shown in chapter 9, the implication is that labor market tightness increases with y and decreases with γ. Moreover, we know from chapter 9 that the equilibrium unemployment rate decreases with the labor market tightness. Therefore, this equation allows us to analyze the impact of taxes on unemployment.

First, it is worth noting that the model illustrates a classic result of the analysis of tax incidence, according to which a \$1 tax has the same impact on unemployment whether it is paid by the firm or by the employee.

It turns out that increases in the level of taxes, T_e and T_f (and thus in the tax wedge), keeping ϕ constant, increase unemployment. Such tax increases are purely equivalent to drops in productivity: they reduce the surplus of jobs. From this perspective, it is important to remark that it is the difference between the production net of tax of employees, $y - T_e - T_f$, and the income of the unemployed, z, that exerts an impact on unemployment. All taxes and transfers that decrease the difference between the production net of tax of employees and the income of unemployed workers, keeping ϕ constant, increase unemployment. This means that increases in taxes that also decrease the income of unemployed workers by the same amount should have no effect on unemployment.

Increases in tax progressivity, either through taxes or transfers, keeping T_e and T_f constant, reduce unemployment because more progressive taxation exerts a downward pressure on wages. This is the consequence of the fact that as progressivity becomes steeper, any wage rise procures a smaller marginal utility for workers and entails a higher marginal cost for the firm. This general mechanism implies that the downward pressure of progressivity on wages is not a mere consequence of the Nash bargaining solution. For instance, it can easily be verified that it also shows up in the competitive search model presented in chapter 5, section 4.2, where firms post wages to attract workers, and in collective wage bargaining models presented in chapter 7.

The Impact of Taxes on Labor Market Participation

To understand the impact of taxes on employment, we must take into account unemployment and labor market participation, or in other words, labor supply at the extensive margin. If we assume that decisions to participate in the labor market result from a trade-off between being an unemployed job seeker and not participating at all, any improvement in the welfare of the unemployed leads to an increase in participation. Let H be the cumulative distribution function of the expected utilities outside the labor market of the entire working-age population. All the individuals whose expected utility outside the labor market is less than the expected utility of an unemployed person V_u decide to participate in the labor market, which entails that the participation rate is equal to $H(V_u)$. As H is necessarily an increasing function, the participation rate increases with the expected utility of unemployed persons. In this framework, the employment rate is equal to $H(V_u)(1-u)$.

From definition (12.12) of the equilibrium value of the expected gains of unemployed workers, it can be deduced that increases in the level of taxes, T_e and T_f, keeping ϕ constant, decrease labor market participation because they decrease the labor market tightness. Accordingly, such tax increases raise unemployment, reduce labor market participation, and consequently exert a negative effect on employment.

The impact of tax progressivity on labor market participation, keeping T_e and T_f constant, is ambiguous. On one hand, steeper progressivity raises labor market tightness, which is favorable to unemployed workers and thus to labor market participation. On the other hand, equation (12.12) shows that steeper progressivity reduces the expected utility of unemployed persons by reducing the workers' share of the surplus generated by jobs. Accordingly, steeper progressivity, keeping T_e and T_f constant, has an ambiguous impact on employment, although it does reduce the unemployment rate. Actually, it can be shown that steeper productivity decreases participation if the bargaining power parameter γ is below the value that satisfies the Hosios condition (defined in chapter 9, section 4) and increases participation otherwise (Lehmann et al., 2013).

The Impact of Unemployment Benefits

In this model, benefits paid to working individuals directly alter $T_e(w)$. This would be the case for instance with in-work allowances, tax credits, and housing and family benefits. Some benefits are paid to individuals who do not work. The so-called unemployment benefits include unemployment insurance benefits, which are represented by z in the model, as well as any "inactive" benefits such as general social assistance or disability benefits, which are not conditional on looking for a job. The impact of increases in z in the above model is to raise wages (through higher V_u), decrease labor market

tightness, and increase unemployment. They also increase labor market participation $H(V_u)$ and thus would have an ambiguous impact on the employment rate $H(V_u)(1-u)$. In the next chapter we will see that in fact many more mechanisms are at stake in the analysis of unemployment insurance. As for inactive benefits, two cases may arise. In the first case, they are paid only to individuals who face real barriers to employment, for either health or social reasons, and cannot participate in the labor market. Then they improve the utilities outside the labor market for these persons but have no impact on participation and employment (assuming that the benefits in question do not alter working decisions made by their partners when the beneficiaries live in couples). In the second case, they are also paid to individuals who face no real barriers to employment, and in that case they would reduce labor market participation. Some individuals whose utility outside the labor market is close to V_u would then prefer to be inactive rather than unemployed with a duty to look for a job, if inactive benefits are generous enough. For that reason, the main challenge for inactive benefit programs is to ensure an appropriate targeting on those who cannot in fact perform any work.

1.2.3 TAKING STOCK

Let us summarize the results obtained to this point. The model of perfect competition has proved useful in analyzing the impact of taxes on labor supply at the extensive margin (to work or not) and at the intensive margin (choice of hours for those who do work). The model of imperfect competition has proved useful in shedding complementary light on the impact of taxes on labor market participation and on unemployment.

All in all, as summarized by table 12.3, the impact of tax changes on the choice of hours of work of employees, on labor market participation, and on unemployment varies according to whether the changes in question are made to average tax or to marginal tax.

Table 12.3, row 1, column 1, shows that increases in average tax, with marginal tax held constant, increase the hours of work of employees (i.e., choice at the intensive margin) due to the income effect when leisure is a normal good, which is the empirically relevant situation. Such increases in average tax reduce labor market participation (choice at the extensive margin) and increase unemployment because they reduce the gap between the income of employees on the one hand and the income of persons either unemployed or inactive on the other hand. Obviously, increases in nonlabor income, like the minimum income, reduce labor market participation.

Table 12.3, row 2, column 1, shows that increases in marginal tax, with average tax held constant, reduce the hours of work of employees (i.e., choice at the intensive

TABLE 12.3
The impact of taxes on hours of work, labor market participation, unemployment rate, and employment.

	(1) Intensive margin: hours of work	(2) Extensive margin: participation	(3) Unemployment rate	(4) Employment
Higher average tax (keeping marginal tax constant)	+	−	+	−
Higher marginal tax (keeping average tax constant)	−	?	−	?

margin) due to the substitution effect between consumption and leisure highlighted in the labor supply model. Such increases in marginal tax reduce the unemployment rate because they exert a downward pressure on wages. Their impact on labor market participation and employment is ambiguous.

1.3 What Empirical Studies Tell Us

Much empirical research has been dedicated to the impact of taxation on hours worked and employment. This research generally makes it possible to estimate the elasticity of labor supply at the intensive and/or the extensive margins. We supplied orders of magnitude for these elasticities in chapter 1, section 3.2.2: the Hicksian elasticities of labor supply at the extensive margin (employment) and the intensive margin (hours) are about 0.25 and 0.3 respectively, so that the elasticity of total hours of work is about 0.5.

Beyond these orders of magnitude, it should be stressed that empirical studies show that elasticities are heterogeneous across demographic groups (Blundell et al., 2013). In particular, it turns out that for low earners, responses are larger at the extensive margin than at the intensive margin (hours of work). Moreover, responses at both the intensive and extensive margins (and both substitution effects and income effects) are largest for women with school-age children and for those aged over 50 (Meghir and Phillips, 2010). Therefore, determining the impact of a specific change in taxes requires a specific evaluation. We begin by presenting the contribution of Eissa and Liebman (1996), which evaluates the impact of a tax credit for single-parent U.S. families with low labor income. Databases and programs allowing readers to replicate the main results of this contribution are available at www.labor-economics.org. This will allow us to analyze the impact of tax changes in some detail, showing in particular that they can have an impact of opposite sign on the labor supply of different groups of workers, depending on their level of labor earnings. We will also review the evaluation, based on microeconomic data, of permament or temporary assistance schemes for low-income households in other countries.

We then present empirical studies dedicated to the macroeconomic impact of taxation. In this realm, the large differences in tax systems across countries have laid the ground for many contributions that analyze whether these tax differences explain cross-country differences in unemployment, employment, and hours worked per adult.

1.3.1 Microeconomic Evidence: The Earned Income Tax Credit in the United States

Historically, the United States has chosen to provide a safety net for families with children. Since 1935, Aid to Families with Dependent Children has supplied cash welfare payments to needy single-parent families. Because the maximum level of benefit is received by families with no income, and because benefits are reduced almost dollar for dollar with additional earnings, the U.S. welfare system is predicted by labor supply theory to discourage the labor force participation and hours of work of single parents. In other words, the loss of benefits induces a high marginal tax rate when resuming work. Since 1975, tax credits have been introduced to reduce the marginal tax rate on the first dollars of labor income of single-parent families. These tax credits have been expanded by tax legislation on a number of occasions, including the widely publicized Reagan Tax Reform Act of 1986, which expanded the earned income tax credit (EITC). The EITC is

a refundable credit; therefore, any credit due in excess of tax liability is refunded to the taxpayer in the form of a tax refund check. Eissa and Liebman (1996) have studied the effects of the fiscal reform enacted in the United States in 1986 on labor force participation rates and hours worked. The expansion of the credit affects an easily identifiable group, single women with children, but is predicted to have no effect on another group, single women without children who are not eligible for the EITC.

Figure 12.10 displays the 1986 (before the reform) and 1988 (after the reform) earned income tax credits (in 1992 dollars) as functions of income. It can be seen that the amount of the tax credit on earned income rose dramatically. The well-being of a taxpayer who does not work is not changed by the reform because no earned income tax credit is available to a taxpayer with zero earnings. Thus, any taxpayer who preferred working before will still prefer working, and some taxpayers may find that the additional after-tax income from the EITC makes it worth entering the labor force. The impact of the EITC on the labor force participation of single taxpayers with children is therefore unambiguously positive.

Eissa and Liebman studied the impact of the Tax Reform Act of 1986 on single women only to avoid difficulties arising from intrafamilial decisions (see chapter 1, section 2.2.2). The control group therefore consisted of single childless women, while the treated group comprised single women with at least one child to care for. Eissa and Liebman (1996) then estimated the changes in the participation rate of each of these two groups. The data were those of the March Current Population Survey for the years 1985–1991 (excluding 1987, which was considered the year of the changeover). The treated and control groups comprised respectively 20,810 and 46,287 individuals. The impact of the reform is evaluated with the difference-in-differences estimator.

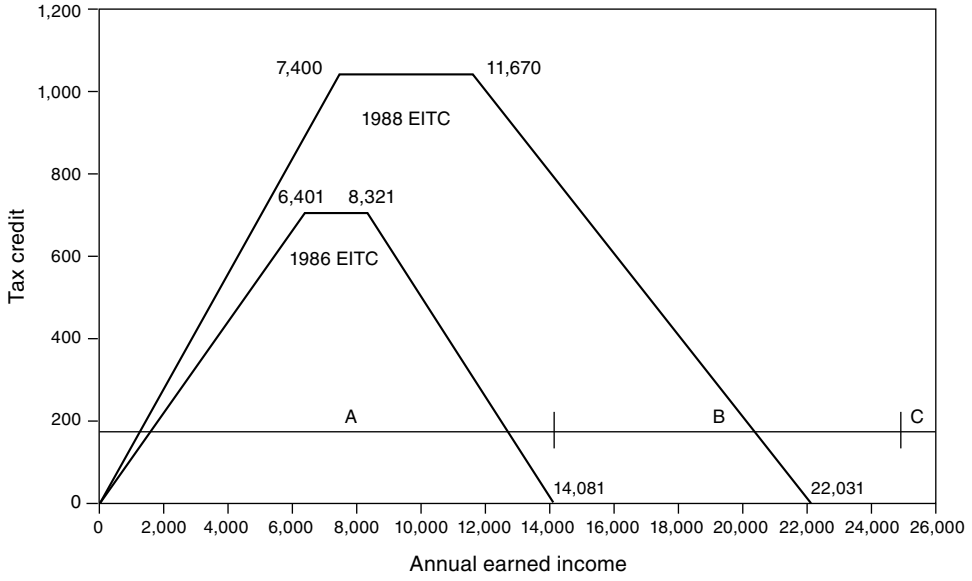


FIGURE 12.10
The expansion of the Earned Income Tax Credit between 1986 (before the reform) and 1988 (after the reform), 1992 dollars.

Source: Eissa and Liebman (1996, figure IV, p. 631).

The Difference-in-Differences Estimator

Let us take a population of individuals of size N, out of which a group $g = T$ (for "treated") of size N_T has been affected by a policy change. This group is composed of single women with children in Eissa and Liebman's setup. The control group $g = C$ of size N_C, composed of single women without children, has not been so affected. Suppose that we want to find out the effects of this policy change on a variable y (for example, participation in the labor market or hours of work). Let us denote by y_{it} the observed value of this variable on an individual i at date t, and let us use δ_{it} to designate the dummy variable, which equals 1 if the policy applies to individual i at date t, and 0 if it does not.

The *potential* outcomes for individual i are denoted by y_{it}^1 and y_{it}^0 for the treated and nontreated situations respectively (see chapter 14, section 3.1, for a detailed presentation of potential outcomes). They are specified as:

$$y_{it}^1 = \alpha_i + \beta + \varepsilon_{it}$$

$$y_{it}^0 = \beta + \varepsilon_{it}$$

where β is a constant, α_i is the effect of the treatment on individual i, and ε_{it} designates an unobserved component. It is possible, and often desirable, to introduce a vector of observable characteristics into this equation. Such a vector is omitted here for the sake of simplicity, but it will be introduced below. The observable outcome, y_{it}, is equal to y_{it}^1 for individuals who are treated at date t and to y_{it}^0 for those who are not treated. Thus:

$$y_{it} = \delta_{it} y_{it}^1 + (1 - \delta_{it}) y_{it}^0$$

which implies, using the two previous equations:

$$y_{it} = \beta + \alpha_i \delta_{it} + \varepsilon_{it} \tag{12.14}$$

The difference-in-differences estimator uses the "common trend" assumption (see also chapter 14, section 3.3), which posits that the difference in the average unobserved components ε_{it} across groups is constant over time, or more formally:

$$\mathbb{E}[\varepsilon_{it} | g, t] = m_g + m_t \tag{12.15}$$

where $g = C, T$ designates the group to which individual i belongs, m_t is a temporal effect common to all individuals, and m_g is a time-invariant, group-specific component, or in other words a group-fixed effect. The common trend assumption rules out the possibility of selection across groups based on unobserved individual specific effects, since it implies that changes in average unobserved individual effects are uniform across the treated and the nontreated groups:

$$\mathbb{E}[\varepsilon_{it} - \varepsilon_{i,t-1} | g = T] = \mathbb{E}[\varepsilon_{it} - \varepsilon_{i,t-1} | g = C] = \mathbb{E}[\varepsilon_{it} - \varepsilon_{i,t-1}] = m_t - m_{t-1}$$

Using the common trend assumption, we can compute the difference-in-differences estimator. Let us denote by Δ the difference operator, by definition $\Delta\kappa_t = \kappa_t - \kappa_{t-1}$ for any variable κ. We then get, from equation (12.14):

$$\Delta y_{it} = \alpha_i \Delta\delta_{it} + \Delta\varepsilon_{it} \tag{12.16}$$

which implies, using the common trend assumption:

$$\mathbb{E}[\Delta y_{it}|g] = \delta_{it}\mathbb{E}[\alpha_i|g] + \Delta m_t$$

Let us therefore suppose, to lighten the notation, that the observations concern only two periods. In period $(t-1)$ (1986 and before) the same economic policy applies to all individuals, while in period t (after 1986, i.e., when the reform is implemented), economic policy is altered for individuals $i \in T$. For individuals $i \in C$ of the control group, there is no alteration. Since the model has only two periods, the last equation implies that:

$$\mathbb{E}[\Delta y_{it}|g = T] - \mathbb{E}[\Delta y_{it}|g = C] = \mathbb{E}[\alpha_i|g = T] \tag{12.17}$$

We see that the average treatment effect on the treated, equal to $\mathbb{E}[\alpha_i|g = T]$, is merely equal to the difference between the changes in average outcomes across groups. It should be remarked that the average treatment effect on the treated (ATT) is different from the average treatment effect (ATE) when it is assumed that the effect of the treatment is potentially heterogeneous across individuals because the potential average effect of the treatment on the untreated may be different from that on the treated. Obviously, if it is assumed that the effect of the treatment is identical for all individuals ($\alpha_i = \alpha$ for all i), the average effect of the treatment on the treated is equal to the average effect of the treatment.

The sample analog of equation (12.17) is the difference-in-differences estimator:

$$\hat{\alpha} = \frac{\sum_{i \in T} \Delta y_{it}}{N_T} - \frac{\sum_{i \in C} \Delta y_{it}}{N_C} \tag{12.18}$$

To construct it, we first calculate the average within each group of the differences between the dates $(t-1)$ and t of the dependent variable y, then we calculate the difference between these two averages.

Labor Market Participation

The first two columns of table 12.4 represent the average of the participation rates for the periods 1984–1986 and 1988–1990 respectively. The third column shows, for each group, the difference between these averages before and after the reform. In this column, the figures 0.024 and 0.000 thus respectively represent the terms $\left(\sum_{i \in T} \Delta y_i\right)/N_T$ and $\left(\sum_{i \in C} \Delta y_i\right)/N_C$ of relation (12.18). The difference-in-differences estimator is then deduced and reported in column 4.

TABLE 12.4
Participation rates of single women. Standard errors in parentheses.

	Pre-TRA86	Post-TRA86	Difference	$\hat{\alpha}$
Treated group	0.729 (0.004)	0.753 (0.004)	0.024 (0.006)	
Control group	0.952 (0.001)	0.952 (0.001)	0.000 (0.002)	0.024 (0.006)

Note: TRA86 refers to the Tax Reform Act of 1986, which expanded the EITC.

Source: Eissa and Liebman (1996, table 2).

Table 12.4 shows that the participation rate increased in the treated group after the reform, whereas it remained constant in the control group. This result suggests that the EITC had a positive impact on the labor supply of single women with children.

When the difference-in-differences estimator is obtained with repeated cross sections, it is important to check that the composition of the group with respect to the unobservable group-fixed effects remains constant. Although unobserved fixed effects cannot be observed by definition, it is desirable to control for observed individual characteristics.[3] To do this, it is possible to obtain the difference-in-differences estimator from the estimation of a regression equation, which includes such controls. Let us first show what type of regression equation leads to the difference-in-differences estimator without these additional controls for the sake of simplicity. The additional controls will then be introduced. This equation can be written:

$$y_{it} = \beta + \alpha(\nu_g \times \mu_t) + \alpha_1 \nu_g + \alpha_2 \mu_t + \varepsilon_{it} \tag{12.19}$$

where ν_g is a group dummy, equal to 1 for women with children ($i \in T$) and to zero for women without children ($i = C$); μ_t is a time dummy equal to 1 at date t (after 1986) and to zero at date $t-1$ (before 1986).

This equation can be written in differences between t and $t-1$, where the time index has been dropped since there are only 2 periods:

$$\mathbb{E}(\Delta y_i) = \gamma + \alpha\delta_i + \Delta\varepsilon_i \tag{12.20}$$

where $\delta_i = 1$ for the treated individuals, corresponding to women with children, and $\delta_i = 0$ for women without children; $\gamma = \alpha_2 \Delta\mu$ is a constant term identical for all individuals. By definition the estimator of ordinary least squares of coefficient α is then given by:

$$\hat{\alpha} = \frac{Cov(\delta, \Delta y)}{Var(\delta)} = \frac{\sum_{i \in T} \left(\delta_i - \overline{\delta}\right)\left(\Delta y_i - \overline{\Delta y}\right) + \sum_{i \in C} \left(\delta_i - \overline{\delta}\right)\left(\Delta y_i - \overline{\Delta y}\right)}{\sum_{i \in T} \left(\delta_i - \overline{\delta}\right)^2 + \sum_{i \in C} \left(\delta_i - \overline{\delta}\right)^2}$$

where $\overline{\delta}$ and $\overline{\Delta y}$ designate respectively the average values of δ and Δy. Since $\delta_i = 1$ for $i \in T$ and $\delta_i = 0$ for $i \in C$, after several simple calculations, it can easily be verified that $\hat{\alpha}$ has the same expression as that defined by equation (12.18).

This estimator can be given a causal interpretation if the error term $\Delta\varepsilon_i$ in equation (12.20) is independent of δ_i, or formally: $Cov(\delta_i, \Delta\varepsilon_i) = 0$, which means that expected changes in error terms over time are identical across the treatment and the control groups. This corresponds to the common trend assumption. Eissa and Liebman add to the regression equation a vector $\mathbf{x}_{it}$ of individual characteristics, including unearned income, number of children, family size, number of preschool children, age and its square and cube, education and its square, and a dummy variable for race (= 1 if nonwhite). They also include year dummies for 1984, 1985, 1989, and 1990. These variables

[3] If individuals are followed over time, forming a panel, then individual fixed effects can also be added to the specification to control for time-invariant unobserved heterogeneity.

control for observable differences in the characteristics of the treatment and control groups that affect the level of labor force participation. Actually, the benchmark specification of Eissa and Liebman is a probit model where y_{it} is equal to one if individual i is employed at date t and y_{it} equals zero otherwise, so that the equation that is estimated is:

$$\Pr(y_{it} = 1) = \Phi\left[\beta + \alpha(\nu_g \times \mu_t) + \alpha_1 \nu_g + \alpha_2 \mu_t + \mathbf{x}_{it}\boldsymbol{\alpha}_3\right] \tag{12.21}$$

where Φ is the cumulative distribution function of the standard normal distribution. Since the probit model is nonlinear, the coefficients of the probit model cannot be used directly to compute the marginal effects. In particular, the coefficient α is not the difference-in-differences estimator (see Ai and Norton, 2003, and the discussion of this issue in Athey and Imbens, 2007). Eissa and Liebman calculate the effect of the Tax Reform Act of 1986 by predicting two probabilities of participation, one with the interaction variable $(\nu_g \times \mu_t)$ set equal to one and the other with the interaction term $(\nu_g \times \mu_t)$ set equal to zero. The measure of the treatment effect is the difference (over the sample of post-1987 women with children) in the two probabilities of participation.[4]

The estimation of equation (12.21) leads to the conclusion that single women caring for at least one child saw their probability of participating in the labor market grow, on average, by 1.9 percentage points, which is of the same order of magnitude as the 2.4 percentage points appearing in the third column of table 12.4. This result provides further evidence that the EITC did have a positive impact on participation. Eissa and Liebman discuss further the interpretation of this result in their paper. In particular they argue that the timing of the observed changes in labor force participation is consistent with this interpretation. They also show that the treatment had its largest effect among people most likely to be eligible for the credit: the estimated effect of the treatment on participation response is 6.1 percentage points for the less than high school sample, 2.6 percentage points for the high school sample, and only 0.4 percentage point for the beyond high school sample.

Hours Worked

Eissa and Liebman also evaluate the impact of the EITC expansion on hours worked. The analysis of hours worked is a little more complex than that of labor market participation because the predicted impact of the EITC expansion on hours of work depends on the taxpayer's income. To observe this, it is helpful to refer to figure 12.10. For most workers in region A (incomes between \$0 and \$14,081), the EITC expansion is predicted to have an ambiguous impact on hours of work, since the lower marginal tax rate has income effects that decrease hours (assuming that leisure is a normal good) and substitution

[4]The justification is that observed participation of the treated at date t minus observed participation at date $t-1$ is:

$$\Phi(\beta + \alpha + \alpha_1 + \alpha_2 + \mathbf{x}_{it}\boldsymbol{\alpha}_3) - \Phi(\beta + \alpha_1 + \mathbf{x}_{it-1}\boldsymbol{\alpha}_3)$$

whereas the *potential* participation of the treated if they had had no treatment at date t minus their observed participation at date $t-1$ is:

$$\Phi(\beta + \alpha_1 + \alpha_2 + \mathbf{x}_{it}\boldsymbol{\alpha}_3) - \Phi(\beta + \alpha_1 + \mathbf{x}_{it-1}\boldsymbol{\alpha}_3)$$

which implies that the effect of the treatment can be defined as:

$$\Phi(\beta + \alpha + \alpha_1 + \alpha_2 + \mathbf{x}_{it}\boldsymbol{\alpha}_3) - \Phi(\beta + \alpha_1 + \alpha_2 + \mathbf{x}_{it}\boldsymbol{\alpha}_3)$$

effects that increase hours. Workers in region B (incomes between $14,081 and $25,000) are predicted to reduce their hours of work either because they are in the expanded "phase-out" region and face a 10% higher marginal tax rate in addition to having their incomes increased or because they have incomes just above the expanded "phase-in" region and might reduce their hours of work to take advantage of the credit. Workers in region C (incomes above $25,000) are unlikely to be affected by the increase in the credit.

To examine how the EITC expansion affected hours of work conditional on working, Eissa and Liebman estimate equation (12.20) with OLS where y_{it} denotes annual hours of work. The same vector of control variables $\mathbf{x}_{it}$ as in equation (12.21) is added to equation (12.20). Eissa and Liebman estimate this equation for individuals who declare positive hours only. This means that they estimate the change in hours conditional on hours of work exceeding zero. This estimation does not allow us to take into account the possibility that individuals induced to take a job by the EITC expansion may work different hours than those who would have worked without the EITC expansion. A way to deal with this issue is to use a system of two equations: one equation that explains hours of work and another that explains participation. In other words, we can add a "selection model" that explains how individuals select between participation and nonparticipation (see chapter 1, appendix 7.5) to the equation that explains hours of work. Eissa and Liebman argue that they do not impose a selection model on the data for two reasons. First, to identify a selection model, one would need a policy shift that affects participation separately from hours of work (see chapter 1, appendix 7.5). The Tax Reform Act of 1986 does not provide us with such a shift. They also argue that inferences in labor supply models are extremely sensitive to the model chosen. It is true that it is difficult to think of a variable that could influence the decision to work but not the hours of work and that the results may be sensitive to the model chosen. From this point of view, the approach of Eissa and Liebman makes sense, but readers should keep in mind that the failure to account for new participants may bias downward estimates of the impact of the EITC expansion on hours, to the extent that new participants are likely to enter the labor force with earnings and hours below those of individuals who would work in the absence of the EITC expansion.

1.3.2 Other Microeconomic Evidence on the Impact of Social Assistance Programs

The use of microeconomic data has been central to the study of the impact of social assistance programs on labor market outcomes. Essentially, these programs may take the form of permanent or temporary in-work benefits, or they may consist of minimum-income benefits.

Permanent In-Work Support

Based on the method presented in the previous section, Eissa and Liebman (1996) estimate that, conditional on working, the EITC expansion induced no significant changes in the hours of work of single women with children. Consequently, by reducing the marginal tax rate for low-wage earners, the EITC expansion appears to have had a positive impact on labor market participation but not on hours of work conditional on working. This result should be interpreted cautiously to the extent that the estimation

of the impact of the EITC expansion on hours of work suffers from potential selection problems, as noted above.

So EITC does have an impact on the extensive margin, that is, the decision to work or not, but evidence is mixed when it comes to the intensive margin (e.g., Eissa and Liebman, 1996; Meyer and Rosenbaum, 2001; Grogger, 2003; Eissa and Hoynes, 2004; Hotz and Scholz, 2006; Hotz et al., 2011; Gelber and Mitchell, 2012). However, Chetty, Friedman, and Saez (2013) have developed an original strategy which leads them to find a significant impact of EITC on the intensive margin. They noticed that in some cities in the United States, individuals have little knowledge about the policy's marginal impact on earnings (the EITC schedule) compared to other cities. To identify the cities with knowledgeable populations, they use the fact that self-employed people have more discretion than salaried workers when it comes to reporting income and are hence able to report earnings at exactly the level that maximizes the amount of transfer. For instance, 7.4% of EITC claimants in Chicago are self-employed and report total earnings at exactly the refund-maximizing level, compared with 0.6% in Rapid City. One of the key identifying assumptions is that these behaviors are indeed linked to the degree of knowledge about the scheme. To support this assumption, the authors show that people moving from low- to high-information cities increase the received transfers, but those moving from high- to low-information cities do not decrease their transfers, suggesting that the difference observed across cities stems from differences in the prevalence of information. Moreover, this behavior is highly correlated with predictors of information diffusion, such as the density of EITC recipients, the availability of professional tax preparers, and the frequency of Google searches for phrases including the word "tax." They then look at wage earners, whose income is easy to verify, and use those living in the low-knowledge cities as a control group compared to those living in high-knowledge cities, who constitute the treated group. However, cross-city comparisons of wage earnings distributions do not definitively establish that the EITC has a causal effect on earnings because there could be other differences across neighborhoods, such as differences in industrial structure or the supply of jobs that could explain the variation in EITC transfers. To take into account this type of omitted variable biases, the authors note that single individuals are essentially ineligible for the EITC, thus creating a natural control group that can be used to control for any differences across neighborhoods that are not caused by the EITC. They find that wage earnings in low-bunching and high-bunching cities track each other closely in the years prior to childbirth but that wage earnings distributions become much more concentrated around the EITC plateau in high-bunching areas, leading to larger EITC transfers in those areas after the first child is born. They estimate an intensive-margin earnings elasticity of 0.31 in the phase-in region of the EITC schedule and an intensive-margin earnings elasticity of 0.14 in the phase-out region on average in the United States. One explanation for the larger responses in the phase-in region is that individuals with very low incomes have higher elasticities than those holding full-time jobs. Another explanation is that, on average, individuals pay more attention to the phase-in and refund-maximizing plateau portions of the schedule than they do to the phase-out region.

Few other permanent in-work support programs such as the EITC have been evaluated because only a few others exist. One similar and large-scale program is the working families' tax credit (WFTC), which was introduced in the United Kingdom in 1999.

The problem with evaluating the impact of the WFTC is that it was rolled out nationally as an entitlement-based program—all those who apply and satisfy the eligibility conditions receive it—and so there is no suitable control group. In addition, at the time the WFTC was introduced, other reforms of the tax and transfer system occurred and affected some of the eligible families, which makes the identification of the impact of the measure even more difficult. For that reason, Brewer et al. (2006) estimate a structural model that predicts the behavior of households following an arbitrary tax and benefit reform, as a means of isolating the contribution of the WFTC to changes in labor supply. To accomplish this, they use equations describing households' labor supply behavior as a function of taxes and benefits, and they estimate the parameters in these equations using microdata from before and after the introduction of the tax credit. They find that by 2002 WFTC had increased the labor supply of single mothers by around 5.1 percentage points, slightly reduced the labor supply of mothers in couples by 0.6 percentage point, and increased the labor supply of fathers in couples by 0.8 percentage point, compared with the program that preceded it.

Temporary In-Work Benefits

If permanent in-work benefits are still rare, temporary ones are much more frequent. These schemes primarily aim at creating incentives for unemployed or inactive benefit recipients to resume work. In Canada a large-scale, controlled experiment tested the impact of such schemes. The "Self-Sufficiency Project" (SSP) was launched in 1992 and enrolled 9,000 single-parent, social assistance recipients. The program offered a significant earnings top-up for up to three years to parents who had already been on welfare for at least one year, if they found a full-time job within 12 months of random assignment. Comparisons between the treatment and control groups show that SSP had significant effects on work in the short run, raising the full-time employment rate and lowering welfare participation by 14 percentage points within the first 18 months of the experiment (Michalopoulos et al., 2002). But the effects of this temporary bonus tend to fade over time: by the third year after random assignment, the difference in welfare participation between the treatment and control groups had fallen to 7.5%, and 52 months after random assignment, and a few months after the subsidy payments had ended, the employment rates and the distributions of wages of the program and control groups were equal (Card and Hyslop, 2005).

A second experiment was also designed in 1994–1995 to measure the effects of the subsidy offer on new welfare entrants. New welfare entrants were informed that if they remained on public assistance for a year they would become eligible to receive a generous earnings subsidy when they resumed full-time work. Those who satisfied the waiting period and then left welfare and began working full-time within the following year were entitled to receive payments for up to 36 months whenever they were off welfare and working full-time. Card and Robins (2005) and Card and Hyslop (2009) find that targeting long-term benefit recipients created a waiting period, increasing welfare participation in the first year after initial entry. But they also find reduced participation in welfare thanks to the subsidy over the 5 following years, notably during the 12 months during which participants need to "lock in" the subsidy by finding work. After that initial period participants tend to keep their jobs as long as the subsidy is paid and even a little after (see figure 12.11).

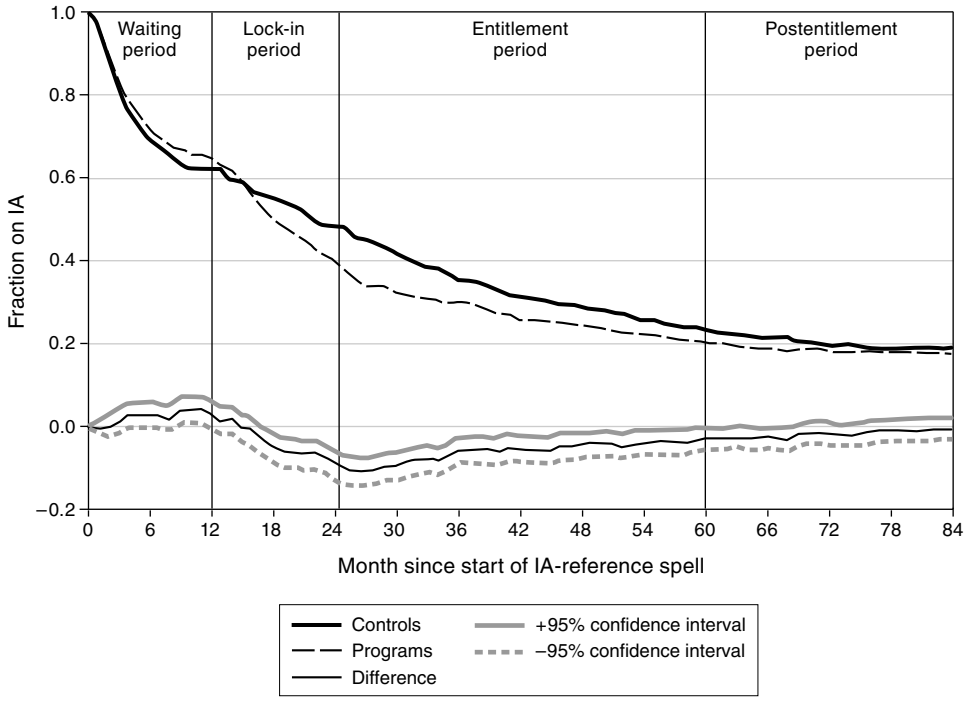


Figure 12.11
The impact of a temporary in-work benefit program in Canada (SPP) on income assistance (IA) receipt.

Source: Card and Hyslop (2009, figure 1a, p. 5).

Minimum-Income Benefits: The Static Approach

Despite the growth of in-work benefits in recent years, the main form of welfare in the majority of developed economies remains income replacement benefits paid to those not working. The impact of these benefits on labor market participation depends on both effective targeting and the efficiency of active programs to help recipients to find jobs (see chapter 14). There is evidence that this type of benefit can reduce the labor supply of those who have the ability to work. For instance, prior to 1989 in Quebec, childless persons younger than 30 years old received substantially less in welfare payments than similar individuals 30 years of age or older ($185 per month compared to $507). The 1989 welfare reform ended this differential rate. Since individuals obviously cannot choose to be above or below the age of 30, this 1989 reset spontaneously generates a control group (those above 30 receiving high benefits before and after the reform) and a treated group (those receiving lower benefits before and higher benefits after the reform), thus minimizing the risk of selection bias that could undermine the evaluation. Using administrative data, Fortin et al. (2004) studied the effect of higher benefits on the duration of welfare spells. They estimate hazard models analogous to the difference-in-differences approach,[5] comparing hazard rates for those above and below age 30, before

[5] See chapter 5, section 3.2, for an example of this method.

and after the reform, and controlling for observed characteristics that could influence the duration of welfare. They find that higher benefits increased the average duration of social assistance benefits among those under 30 by 4 to 8 months (depending on subgroups).

Administrative data typically allow us to follow individuals over time and to study precisely the dynamics of benefit receipt, but they include no more than limited information on beneficiaries' characteristics (Fortin et al. are only able to control for age, gender, education, region, and immigrant status). Now, the marital status or the household's income level may vary over time and have a strong impact on employment. Besides, if important unobservable characteristics are correlated with age, then studying behavior just below and just above age 30, rather than making broader comparisons of all under-30s and over-30s of any age, can improve the identification. For these reasons, Lemieux and Milligan (2008) used survey information and a regression discontinuity design to study the same reform.[6] They find that generous social assistance benefit reduces the employment probability of less-educated men without children by 3 to 5 percentage points (see figure 12.12). This significant but still relatively modest adjustment compared with a 175% increase in benefits is consistent with

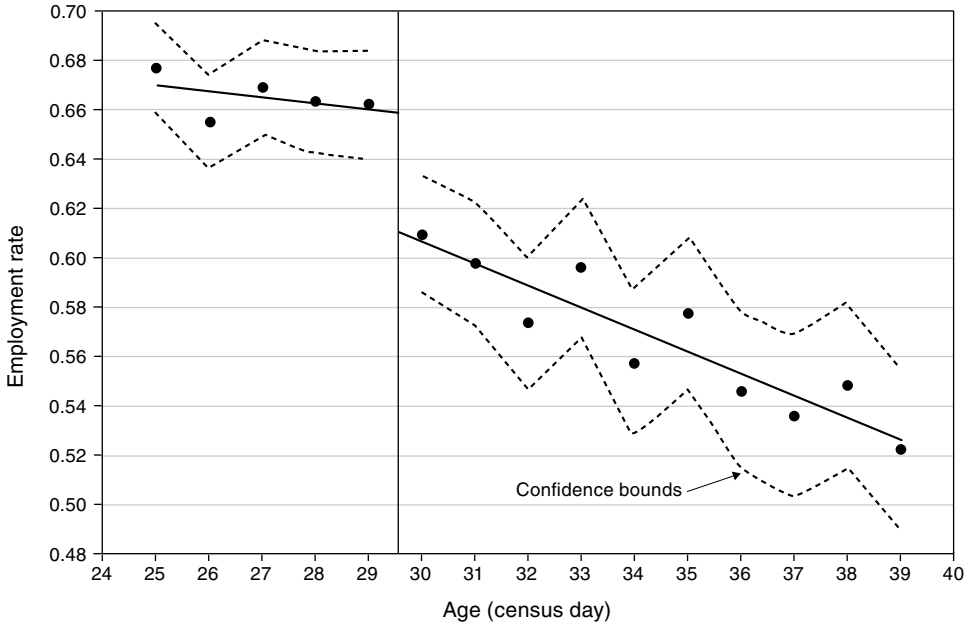


Figure 12.12
The effect on employment of an increase in social assistance benefits at age 30 in Quebec.

Source: Lemieux and Milligan (2008, figure 3, p. 816).

[6] See chapter 7, section 4.2, for a formal presentation of this method. Lemieux and Mulligan (2008) also run a difference-in-differences model on the same data and find results similar to those obtained with the regression discontinuity.

relatively modest behavioral effects. They also find that, as expected, the take-up of social assistance increases when the amount of benefit rises.

Similar results have been obtained for continental Europe. For instance, in France the main social assistance scheme, the Revenu Minimum d'Insertion (RMI), was introduced in 1989. But a specific region, namely Alsace, already had a system of social assistance similar to the RMI before it was introduced. Chemin and Wasmer (2012) run a difference-in-differences between low-income individuals living in Alsace and in the rest of France, after and before the reform. They find that the reform was associated with a 3% fall in employment, and the effect was even stronger for single parents. Another feature of the RMI is that childless single individuals under age 25 were not eligible: their only eligibility was for housing allowance. As a result, out-of-work payments would increase by 160% as they turned 25 years old. Bargain and Doorley (2011) analyze this difference in a regression discontinuity setting based on French census data. They find that the RMI reduces the participation of uneducated single men by 7%–10% at age 25, but no significant effect can be identified for educated individuals.

In the United States, the Welfare Reform legislation in 1996 introduced time limits that for most families restricted the maximum duration of receipt to 60 months of federally funded benefits, over one or several spells. Imposing time limits should reduce the benefit receipt rates automatically, once recipients exhaust their benefits. But it could also create incentives to exit programs in anticipation of the time limit in order to preserve their benefits for the future. Since states implemented time limits at different dates, the differences in the timing provide a means of estimating the effects of the reform on the rate of receipt. Grogger (2002) shows that time limits had only limited behavioral effects: most social assistance dependent families were not responsive to limits, except for those with young children, for whom the effects were substantial.

These results are difficult to generalize, especially because the estimated effect is an average treatment effect on the treated, sometimes evaluated for very narrow populations (e.g., childless, low-educated, young men at around age 30). However, they suggest that an increase in minimum income benefit does reduce labor supply at the extensive margin for some groups, which include low-educated, single parents and young men.

Minimum-Income Benefits: The Dynamic Approach

The previous studies compare labor market status in a static manner across groups. But there is also a dynamic effect of income replacement programs, as it is commonly observed that rates of benefit receipt are greater for individuals who have received benefits in the past than for individuals who have not: receipt history matters. Actually an individual's benefit receipt history can have a strong association with current receipt. This suggests that there is a form of "state dependence" in benefit receipt. The odds of employment might be reduced over time just because of the stigma associated with the receipt of social assistance or because of lower self-confidence, job search effort, or motivation. Measuring the degree of state dependence requires the researcher to control for observable and unobservable characteristics that may influence the chances of receiving benefits. Such characteristics are indeed likely to differ between recipients and nonrecipients and over time. Failure to compare otherwise identical individuals could lead to biased estimates of structural state dependence: previous benefit receipt might look like a determinant of future benefit receipt just because it is a proxy for temporally persistent characteristics (Heckman, 1981). This can be controlled for using panel data

sets containing sufficient information on the characteristics and labor market history of beneficiaries, while following the same individuals over time. When these data are available, one approach is to regress the following equation:

$$y_{it} = \alpha + \beta y_{i,t-1} + \mathbf{x}_{it}\boldsymbol{\gamma} + \eta_i + \varepsilon_{it} \tag{12.22}$$

where y_{it} is a dummy that has value 1 if individual i is receiving benefits at date t and 0 otherwise, $\mathbf{x}_{it}$ is a vector of exogenous individual characteristics, η_i is an individual time-invariant factor that is not directly observed by the econometrician, and ε_{it} is a random error. If $\beta \neq 0$, there is state dependence. We can also bring time-specific effects into this equation to control for the macroeconomic situation and any policy reform. Note that this equation implicitly assumes that only receipt in the previous period matters and not receipt in earlier years. This assumption is commonly used in this literature. One complication is that individuals whose unobserved characteristics make them more prone to receive benefits, other things being equal, are also more likely to receive benefits in the first year in which they are observed. That is, y_{i1} is likely to be correlated with the unobserved factor η_i, inducing a correlation between the ε_{it} and the lagged dependent variable y_{it-1}. Ignoring this would lead to biased estimates of the parameters. Several methods can be used to deal with this "initial conditions" problem (see Wooldridge, 2010, chapter 15, for more detail). For instance, Wooldridge (2005) relies on the assumption that once we include the initial value of the outcome variable y_{i0} and the lags and leads of all explanatory variables appearing in $\mathbf{x}_{it}$ as additional regressors in equation (12.22), all remaining unobserved heterogeneity η_i is uncorrelated with the outcome in the initial period. In practice, a simplification widely adopted is to include in equation (12.22) the vector of individual longitudinal averages of all time-varying observed characteristics in $\mathbf{x}_{it}$ instead of all past and future values.

This equation can be estimated by the OLS (linear probability model), but most often random-effects probit models are used. For instance, adopting this approach and using data from the British Household Panel Survey between 1991 and 2005, Capellari and Jenkins (2014) find that the risk of receiving social assistance in one year is 14 percentage points higher if social assistance was also received in the previous year, even after controlling for observed time-varying and unobserved time-invariant characteristics. Similarly, exploiting the Survey of Income and Program Participation (SIPP) in the United States, Chay and Hyslop (2014) find that individual characteristics explain about 50% to 70% of the persistence in welfare receipt and the rest is state dependence. Studying social assistance receipt in Sweden in the 1990s, Hansen and Lofstrom (2008) find that differences in observable characteristics only account for a small part of the differences in the frequency of benefit receipt between migrants and natives, but that state dependence in benefit receipt may be higher for migrants. This approach has also been used to study the presence of state dependence in the phenomena of unemployment and labor market participation (see Hyslop, 1999, and Stewart, 2007).

1.3.3 Macroeconomic Evidence: Why Do Americans Work So Much More Than Europeans?

The macroeconomic literature studies the impact of taxation on unemployment, employment, and hours worked at the level of the entire economy. It suggests that different systems of mandatory contribution in different countries explain a large portion of the spread in the number of hours worked per person.

Macroeconomic Time Series and Panel Data

In the wake of the contributions of Layard et al. (1991) and Nickell et al. (2005), one strand of research uses cross-country panel data to investigate the impact of average rates of taxation on unemployment and employment. These papers generally find that an increase in fiscal pressure as measured by the tax wedge does increase unemployment and reduce employment, though fiscal pressure has less impact in the Scandinavian countries (see Alesina and Perroti, 1997, and Daveri and Tabellini, 2000). Still, it is a delicate matter to interpret this correlation as a causal relation, inasmuch as there may be numerous reasons for variations in taxation to be correlated with events imperfectly observed by the econometrician—events that also exert influence on labor market performances. Another strand investigates the impact of tax progressivity on wages, using time series data at the macroeconomic level, the sectoral level, or for groups of workers (Malcomson and Sator, 1987; Lockwood and Manning, 1993; Holmlund and Kolm, 1995; Hansen et al., 2000; Lehmann et al., 2013). This research generally brings to light a negative correlation between the progressivity of taxes and net wages, in conformity with the theoretical predictions set out above. And for the same reasons as before, it is a tricky matter to interpret these results as a sequence of cause and effect. As well, wage changes that move the averages may be the upshot of composition effects arising out of variations in the composition of jobs induced by changes in progressivity.

Overall, correlations obtained on the basis of macroeconomic data tend to confirm theoretical predictions to the effect that an increase in the tax wedge leads to an increase in unemployment and a reduction in employment, while an increase in progressivity exerts downward pressure on wages and unemployment. Research presented in chapter 1, section 3.2.3, shows for that matter that the impact of taxes on hours worked as estimated on macroeconomic data is close to that estimated on microeconomic data. Nevertheless, the very nature of macroeconomic data, which bear on averages at a highly aggregate level and which cover countries numbering no more than a few dozen at best, makes it advisable to use prudence in interpreting the results of this research.

Another approach relies on models with microfoundations to explain cross-country differences in labor market performance. It has generated a number of publications following the article of Prescott (2004).

The Contribution of Prescott (2004)

The paper by Eissa and Liebman (1996) on the EITC presented above, like the rest of the empirical literature, indicates that taxes do have a significant negative impact on labor supply. Such a result might signify that variation in the rates at which tax is levied explains, to some degree at least, the differences in employment rates and hours worked that we observe across countries. This problem has been a focus of debate since the publication of the paper by Prescott (2004), which contends that differences in the time path of taxation between the United States and Europe explain why Europeans (or to be precise, the inhabitants of France, Germany, Italy, and the United Kingdom, for Prescott's purposes) were spending on average 30% fewer hours at work in the middle of the 1990s than Americans were, whereas the same Europeans had been working just as long, or even a bit longer, at the start of the 1970s.

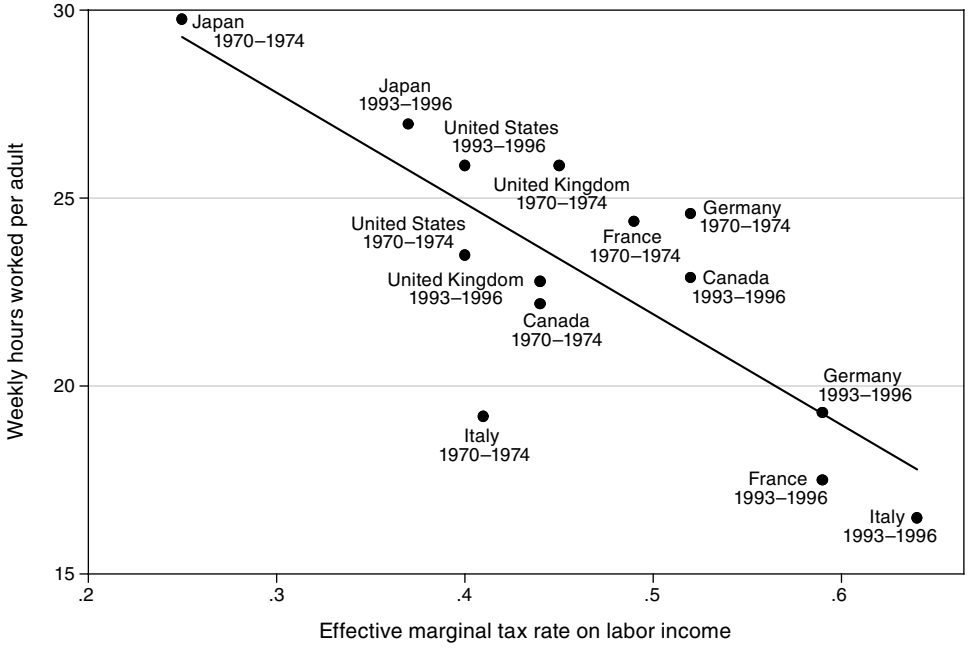


Figure 12.13
Taxes on labor income and weekly hours worked per individual in the age range 15–64 in 1970–1974 and 1993–1996.

Note: Taxes on labor income take in consumption taxes, income tax, and social security contributions; see equation (12.25) for a precise definition.

Source: Prescott (2004).

Figure 12.13 shows that for the ensemble of G7 countries which make up Prescott's sample, there is indeed a very marked decreasing relation between rates of taxation on wages earned at work and the number of hours worked per adult (i.e., per person in the age range 15–64), from the start of the 1970s to the 1990s. In the United States, where the duration of hours worked has practically remained stable, taxes have remained practically unvaried. Conversely, tax rates have risen strongly in most other countries, where the duration of hours worked has on average fallen more than in the United States. Figure 12.14 shows that this decreasing relation between hours worked and fiscal pressure also obtains when we consider a larger sample, made up of 15 OECD countries, and a longer period, from 1970 to 2010.

To show that differences in hours worked across countries may reflect differences in taxation, Prescott uses a model of intertemporal labor supply, of which we present here a static version that highlights the main thrust of the argument. The preferences of the representative individual are described by the utility function:

$$U = \ln c + \alpha \ln(1 - h)$$

where c and h designate consumption and hours worked, or more precisely the proportion of disposable time dedicated to working. The parameter $\alpha > 0$ specifies the value

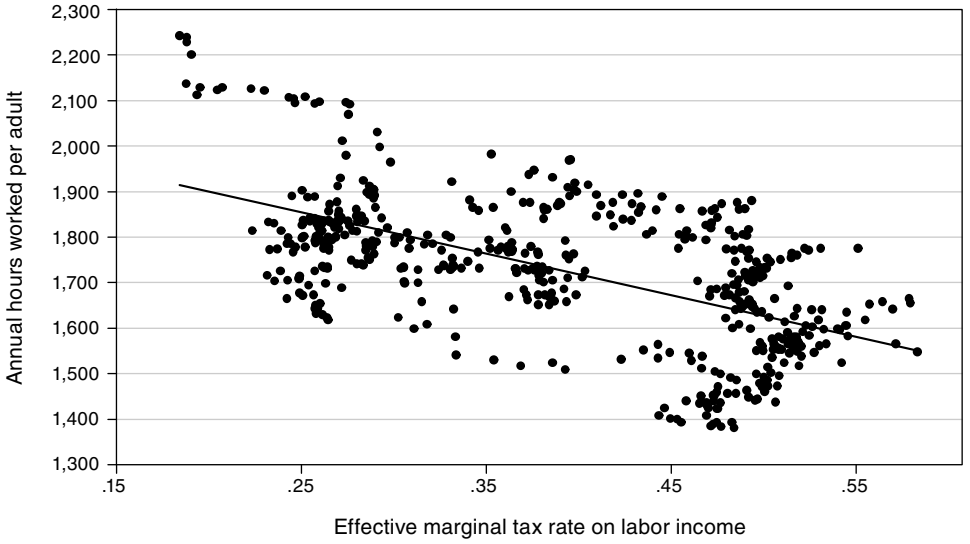


Figure 12.14

Taxes on labor income and annual hours worked in 15 OECD countries (Australia, Austria, Belgium, Canada, Finland, France, Germany, Italy, Japan, the Netherlands, Spain, Sweden, Switzerland, the United Kingdom, and the United States) over the period 1970–2010.

Note: Taxes on labor income take in consumption tax, income tax, and social security contributions; see equation (12.25) for a precise definition. Each dot corresponds to a country-year observation.

Source McDaniel (2011, and www.caramcdaniel.com) and OECD.

of leisure time for the household. Each unit of labor produces one unit of good, which implies, together with the zero profit condition, that the wage is equal to 1, assuming that firms pay no tax. Taxes on consumption and on labor earnings are assumed to be proportional for the sake of simplicity. The budget constraint of the consumer is:

$$(1+\tau_c)c = h(1-\tau_h) + T \tag{12.23}$$

where τ_c and τ_h stand respectively for consumption and the rate of tax on labor earnings, and T denotes lump-sum transfers from the government. The budget constraint (12.23) gives the value of c as a function of h. Carrying this value into the utility function U and deriving with respect to h, we find that the optimal value of h verifies:

$$\frac{(1-\tau_h)}{h(1-\tau_h)+T} - \alpha\frac{1}{1-h} = 0 \iff \frac{\frac{1-\tau_h}{1+\tau_c}}{c} = \alpha\frac{1}{1-h}$$

This equation may also be written:

$$(1-\tau) = \frac{\alpha c}{1-h} \tag{12.24}$$

where

$$\tau = \frac{\tau_c + \tau_h}{1 + \tau_c} \tag{12.25}$$

is called the *effective marginal tax rate* on labor income, which is the fraction of labor income that is extracted in the form of taxes. It is this tax rate which is displayed in figures 12.13 and 12.14.

The budget constraint of the government implies that lump-sum transfers T are equal to tax receipts, which implies that $T = \tau_c c + h\tau_h$. Using the budget constraint (12.23) of the household, we find that $c = h$, that is, that consumption equals labor income. Substituting this equality into (12.24), we get the equilibrium number of hours worked:

$$h = \frac{1}{\frac{\alpha}{1-\tau} + 1} \tag{12.26}$$

This equation shows that the duration of hours worked does decrease with the effective marginal tax rate. Prescott computes the average across-country value of τ and sets the value of α equal to 1.54, in order to match the average number of weekly hours worked in the model with the actual value for the G7 countries. This allows him to show that across-country variation in taxes explains most of the differences in hours worked.

A question that arises is whether the elasticity of labor supply implied by the calibrated model is compatible with the usual microestimates of the elasticity of labor supply. To address this question, the first thing to note is that in Prescott's model changes in taxes are compensated by changes in transfers, implying that changes in taxes have no income effects. As shown above, we always have $c = h$ ex post, once taxes and lump-sum transfers T have been paid. Accordingly, equation (12.26) allows us to compute the *Hicksian* elasticity of labor supply with respect to net labor income $(1 - \tau)$, equal to:

$$\frac{\mathrm{d}\ln h}{\mathrm{d}\ln(1-\tau)} = \frac{\alpha}{\alpha + 1 - \tau}$$

The reader will see that the Hicksian elasticity of labor supply depends on the parameter α and on tax rates. In Prescott's data the average value of $1 - \tau$ amounts to 0.53. Accordingly, the Hicksian elasticity of labor supply predicted by the model is about 0.74, which is only a little greater than the average of estimated values of the Hicksian elasticity reported in chapter 1. Hence the model fits the data surprisingly well.

Beyond Prescott (2004)

Still, it remains the case that the relation between hours worked and taxation observed at the macroeconomic level that permits us to deduce this value for elasticity is far from robust. If we interpret the data displayed in figure 12.13 on the assumption that each point describes a stable situation close to a stationary state, it is possible to estimate the Hicksian elasticity of labor supply from the data in this figure by simply regressing the logarithm of hours worked on the logarithm of $1 - \tau$ (see Chetty, Manoli et al., 2013).

We then arrive at the equation (standard errors are in parentheses):

$$\ln h = \underset{(0.10)}{3.57} + \underset{(0.15)}{0.70}\ln(1-\tau),\ R^2 = 0.62$$

which does in fact yield an elasticity of 0.7—very near, by construction, to that obtained by Prescott. With the data in figure 12.14, which takes in 15 countries over the period 1970–2010 and comprises 505 observations instead of 14, the estimated Hicksian elasticity with simple OLS amounts to 0.3. But when country-fixed effects (to account for time-invariant country specificities) and year-fixed effects (to account for macroeconomic effects common across countries) are introduced, we get a coefficient with opposite sign (implying that hours increase when taxes rise), equal to −0.1 and different from zero at the 1% level of confidence.

This is not surprising. The fact is, OLS estimators are certainly biased for at least three reasons. First, hours of work are influenced by many other factors besides taxes. From this perpective, Alesina et al. (2005) argue that the cross-sectional relationship between taxes and hours worked is the result of omitted variables that are correlated with taxes and hours worked. Using a panel of countries, they show that the correlation between taxes and hours worked disappears once unionization and employment protection are included in the regression. Another example is given by Rogerson (2007), who shows that differences in the spending patterns of governments can account for the large labor supply in European Nordic countries, in spite of high taxes. In the European Nordic countries a larger portion of public expenditures is devoted to provision of family services, child care, and transfers that are conditional on working. These public expenditures reduce the cost of work and thus reduce the distorting effects of taxes. Accordingly, depending on the nature of public expenditure, taxes can have different distorting effects (see also Ljungqvist and Sargent, 2006).

Second, taxes are likely endogenous: they are influenced by factors that also influence labor supply. For example, a trend toward greater labor productivity may reduce labor supply on account of the income effect and also increase the demand for collective goods financed out of tax revenue (Wagner's law). Such a phenomenon would induce a negative correlation between taxes and hours worked. McDaniel (2011) argues that this phenomenon has indeed played a role in the case of the OECD countries since the 1970s: she estimates, using a calibrated growth model extended to include home production and subsistence consumption, that the primary force driving changes in market hours is the changing labor income tax rates but that productivity catch-up relative to the United States is an important secondary driver.

Third, the impact of taxes on total hours worked in the economy depends on the composition of the population. This perspective has been foregrounded by Rogerson and Wallenius (2009), who provide a life-cycle model of labor supply that represents the choices at the extensive margin of young people when they decide to enter the labor force and of older workers when they decide to retire. These choices, which explain a large share of the across-country differences in hours worked (see Blundell et al., 2013), can exert an important effect on the macroelasticities of labor supply.

Let us show this in a simple version of the model of Rogerson and Wallenius. Time is continuous and lifetime is normalized to one. The individual is endowed with one unit of time at each instant. Letting a denote age, the representative individual

has preferences with respect to paths for consumption $\{c(a)\}$ and hours worked $\{h(a)\}$ given by:

$$\int_0^1 \{\ln c(a) + \alpha \ln [1 - h(a)]\}\, \mathrm{d}a \tag{12.27}$$

where $\alpha > 0$ specifies the value of leisure time. It is assumed for the sake of simplicity that individuals do not discount future utility. Labor is the only factor of production. One unit of labor produces one unit of good. The individual of age a who works h hours produces:

$$\ell(h, a) = e(h)g(a) \quad \text{where} \begin{cases} e(h) = \max(h - h_f, 0) \\ g(a) = \frac{1}{2} - \left|\frac{1}{2} - a\right| \end{cases} \tag{12.28}$$

where $h_f > 0$ stands for a fixed cost of work (one needs to work at least h_f hours to start producing), and $g(a)$ defines the productivity age profile. For the sake of simplicity, we consider here a piecewise linear productivity profile that is symmetric around mid-life. When very young (a close to zero) or very old (a close to one), the returns to hours of work go to zero, while they reach a maximum, equal to $1/2$, at mid-life.

Rogerson and Wallenius assume free entry, which implies that workers get a gross wage equal to their productivity. The net wage is equal to the gross wage times $(1 - \tau)$ where τ denotes the proportional tax rate. As in Prescott's framework, we assume that tax receipts are transferred to individuals in a lump-sum fashion. The transfer is denoted by T. Moreover, individuals can lend and borrow at no cost. Accordingly, individuals maximize the utility (12.27) subject to the budget constraint:

$$\int_0^1 c(a)\mathrm{d}a = \int_0^1 (1 - \tau)\ell\,[h(a), a]\, \mathrm{d}a + T$$

The solution to this problem for the path of hours is (see appendix):

$$h(a) = \begin{cases} 1 - \frac{\alpha c}{(1-\tau)g(a)} & \text{if} a \in [\frac{\alpha c}{(1-h_f)(1-\tau)}, \quad 1 - \frac{\alpha c}{(1-h_f)(1-\tau)}] \\ 0 & \text{otherwise} \end{cases}$$

where $c = \int_0^1 \ell\,[h(a), a]\mathrm{d}a$ denotes total consumption during lifetime, which does not depend on a particular age. This solution is represented in figure 12.15. It turns out that the combination of fixed cost of labor and productivity age profile implies that individuals choose to work at ages where productivity is sufficiently high. With the productivity profile, which starts from zero at age zero and increases until mid-life, then decreases back to zero [see equation (12.28)], there always exists a threshold age below which individuals decide not to work and another age for which they retire.

As shown in figure 12.15, when taxes are higher, individuals start their careers older and retire younger; moreover, they work fewer hours when they are employed. The model of Rogerson and Wallenius shows that the elasticity of hours worked in the population depends on the proportions of young and old individuals who change their behavior when taxes move. If these proportions are large, the macroelasticity of

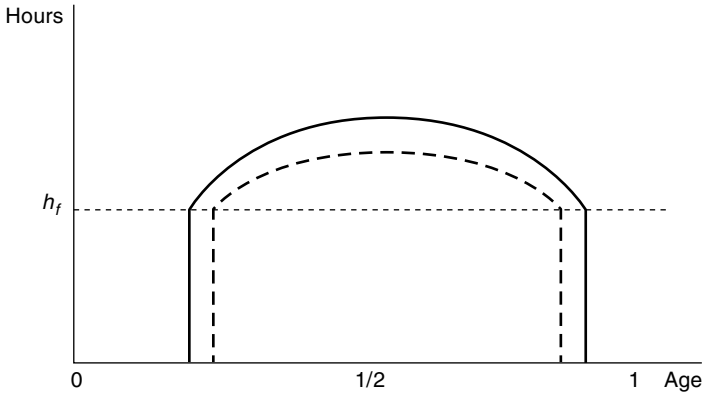


FIGURE 12.15
The impact of an increase in taxes in the model of Rogerson and Wallenius (2009).

Note: The bold continuous line displays hours worked before the tax increase and the bold dotted line displays the hours worked after the tax increase.

hours worked can be large even though the elasticity at the intensive margin is small on average. More generally, the macroelasticity of hours worked depends on the composition of the population, which can comprise demographic groups with very heterogeneous behaviors, as stressed by Blundell et al. (2013).

Overall, the debate centered on Prescott's article instructs us that differences between countries in the taxation of labor income can explain large differences in hours worked, for verisimilar values of the Hicksian elasticity of labor supply. In practice, the impact of taxes on hours worked depends on the bundle of features of each separate fiscal system, for there is a wide range of average and marginal rates varying with amount and kind of income and with types of households and their share in the population. In light of this, it would be erroneous to conclude that a country with a higher *effective marginal tax rate* on labor income, measured as it is in Prescott's type of study—that is, as an average value—has a tax system that is necessarily more detrimental to the supply of hours worked.

2 THE MINIMUM WAGE

Minimum wage legislation exists in 25 OECD countries. Such legislation has generally been framed to achieve the goal of compressing wage inequality. But the effectiveness of the minimum wage as a policy lever for achieving income redistribution is often doubted, since by raising the cost of labor it can have negative impacts on output and employment. Economic analysis suggests that the effects of the minimum wage on employment actually depend on the initial level of the minimum wage. When it is set relatively low to start with, subsequent increases are not necessarily unfavorable to employment. But if the minimum wage is set relatively high to start with, subsequent increases likely do exert a negative impact on hiring. This analytical result receives a degree of confirmation from empirical research.

To point out that the minimum wage can have positive effects on employment does not put an end to the question of whether or not it can be justified, since there may be more efficient policy levers available, like taxation, when it comes to improving resource allocation and redistributing income.

2.1 A Constraint of Varying Strength from Country to Country

Minimum wage legislation and its incidence vary greatly from country to country, but it covers populations that are much alike everywhere.

2.1.1 The Minimum Wage: Legal Aspects and Orders of Magnitude

Minimum wages, set by law or by collective agreement, exist in all European Union countries and a large number of OECD ones. The legislation governing them, however, varies widely. The legal (i.e., set by law) minimum wage may be regional (the United States, Canada, Japan) or national (France, the Netherlands, the United Kingdom since April 1999). It can also be set exclusively by collective agreements instead of law and then varies according to industry (Austria, Italy, Germany, and the Nordic European countries). In these countries, it is usually considered that there is no minimum wage, that is, no national/regional wage floor. Very often the age of the beneficiary makes a difference; for example, a minimum wage set at a reduced rate for young people exists in Australia, Belgium, the Netherlands, and the United Kingdom. The minimum wage can be set on an hourly, daily, or monthly basis. Everywhere the public authorities govern the mode of its calculation, but it can also be bargained over between employers and employees. From one country to another, the minimum wage may be reset according to inflation (Belgium, France) and/or the evolution of the average wage (France, Japan, Spain), and sometimes even according to criteria thought to reflect the impact of the minimum wage itself on employment (the Netherlands, Spain, the United Kingdom). In the United States, minimal hourly wages are set by law at the federal and state levels, and there is no automatic indexation to inflation or the average wage. The upshot of these various regulations is wage floors that can range widely from one country to another and for various age groups, professions, regions, and sectors, a situation depicted in figure 12.16.

The amount of the minimum wage (the normal amount, if reduced rates exist) also varies significantly across countries. In order to make international comparison possible, the relative size of the minimum wage is often measured by the Kaitz index, which represents the ratio of the minimum wage to the average wage. Figure 12.17 gives the value of this index for 25 OECD countries.

Minimum wage levels are clearly set lower in the United States than they are in the other OECD countries. In the United States they reach just 28% of the average wage, compared with over 45% in Australia, France, and New Zealand. Another useful indicator, notably for employers, is to calculate the labor cost at the minimum wage, including employers' social security contributions. Figure 12.18 shows that the cost per hour, in U.S. dollars, is around or above $15 in Australia, about $14 in France and the Netherlands, and around or below $10 in the United States and the United Kingdom. These differences in the cost of labor reflect in part variation in the productivity of

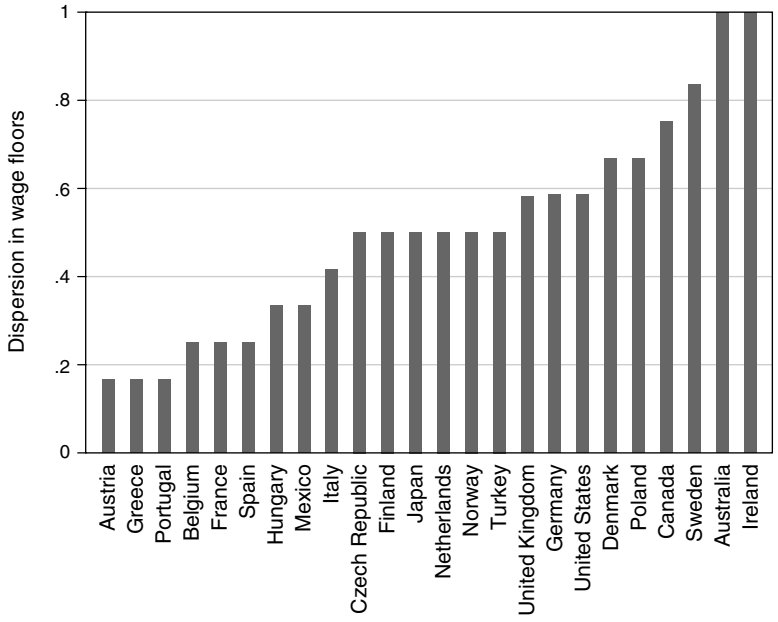


FIGURE 12.16

Degree of dispersion in wage floors by ages, qualifications, regions, sectors, or occupations, 1980–2000. The indicator of dispersion in wage floors measures the extent of dispersion and derogations in minimum wage setting. Minimum wage can differ by ages, qualifications, regions, sectors, or occupations. A more constraining minimum wage legislation is one that leaves little room for derogations and dispersion. This characteristic is measured by constructing two subindexes for age dispersion and other kinds of derogations. The subindexes are ranked between 0 and 1, a higher score indicating that the country provides more derogation. The subindex of dispersion across ages is constructed as follows. The score is equal to 0 if there is no provision at all for subminimum wages. It is equal to 0.5 if derogations are restricted to workers younger than 18 years old or if the derogation is less than half the official minimum wage. And it takes on the value 1 if the derogations can be extended to people older than 18 years or/and if the subminimum wages are lower than half the standard wage floor. The subindex for other derogations equals 0 if the minimum wage is allowed to differ along at least the three dimensions of regions, sectors, and occupations, 0.67 if there are two types of distinctions, 0.33 for one type of distinction, and 0 if no dispersion at all is allowed. The indicator of dispersion in wage floors is the average of these two subindexes.

Source: Aghion et al. (2011).

labor: countries like Mexico, Chile, and Turkey, but also a number of Eastern European countries, feature labor costs below $5.

The evolution of the minimum wage has varied greatly from one country to another; in figure 12.19 it is shown for several OECD countries. France and Japan have seen the real value of their minimum wage rise constantly from the 1970s to the late 2000s. In France the purchasing power of the gross hourly minimum wage went from $4 to $10 (in constant 2011 U.S. dollars) between 1970 and 2012; in other words, it multiplied 2.5 times. In the Netherlands, Canada, and the United States, however, the real value of the minimum wage declined between the early 1980s and the mid-1990s and has not recovered since. In the United States the purchasing power of the hourly minimum wage was less in 2012 than it was in 1970, although it had been rising until

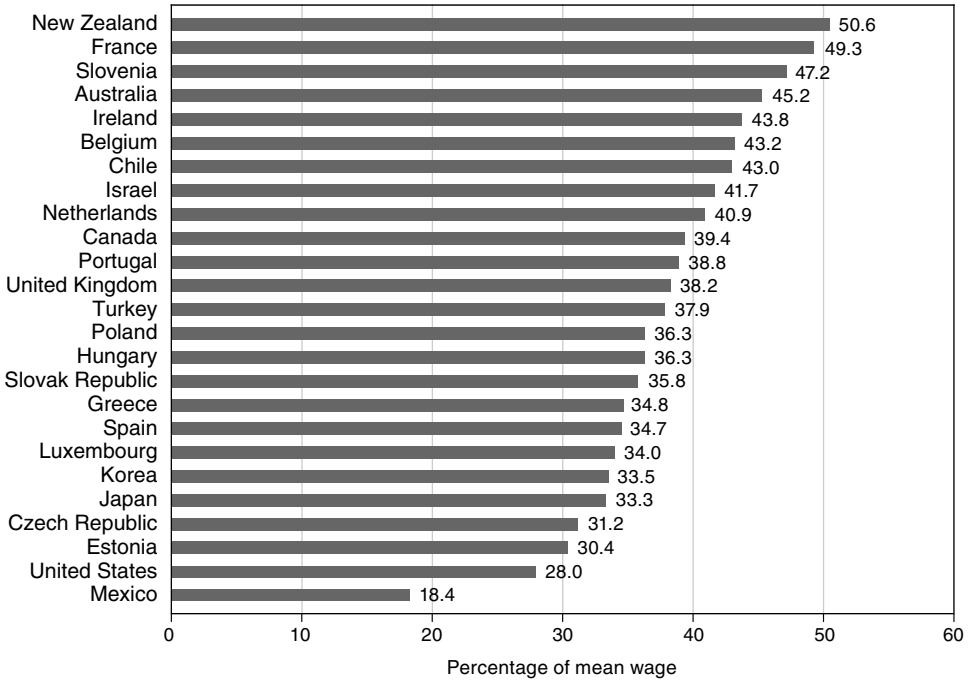


Figure 12.17

The Kaitz index (minimum wage as a percentage of mean wage), in 2011 in 25 OECD countries with national/regional minimum wages.

Source: OECD Earnings database.

the 1970s. In Canada the amount of the minimum wage was less in 2012 was quite close to what it was in the mid-1970s. The evolution of the ratio of minimum wage to average wage confirms this tendency, for this ratio has decreased in Canada, the Netherlands, and the United States, while in France it has stabilized at a relatively high level.

2.1.2 The Populations Concerned

The populations employed at minimum wage possess particular characteristics, which recur in all countries. Table 12.5 sets out some of these characteristics for Australia, France, the Netherlands, the United Kingdom, and the United States. In 2011 the proportion of workers being paid minimum wage was approximately twice as high in France as it was in the United States and the United Kingdom. The share of workers paid at the minimum wage yields an indication of how constraining the regulation is. In countries where a smaller share of workers are paid at the minimum wage, a larger share of jobs have productivities that allow remunerations above the wage floor. Even though the proportion of workers paid at the minimum wage differs across countries, the composition

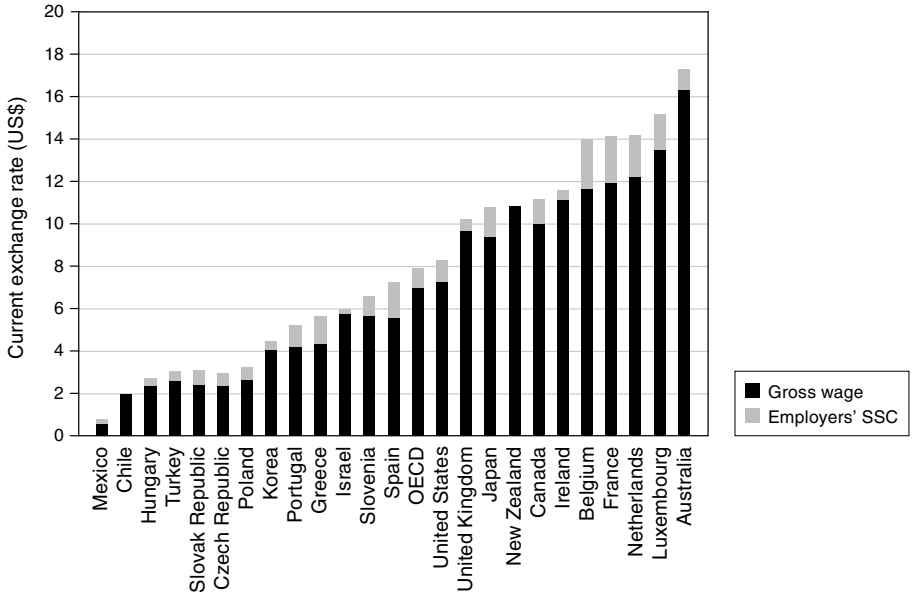


FIGURE 12.18
The labor cost at the minimum wage in 24 OECD countries in 2012.

Note: SSC = Social security contributions. OECD refers to the nonweighted average of labor costs among the 24 countries.

Source: Calculations from the OECD Earnings database. Data not available for Estonia.

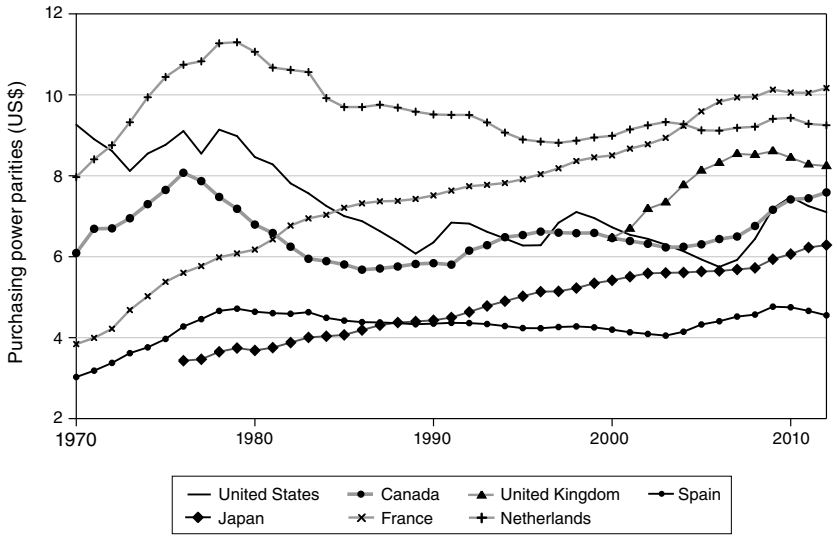


FIGURE 12.19
Change in real minimum wage (U.S., PPP) 1970–2012 in 7 OECD countries.

Note: Gross real hourly minimum wages are calculated first by deflating the series using the consumer price index taking 2011 as the base year. The series are then converted into a common currency unit (USD) using purchasing power parities (PPPs) for private consumption expenditures in 2011. The minimum wage was introduced in the United Kingdom on 1 April 1999.

Source: OECD Earnings database.

Table 12.5
Share of workers (in percentage) at the minimum wage for different types of labor force.

Country	Year	Total	Men	Women	Under 25*	Part-time	Retail
Australia	2010	4.1	4.3	3.9	10.2	—	4.0–7.7
France	2011	11.1	8.0	13.9	29.6	21.4	15.8
Netherlands	2005	4.0	3.2	4.9	12.8	4.9	—
United Kingdom	2011	4.4	3.5	5.0	12.5	9.5	9.0
United States	2011	5.2	3.9	6.4	13.1	12.9	7.7

Note: * Younger than 21 years for the United Kingdom. For the United States the statistics in this table relate only to workers who are paid hourly rates. Salaried workers and other workers who are not paid by the hour are not included, even though some have earnings that, if converted to hourly rates, would be at or below the minimum wage; consequently, the estimates presented in this table likely understate the actual number of workers with hourly earnings at or below the minimum wage. However, the degree of understatement is likely small (see www.bls.gov/cps/cpswom2011.pdf). France: End-2011 figure for the total, averages over 2010 for the breakdowns.

Source: Bureau of Labor Statistics for the United States, based on CPS data (www.bls.gov/cps/minwage2011tbls.htm), Low Pay Commission for the United Kingdom based on ASHE, considered by ONS to provide the best source of structural earnings information (www.lowpay.gov.uk/), DARES, French Ministry of Labor for France based on employer surveys Ecmoss and Acemo (http://travail-emploi.gouv.fr/IMG/pdf/2012-065.pdf and http://travail-emploi.gouv.fr/IMG/pdf/2012-095.pdf), Netherlands: Salverda (2008, table 2, p. 15), based on the structure of earnings survey from employers. Australia, based on the employers EEH survey considered to be more reliable on earnings than houshold surveys (Bray, 2013, tables 3 and 5, and figure 5), estimates for youth based on proportions observed in the HILDA household survey.

of the populations is much alike.[7] These are mainly persons without a secondary-school diploma or university degree, and the majority are women and youth. Almost 30% of those 25 and younger in France are paid at minimum wage—a reminder of the large dimensions the phenomenon assumes there. Workers paid at minimum wage are likewise overrepresented in the commercial field (especially retail and the hotel and restaurant trades) and in part-time jobs.

2.1.3 A First Look at the Macro Evidence

At first glance, countries featuring high minimum wages over the last 40 years also featured higher employment and lower unemployment both for the working-age population and for youth. This can be seen for youth over the 1970–2012 period in the two panels on the left-hand side of figure 12.20: here the averages of the minimum wage (in US$ and purchasing power parity) are represented on the vertical axis, and the employment rate (top left) and the unemployment rate (bottom left) on the horizontal axis. Now, it is very difficult to infer any causal relationship from these charts, which are purely descriptive. For one thing, both the minimum wage and labor market outcomes may have been influenced by the specificities of each economy over that period, especially the level of productivity and any policies influencing productivity. This is a classic

[7] Comparisons of minimum wage incidence across countries are difficult to carry out because many surveys are imprecise about exact levels of remuneration. Surveys for which the source is employers' declarations are usually considered more accurate.

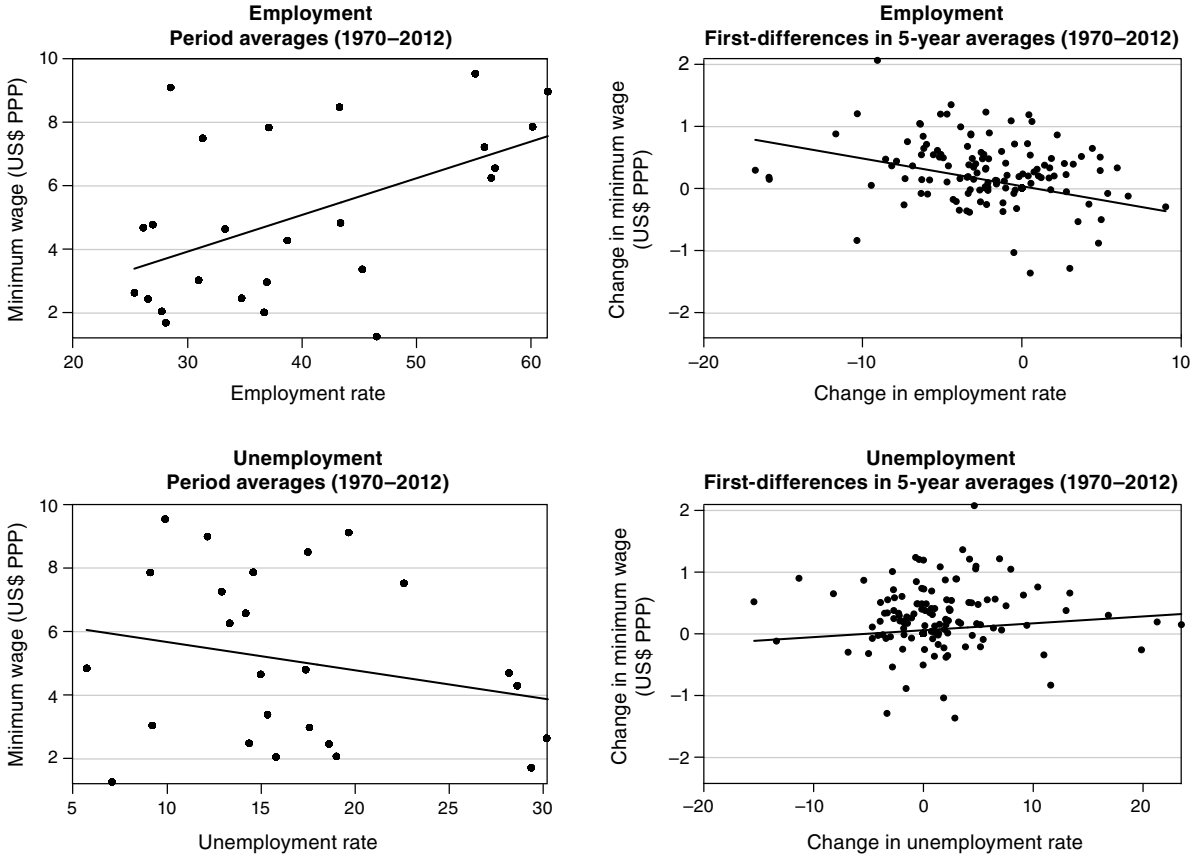


Figure 12.20
Unemployment, employment, and minimum wage for youth (15–24) in 25 OECD countries.

Note: The minimum wage is the gross wage in real terms, in US$ at purchasing power parity (PPP).

Source: OECD Labor Force and Earnings databases.

problem of simultaneity, and we do not control for these specificities here. For another thing, in countries with high levels of employment (or low levels of unemployment), governments may feel able to afford more generous income policies, such as a high minimum wage. This is the classic problem of reverse causality. Actually, when we turn to a longitudinal analysis, the relationships are reversed, as shown on the right-hand side of figure 12.20, where the first-differences in the 5-year averages of the minimum wage, employment, and unemployment in the same countries are represented. Using 5-year averages minimizes the interference of short-run variations in GDP on the relationship. What we see is that increases in the minimum wage are associated with decreases in the employment rate and slight increases in the unemployment rate for youth. The relationship is stronger with the employment rate. If we regress changes in employment or unemployment rates on changes in minimum wage, using a fixed-effect model to account for country unobservable characteristics, we find that on average a $1

increase in the hourly minimum wage decreases employment by 3 points and increases unemployment by 1 point, and these estimates are significant at the 1% level.

Again these elasticities are very large and this type of evidence should be approached very cautiously because we would need to control for omitted factors and reverse causality issues. These limitations mean that these correlations cannot be interpreted in terms of a causal impact of the minimum wage. We will see below how labor economists deal with this issue.

2.2 Minimum Wage and Employment

The effects of the minimum wage depend on the characteristics of the labor market to which it applies. The model of the perfectly competitive labor market highlights the negative aspects of the minimum wage for employment. In this setting, the minimum wage always reduces labor demand and then employment and production if it is set above the wage that would have prevailed in equilibrium. However, other theoretical setups, like the monopsony model and the search and matching model, highlight situations in which a rise in the minimum wage does lead to an increase in hiring. More precisely, the analysis of a labor market with frictions suggests that the effects of the minimum wage on employment actually depend on the initial level of the minimum wage. When it is set relatively low to start with, subsequent increases are not necessarily unfavorable to overall employment and production. But if the minimum wage is set relatively high to start with, subsequent increases likely do exert a negative impact on hiring. These analytical results receive a degree of confirmation from empirical research.

2.2.1 What the Monopsony Model Tells Us

In a situation of monopsony, there is a sole purchaser, in this case the employer, confronting many vendors, here the wage earners, who are selling their labor services. On a monopsonistic market, the workers cannot drive firms into competition and so cannot obtain a remuneration equal to their marginal productivity. The monopsonist profits from her privileged situation to impose a wage inferior to marginal productivity. The monopsony model goes back to the work of Robinson (1933) and Stigler (1946). More recently, Manning (2003) has stressed its importance for better understanding how the labor market functions. We will start by describing the functioning of a monopsonistic market. Then we will establish that the existence of monopsony power is only conceivable in the presence of some kind of barrier that prevents others from gaining access to the labor market that it dominates.

The Monopsony Model

The simplest model has a firm employing a number L of workers and using a technology represented by an increasing and concave production function $F(L)$. Labor supply, denoted $L^s(w)$, is taken to increase with respect to the wage w.

In this setting, when the firm decides to pay wage w, it knows that its level of employment will be $L^s(w)$; its profit is then written:

$$\Pi(w) = F[L^s(w)] - wL^s(w)$$

The equilibrium values w^M and L^M of the wage and employment are found by maximizing this expression of profit with respect to w. We arrive at:

$$F'(L^M) = w^M\left(1 + \frac{1}{\eta_w^L}\right) \quad \text{and} \quad L^M = L^s(w^M) \tag{12.29}$$

In this relation, the positive quantity $\eta_w^L = wL^{s\prime}(w)/L^s(w)$ designates the wage elasticity of labor supply. Equation (12.29) conveys the usual equality between the marginal productivity of labor and the marginal cost of this factor. In a monopsony situation, this marginal cost is higher than the wage because the elasticity of the labor supply with respect to this variable is positive. A monopsony pays the marginal employee at a level beneath his productivity; that is how the monopsony's gain is realized. This result also means that in the (L, w) plane, the curve with equation $F'(L) = w\left[1 + \left(1/\eta_w^L\right)\right]$ is situated below the labor demand curve $L^d(w)$ defined by $F'(L) = w$. Since employment is determined by the labor supply, the wage paid by the monopsony is below the competitive wage w^c that would equalize labor supply $L^s(w)$ with the labor demand $L^d(w)$ issuing from firms in a competitive market. This is the situation portrayed in figure 12.21.

Minimum Wage in the Monopsony Model

Knowing the labor supply that it faces, the monopsony affects the equilibrium wage directly by deciding on its volume of hires. If the supply of labor swells as wages rise, the monopsony is given an incentive to restrict its hires so as to get the benefit of low wages. In this context, a monopsonist firm chooses the lowest wage that lets it attract a number of workers sufficient to reach the desired output at minimal cost.

Robinson (1933) and Stigler (1946) had already noted that, in this setting, there is a theoretical possibility that a wage rise is accompanied by a rise in employment. Figure 12.21 does indeed indicate that if the minimum wage $\bar{w}$ lies somewhere between the wage w^M chosen by the monopsony and the competitive wage w^c, any rise in $\bar{w}$

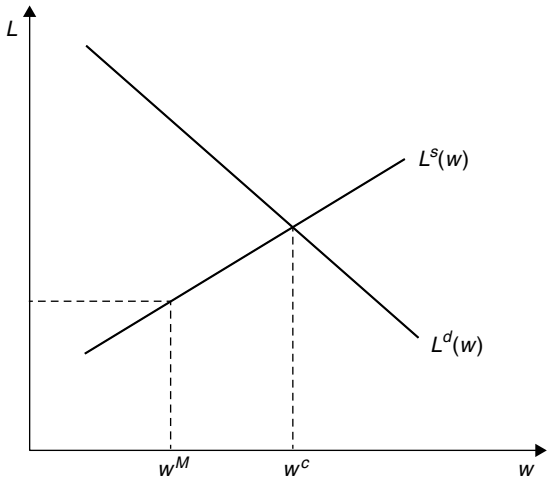


FIGURE 12.21
Employment and wage in the monopsony model.

allows the level $L^s(\bar{w})$ of employment to be increased. If, however, the minimum wage is greater than the competitive wage, the impact on employment is evidently negative, since no firm would consent to operate at wage levels that would bring it only losses. So the monopsony model suggests that the relationship between the minimum wage and employment is not monotonic, but increasing for low values of the minimum wage and decreasing for higher ones. The minimum wage may therefore affect employment positively in certain markets and negatively in others.

Thus the monopsony model brings out the possibility of a nonmonotonic relationship between the minimum wage and employment. The importance of this possibility should however be kept in perspective, for at least three reasons. In the first place, pure monopsony situations such as the one that has just been considered are very uncommon; they occur principally in specific geographical areas where mobility is low and the number of firms small. In the second place, the minimum wage acts positively on employment only when it lies *beneath* the competitive wage, in other words for wage levels probably a lot lower than those that exist in many European countries. Finally, the impact on employment of a rise in the minimum wage is all the stronger, the greater the wage elasticity of the labor supply. But as we saw in chapter 1, labor supply has little elasticity on average.

A number of studies have enriched the monopsony model by giving different foundations to the labor supply function. If manpower is mobile, for example, and information costly, workers sometimes have an interest in refusing job offers when the wage is too low, since they may hope to obtain other and better offers. The firm must then choose a wage level that allows it to attract a sufficient number of workers, in order to minimize hiring and firing costs. The work of Burdett and Mortensen (1998) and Masters (1999) has developed this idea. Drazen (1986) and Rebitzer and Taylor (1995) have proposed variants of the monopsony model grounded in the theory of efficiency wage. The former focus on problems linked to the quality of workers and the latter on checking up on what is actually getting done. Starting with the efficiency wage model of Shapiro and Stiglitz (1984), Rebitzer and Taylor (1995) assume that the probability of checking up on what an employee is accomplishing diminishes as the size of the firm's workforce grows. This hypothesis entails an increasing relation between employment and wages, for the latter rise when the probability of effective supervision falls (see chapter 6, section 4.3). Employers would then have an incentive to limit employment in order to keep wage costs down. In this setting, the minimum wage might have a positive impact on employment. Manning (1995) offers a systematic analysis of different efficiency wage models and shows that there are many cases in which the minimum wage exerts a positive effect on employment.

In the matching model developed in chapter 9, firms also have some monopsony power, since the employees are paid below their marginal productivity. Under the circumstances, we can expect that simple enrichments of the basic search and matching model will help to explain a positive linkage between the minimum wage and employment.

2.2.2 Minimum Wage in Labor Markets with Frictions

The search and matching model of chapter 9 is useful for analyzing the impact of the minimum wage on employment and on unemployment. As regards employment, we show that the matching model leads to conclusions similar to those of the monopsony

model: minimum wage hikes can increase employment if the minimum wage is sufficiently low, but they systematically decrease employment when the minimum wage surpasses a certain threshold.

An upward reset of the minimum wage increases the gap between the expected gains of employed and jobless persons. Thus it may provide an incentive for the latter to search harder for work, increase the exit rate from unemployment, and so help to lower unemployment. Obviously, the minimum wage also exerts a negative impact on employment because it raises the cost of labor. Taking job search effort into account suggests that, overall, the minimum wage has effects on unemployment that run counter to one another. The matching model allows us to shed light on the impact of the minimum wage in this context.

In what follows, we start by studying the impact of the minimum wage on welfare and labor market participation. We then analyze the impact of the minimum wage on job search effort.

The Influence of the Minimum Wage on Welfare and Labor Market Participation

Let us assume that decisions to participate in the labor market result from a trade-off between being an unemployed job seeker and not participating at all: then any improvement in the welfare of the unemployed leads to an increase in participation. Let H be the cumulative distribution function of the utilities expected outside the labor market by the entire working-age population. All the individuals whose expected utility outside the labor market is less than the expected utility of an unemployed person V_u decide to participate in the labor market, which entails that the participation rate is equal to $H(V_u)$. As H is necessarily an increasing function, the participation rate increases with the expected utility of unemployed persons.

Let us recall that in the basic search and matching model, the expected utility of unemployed persons V_u and that of employed persons V_e are defined by the two equations:

$$rV_e = w + q\,(V_u - V_e) \tag{12.30}$$

$$rV_u = z + \theta m(\theta)\,(V_e - V_u) \tag{12.31}$$

where w denotes the wage, r the interest rate, z the instantaneous income of unemployed persons, q the exogenous job destruction rate, θ the labor market tightness, and $\theta m(\theta)$ the job finding rate.

Profits Π_e and Π_v respectively expected from a filled job and a vacant one are written:

$$r\Pi_e = y - w + q(\Pi_v - \Pi_e) \quad \text{and} \quad r\Pi_v = -h + m(\theta)(\Pi_e - \Pi_v) \tag{12.32}$$

where h designates the cost of a vacant job, and y productivity.

When the free entry condition $\Pi_v = 0$ is satisfied, these two equalities yield the following relationship between w and θ, which is interpretable as a labor demand:

$$\frac{h}{m(\theta)} = \frac{y - w}{r + q} \tag{12.33}$$

If w represents a minimum wage that applies to all workers, this equation completely determines the equilibrium value of the labor market tightness θ. As we have $m'(\theta) < 0$ and $[\theta m(\theta)]' > 0$, it results that the labor market tightness θ is a decreasing function of the minimum wage w, and so is the job finding rate $\alpha = \theta m(\theta)$. A hike in minimum wage degrades the profitability of a job, so firms post fewer vacancies and the job finding rate falls off.

We will now observe that maximization of the expected utility of an unemployed person with respect to wage w, subject to the labor demand (12.33) constraint, gives labor tightness identical to that obtained at the outcome of decentralized wage bargaining for which the bargaining power of workers, measured by the share γ of the surplus they get, is equal to the elasticity $\eta(\theta) = -\theta m'(\theta)/m(\theta)$ of the matching function with respect to the unemployment rate. In chapter 9, section 3.4, we showed that the equilibrium value of labor market tightness in a decentralized economy where wages are bargained over is defined by the equation:

$$\frac{(1-\gamma)(y-z)}{r+q+\gamma\theta m(\theta)} = \frac{h}{m(\theta)} \tag{12.34}$$

Let us write the expected utility of unemployed workers, using equations (12.30) and (12.31), as follows:

$$rV_u = \frac{(r+q)z + \theta m(\theta) w}{r+q+\theta m(\theta)}$$

Then, using the labor demand (12.33) to eliminate w from this expression, we get:

$$rV_u = \frac{\theta m(\theta) y + (r+q)z - \theta(r+q)h}{r+q+\theta m(\theta)}$$

The value of labor market tightness that maximizes rV_u satisfies the first-order condition $\partial rV_u/\partial\theta = 0$, which can be written:

$$\frac{[1-\eta(\theta)](y-z)}{r+q+\eta(\theta)\theta m(\theta)} = \frac{h}{m(\theta)} \tag{12.35}$$

Comparison of equations (12.34) and (12.35) shows that the expected utility of unemployed workers is maximized when the minimum wage is set at a level that corresponds to the wage level of the decentralized economy in which the bargaining power parameter satisfies the Hosios condition. This result is illustrated in figure 12.22, where the wage that emerges from decentralized equilibrium gives unemployed persons a maximal expected utility only if the Hosios condition ($\gamma = \eta(\theta)$) is satisfied. The level of the negotiated wage when the Hosios condition is met is denoted by w^*.

If $w < w^*$, we see that any increase in the minimum wage increases participation, equal to $H(V_u)$, *and* the unemployment rate, and that it has an ambiguous impact a priori on employment, equal to $H(V_u)(1-u)$. In consequence, when the bargaining power of workers is too low to satisfy the Hosios condition ($\gamma < \eta(\theta)$), an increase in the minimum wage improves the welfare of the unemployed. As the welfare of the unemployed reaches a maximum when the Hosios condition is fulfilled, this remark implies that

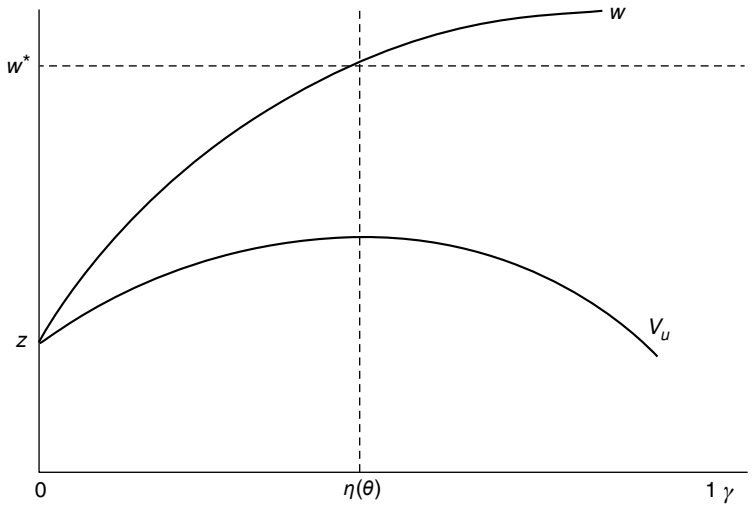


FIGURE 12.22
The wage and the expected utility of unemployed workers as functions of the bargaining power parameter in the search and matching model.

minimum wage hikes can improve labor market efficiency. (Flinn, 2006, 2010, reaches the same conclusion in a job matching model estimated for young labor market participants in the U.S. economy.)

On the other hand, if $w \geq w^*$, any increase in the minimum wage entails a decline in labor market participation (because V_u decreases) and an increase in unemployment, which necessarily leads to a fall in employment.

The Influence of the Minimum Wage on Job Search Effort and Unemployment

We have shown that the minimum wage can increase labor market participation in the search and matching model when the bargaining power of workers is small relative to the elasticity of the matching function with respect to unemployment. However, in the model just presented, the minimum wage always increases unemployment. This result does not necessarily hold when the search effort of workers is endogenous. For the sake of simplicity, this result is shown in a model where the arrival rate of job offers is exogenous.

We consider the model with endogenous job search effort studied in chapter 5, section 2.2.3. In this model, the intensity of the job search is designated by the scalar e, which can be interpreted as the amount of time and/or the intensity of the effort devoted to search. The notion that more job offers should result from greater effort devoted to search amounts to postulating that the rate at which offers arrive increases with e**.** For the sake of simplicity and without loss of generality, we postulate that the arrival rate of job offers is αe and we assume that the wage distribution is degenerated to a single wage, denoted by w. The parameter $\alpha > 0$ we interpret as an indicator of the state of the labor market, independent of individual efforts. We denote by $\phi(e)$ the cost arising from the search effort e, with $\phi' > 0$ and $\phi'' > 0$. So the instantaneous utility of a job

seeker will be written $z - \phi(e)$. Jobs are destroyed at rate q. In this setting, the expected discounted utilities of a job seeker and a job holder, respectively denoted V_u and V_e, verify equations:

$$rV_e = w + q(V_u - V_e) \tag{12.36}$$

$$rV_u = \max_e \; z - \phi(e) + \alpha e(V_e - V_u) \tag{12.37}$$

The optimal search effort is such that the marginal cost of performing the search is equal to its marginal return:

$$\phi'(e) = \alpha(V_e - V_u)$$

This equation indicates that the optimal search effort increases as the difference grows between the expected utility of a job holder and that of a job seeker. As this difference itself grows with w,[8] a wage increase drives up job search effort and thus the job finding rate.

At stationary equilibrium, the unemployment rate u is found by equalizing the flow of entries into and exits from unemployment. Assuming that the labor force is of constant size normalized to 1, the number of jobs destroyed per unit of time is equal to $(1 - u)q$. The exit rate from unemployment being equal here to $e\alpha$, the number of jobs created per unit of time takes the value $ue\alpha$. Equalization of the flows of entries into and exits from unemployment then yields the stationary value of the unemployment rate u as a function of the equilibrium values of e:

$$u = \frac{q}{q + e\alpha} \tag{12.38}$$

We see that a hike in the wage, which increases the search effort and the job finding rate, decreases the unemployment rate. Hence a hike in the minimum wage may, by boosting the search effort of job seekers, boost employment, since it increases the rate of return to holding a job. For their part, firms may have an interest in hiring workers as long as their wage is equal to their productivity. We have seen, in the equilibrium model of job search, that wages in general lie well below productivity. It is thus possible to increase wages beyond their equilibrium value while permitting firms to continue to make profits, as in the monopsony model. Nonetheless there exists, again as in the monopsony model, an upper bound to wage growth, past which wages surpass marginal productivity and the firm no longer has an interest in keeping its jobs (van den Berg and Ridder, 1998; Masters, 1999; and see Flinn, 2010, for a thorough analysis of the minimum wage in a labor market with frictions).

[8] More precisely, we can write on the basis of (12.36) and (12.37):

$$(r + q)(V_e - V_u) = w - \left[\max_e z - \phi(e) + \alpha e(V_e - V_u)\right]$$

Differentiating this equation with respect to w, we get:

$$\frac{\mathrm{d}(V_e - V_u)}{\mathrm{d}w} = \frac{1}{r + q + \alpha e} > 0$$

All in all, the analysis of the consequence of the minimum wage in labor markets with frictions shows that the minimum wage can improve employment and decrease the unemployment rate when the minimum wage is sufficiently low. However, a high minimum wage is detrimental to employment and increases the unemployment rate.

2.3 The Employment Impact of the Minimum Wage in Light of Empirical Research

Different approaches are used to assess the impact of the minimum wage on employment. One approach consists of analyzing the correlations between employment and the minimum wage on the basis of time series. Other approaches analyze the outcome of "natural experiments" with a difference-in-differences estimator through a comparison of the employment of workers affected by a change in minimum wage and the employment of similar workers who have not been affected by this change.

2.3.1 Time Series Studies

Until the end of the 1990s, a large majority of empirical studies adopted a methodology which consists of bringing out possible correlations between variations in employment and the minimum wage while controlling for the other factors that might affect employment. These studies make use of the temporal evolution (or "time path") of the minimum wage, as well as differences in its level as between industries and/or geographical regions. They generally conclude that the minimum wage has a negligible impact on employment, except perhaps for youth employment. For example, the OECD study (1998, chapter 2) of nine countries (Belgium, Canada, France, Greece, Japan, the Netherlands, Portugal, Spain, and the United States) for the period 1975–1996 finds that a rise of 10% in the minimum wage entails a fall of between 2% and 4% in employment among *those under 20 years old*. This conclusion echoes that of Brown et al. (1982), who published a review of existing research on the employment effects of the minimum wage in the United States and argued that "time-series studies typically find that a 10 percent increase in the minimum wage reduces teenage employment by one to three percent." Bazen and Marimoutou (2002) reach similar conclusions by reporting statistically significant negative effects of the minimum wage on teenage employment on U.S. data over the period 1954–1999, with an elasticity of $-.11$ in the short run and $-.27$ in the long run. On the other hand, the minimum wage is shown to have no effect on the employment of workers 25 years of age and older. Dolado et al. (1996) come to the same type of conclusion for the European Union countries, suggesting that the minimum wage reduces youth employment but increases total employment, while pointing out that the dimensions of this effect are slight. Bassanini and Duval (2006) find no significant impact of the minimum wage on unemployment or employment, but their estimates suggest that a high tax wedge has greater adverse effects on unemployment when the minimum wage is high and prevents wage adjustments.

2.3.2 Studies Based on Natural Experiments

We saw in previous chapters that the method of natural experiments consists of exploiting exogenous changes in the economic environment of certain agents in order to compare their reactions to those of other (in principle identical) agents who have not

undergone these changes. This method has been used since the early 1990s to study the impact of minimum wage on employment. One approach consists of identifying geographic zones where the minimum wage varies differently for reasons independent of changes in employment. We rely on the paper of Card and Krueger (1994) to present this approach. Databases and programs allowing readers to replicate the main results of this contribution are available at www.labor-economics.org. Another approach consists of comparing the employment trajectories of individuals directly affected by hikes in the minimum wage with those of individuals earning remunerations slightly above the minimum, who are therefore not directly affected when it is raised.

The Impact of the Minimum Wage in the Fast-Food Industry

In a well-known paper, Card and Krueger (1994) studied the impact of increases in the minimum wage in New Jersey in 1992. From 1 April 1991, the minimum wage was $4.25 per hour in the neighboring states of Pennsylvania and New Jersey. In New Jersey the minimum wage increased to $5.05 per hour on 1 April 1992 but remained unchanged in Pennsylvania. The 1992 increase gave New Jersey the highest state minimum wage in the country. Card and Krueger analyzed the effect of the minimum wage on 410 employment outcomes in fast-food restaurants in New Jersey (the treatment group) and eastern Pennsylvania (the control group).

The choice of the fast-food industry was driven by several factors related to employment and wages. First, fast-food outlets are a leading employer of low-wage workers. Second, fast-food outlets comply with minimum wage regulations. Third, the job requirements and products of fast-food restaurants are relatively homogeneous, making it easier to obtain reliable measures of employment, wages, and product prices. Fourth, the absence of tips greatly simplifies the measurement of wages in the industry. Moreover, since New Jersey is a relatively small state with an economy that is closely linked to nearby states, the control group of fast-food outlets in eastern Pennsylvania forms a relevant basis for comparison with the experiences of restaurants in New Jersey.

Card and Krueger conducted two surveys by telephone. In late February and early March 1992, a little over a month before the scheduled increase in New Jersey's minimum wage, 410 restaurants out of the 473 present in the sample responded to the interviews. The second survey was conducted in November and December 1992, about eight months after the minimum wage increase. Only the 410 outlets that responded in the first survey were contacted in the second round of interviews.

Figure 12.23 displays the distribution of starting wages by restaurant in Pennsylvania and in New Jersey in February–March 1992, before the wage hike in New Jersey. It is apparent that many outlets had starting wages between $4.25 and $5.05 in both states. Both states also had a large share of businesses where the starting wage equaled the minimum wage, amounting to $4.25 in that period. In New Jersey, 91% of restaurants had starting wages below $5.05. The corresponding figure was 94% in Pennsylvania. In November–December 1992, following the minimum wage hike to $5.05 in New Jersey, figure 12.24 shows that no restaurant had a starting wage below that value in New Jersey[9] but that most restaurants (90%) still had starting wages below

[9] Actually, the starting wage is equal to $5.00 in 1 restaurant in the second survey in New Jersey.

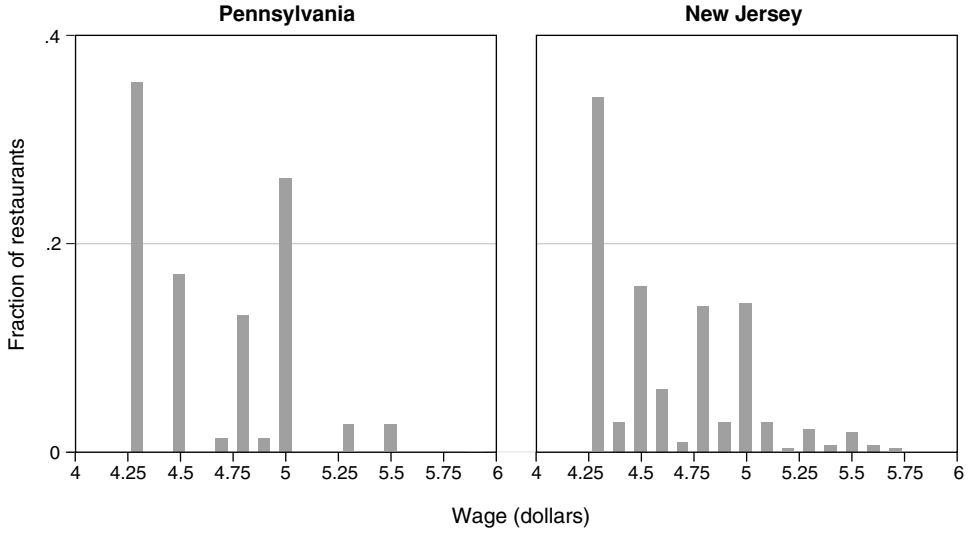


FIGURE 12.23
Distributions of starting wages in February–March 1992.

Source: Card and Krueger (1994).

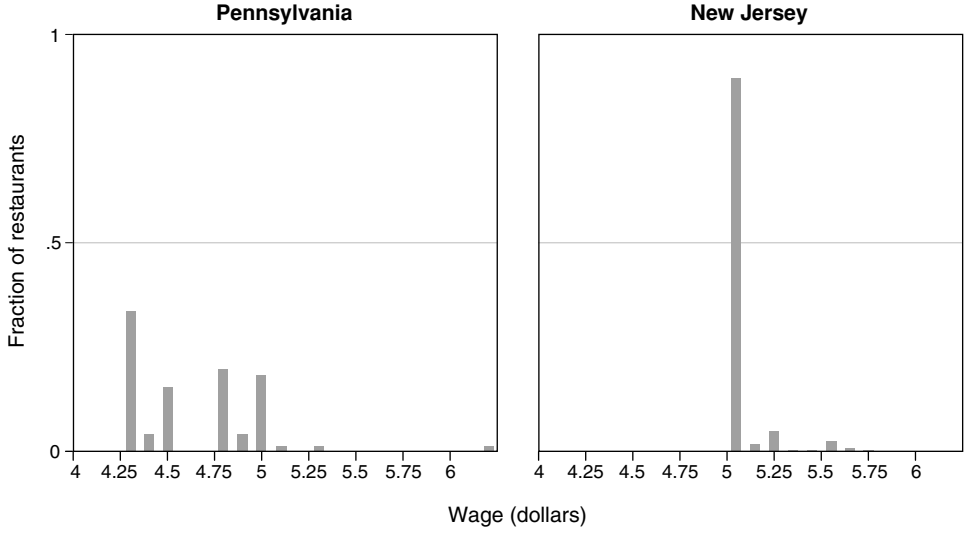


FIGURE 12.24
Distributions of starting wages in November–December 1992.

Source: Card and Krueger (1994).

that value in Pennsylvania. Moreover, the new minimum wage appears to have been very binding in New Jersey, since the starting wage equals \$5.05 in 89% of restaurants across the whole state. Thus the hike in the minimum wage in the state of New Jersey had a very significant impact on hiring wages.

TABLE 12.6
Average employment per outlet and by state before and after the rise in the New Jersey minimum wage.

	Pennsylvania (PA)	New Jersey (NJ)	Difference NJ-PA
Employment before	23.63 (1.50)	20.51 (0.54)	−3.11 (1.34)
Employment after	21.50 (1.04)	20.71 (0.54)	−0.78 (1.22)
Change in mean employment	−2.13 (1.38)	0.20 (0.48)	2.33 (1.19)
N	67	290	

Notes: Standard errors are shown in parentheses. The sample consists of all outlets with available data on employment and wages in both surveys. Employment is full-time equivalent; employment counts each part-time worker as half a full-time worker. Employment at six closed outlets is set to zero. Employment at four temporarily closed outlets is treated as missing.

Source: Card and Krueger (1994) data set.

TABLE 12.7
Average employment per outlet in New Jersey before and after the rise in the New Jersey minimum wage.

	Low wage = \$4.25	High wage ≥ \$5.00	Difference low − high
Employment before	19.44 (0.78)	22.17 (1.19)	−2.72 (1.37)
Employment after	20.65 (1.04)	20.01 (1.05)	0.64 (1.51)
Change in mean employment	1.20 (0.83)	−2.16 (1.01)	3.36 (1.29)
N	94	67	

Notes: Standard errors are shown in parentheses. The sample consists of all outlets with available data on employment. Employment is full-time equivalent; employment counts each part-time worker as half a full-time worker. Employment at closed outlets is set to zero. Employment at temporarily closed outlets is treated as missing.

Source: Card and Krueger (1994) data set.

Table 12.6 describes the path of average employment by restaurant in the two states. Average employment shrank in both, because the American economy was in recession.[10] But employment shrank less in restaurants in New Jersey than it did in Pennsylvania restaurants. If the increase in minimum wage in New Jersey in April 1992 is the only significant event that exerted an uneven effect on the functioning of the labor markets of the two states between February and December 1992, this result means that the minimum wage increase had a positive effect on employment. Still, there might exist unobservable events that affected the two states differently. To take this possibility into account, Card and Krueger compare the path of employment in the New Jersey restaurants that already had a hiring wage equal to or higher than \$5.00 at the outset, and those where the hiring wage lay below this threshold and were thus constrained to raise their wages. If the hike in minimum wage did have a positive impact on employment, we ought to observe an increase in employment in the restaurants where the hiring wage was initially low relative

[10] The figures are slightly different from those of Card and Krueger (1994, table 3) because the sample consists of all outlets with available data on employment and wages in both surveys, whereas that of Card and Krueger consists of all outlets with available data on employment in both surveys. We made this choice in order to retain the same sample throughout our analysis.

to those where the hiring wage was equal to or higher than \$5.00. Table 12.7 shows that this is indeed the case: employment in the restaurants with lower hiring wages grew by 3.36 workers relative to those with higher hiring wages. This result is significantly different from zero at the 5% threshold. Of course this difference in employment paths might derive from the fact that, as a general rule, growth in employment should be stronger in the restaurants where the hiring wages were lower. To test that possibility, the difference in the path of employment between places with lower and higher hiring wages in Pennsylvania may be compared with the same difference in New Jersey. If relative employment in low-hiring-wage restaurants increases more than relative employment in high-hiring-wage restaurants in both Pennsylvania and New Jersey between February–March and November–December 1992, that means that the results presented in table 12.7 have little probative value. In fact, though, employment in low-hiring-wage restaurants did not rise significantly as compared to employment in high-hiring-wage restaurants in Pennsylvania. To be precise, the difference is positive, equaling 1.69 (as against 3.36 in New Jersey), but with a standard deviation of 2.97, which means that the difference is not significantly different from zero in Pennsylvania at the 10% threshold. Still, the fact that this difference is not significantly different from zero might derive from the small number of observations in Pennsylvania, where 23 high-hiring-wage outlets and 26 low-hiring-wage outlets were observed (whereas in New Jersey the figures were 67 and 94, respectively, as table 12.7 shows).

Card and Krueger then estimate a linear model to take into account variables that might have influenced the path of employment. The difference-in-differences estimator based on comparison of the paths followed by employment in New Jersey and Pennsylvania respectively is obtained on the basis of the equation:

$$\Delta \ell_i = a + \mathbf{x}_i \mathbf{b} + cJ_i + \varepsilon_i \tag{12.39}$$

where $\Delta \ell_i$ designates the variation in employment in restaurant i between February–March and November–December 1992, $\mathbf{x}_i$ is a vector of the characteristics of restaurant i comprising dummy variables for chain type and whether or not the outlet is company owned, and dummy variables for two regions of New Jersey and two regions of eastern Pennsylvania, J_i is a dummy variable equal to 1 if restaurant i is located in New Jersey and equal to zero if not; and ε_i is an error term of zero average.

The difference-in-differences estimator based on comparison of the paths of employment in the low-hiring-wage and high-hiring-wage restaurants in the two states is obtained on the basis of equation:

$$\Delta \ell_i = a' + \mathbf{x}_i \mathbf{b}' + c' gap_i + \varepsilon_i' \tag{12.40}$$

The variable gap_i measures the constraint induced by the minimum wage and is defined by:

$$gap_i = \begin{cases} = 0 & \text{for outlets in Pennsylvania} \\ = 0 & \text{for outlets in New Jersey where} w_{i1} \geq \$5.05 \\ = \frac{\$5.05 - w_{i1}}{w_{i1}} & \text{for outlets in New Jersey where} w_{i1} < \$5.05 \end{cases}$$

where w_{i1} denotes the starting wage of outlet i in February–March 1992.

TABLE 12.8
Reduced form models for changes in employment.

	1	2	3	4	5
New Jersey dummy	2.33 (1.19)	2.30 (1.20)			
Initial wage gap			15.65 (6.08)	14.92 (6.21)	11.81 (7.39)
Control for chain and ownership	no	yes	no	yes	yes
Control for regions	no	no	no	no	yes

Note: Standard errors are given in parentheses. The sample consists of 357 outlets with available data on employment and starting wages in waves 1 and 2. The dependent variable in all models is change in full-time employment. All models include an unrestricted constant (not reported). Initial wage gap: Proportional increase in starting wage necessary to raise starting wage to new minimum rate. For outlets in Pennsylvania the wage gap is 0. Control for chain and ownership: Three dummy variables for chain type and whether or not the outlet is company owned are included. Controls for region: Dummy variables for two regions of New Jersey and two regions of eastern Pennsylvania are included.

Source: Card and Krueger (1994) data set.

Table 12.8 presents the results of these estimations. The first two columns give the results of the estimation of equation (12.39). The estimate in column (1) of table 12.8, which does not include any control, is directly comparable to the simple difference-in-differences of employment changes in column 3, row 3 of table 12.6. Column (2) of table 12.8 takes into account the control for chain and ownership. We see that the difference-in-differences estimator is not significantly altered. Columns (3) to (5) present the results of the estimation of equation (12.40), which includes the variable gap_i. These results indicate that the firms that felt the constraint of the minimum wage hike experienced a rise in employment relative to the firms that did not undergo this constraint, even when controls for chain and ownership and for regions are taken into account.

Overall, the results of Card and Krueger (1994) suggest that the increase in the minimum wage may have had a positive impact on employment when this wage was low to start with, as it was in New Jersey.

The Debate in the Wake of Card and Krueger's Paper

A debate arose in the wake of the study of Card and Krueger (1995). Essentially it bore on three points.

First, Kennan (1995) and Dolado et al. (1996) have emphasized that the interpretation of the results demands caution, inasmuch as consumers of fast food are not necessarily representative of the population as a whole. It is in fact probable that persons earning minimum wage patronize fast-food restaurants more frequently than those earning higher wages, and so, on the assumption that hamburgers, cheeseburgers, and carbonated soft drinks are normal goods, a higher minimum wage will increase the purchasing power of those who regularly consume them—and this in turn will entail a rise in production and employment in fast-food places, despite the increase in the cost of labor.

Second, there is the question of the adequacy of the control groups used in the studies. In particular, Deere et al. (1995) pointed out that teenage employment rates in New Jersey diverged significantly from those in Pennsylvania beginning in 1988, casting doubt on Card and Krueger's claim that Pennsylvania represents a sensible control

group with which to compare New Jersey. Dube et al. (2010) suggest that this issue is important in practice. They use policy discontinuities at all U.S. state borders between 1990 and 2006 to identify the effects of minimum wages on earnings and employment in restaurants and other low-wage sectors. Their approach generalizes the case study method by considering all local differences in minimum wage policies between 1990 and 2006. They compare all contiguous county-pairs in the United States that straddle a state border. This approach makes it possible to compare individuals belonging to either the treatment or the control group, but all living in areas with closely similar employment conditions, controlling for county-pair–specific time effects and county-fixed effects. They find strong earnings effects but no employment effects of minimum wage increases. In the same spirit, Allegretto et al. (2011) use information on state-level minimum wages and individual-level data on teens from the Current Population Survey (CPS) from 1990 to 2009. When they estimate a model that includes state- and period-fixed effects and other usual controls, they find a negative employment effect of minimum wages. But when they include state-specific linear trends, the estimated employment and hours elasticities become indistinguishable from zero. These results would seem to confirm the findings of Card and Krueger. However, the detailed analysis of the papers of Dube et al. (2010) and Allegretto et al. (2011) by Neumark et al. (2013) argues that the results of these papers rely on questionable choices of control groups (states or counties). In particular, Neumark et al. show that identifying minimum wage effects from the variations within contiguous cross-border county-pairs, or states in the same census division, does not isolate the most reliable information. For instance, it is possible that the minimum wage might have changed in the areas included in the control groups in a short time interval before or after the minimum wage changed in the treatment groups. Using the same data as those of Dubb et al. (2010) and Allegretto et al. (2011) with different choices of control groups, Neumark et al. (2013) find significant negative employment effects of the minimum wage.

Third, the reliability of the data used in these case studies has been questioned. For instance, Neumark and Wascher (2000) critique the data of Card and Krueger (1994), which derives from telephone interviews. Neumark and Wascher carry out the same exercise as Card and Krueger but utilize administrative payroll records for the same fast-food restaurants in the same states. Contrary to Card and Krueger, they find that the minimum wage reduced employment in New Jersey. Nonetheless, Card and Krueger (2000), this time using a larger sample of administrative payroll records than that of Neumark and Wascher, obtain results that confirm, to some extent, their earlier work.

The Impact of the Minimum Wage on the Transition Probabilities into and Out of Employment

Individual longitudinal data make it possible to follow the labor market histories of persons whose wages are at or close to minimum wage, and they have the advantage of registering the impact of changes to the minimum wage on the populations actually affected by this level of compensation. Studies exploiting this type of data find that changes to the minimum wage do have a significant effect on employment among this class of worker.

Studies grounded in individual longitudinal data have achieved more precise assessments of the effects of minimum wage. The comparative study of Abowd et al. (2000) on France and the United States is an illustration of this. It exploits the fact

that during the 1980s the minimum wage advanced in real terms in France, while it receded in the United States. For France, the authors analyze the histories of individuals whose current wage fell *below* the minimum wage in the interval between one increase in the minimum wage and the next. They show that such persons had a higher probability of losing their jobs than those whose wage was not overtaken by the minimum wage. For example, young people 21–25 years old whose wage was marginally higher than the latest value of the minimum wage (i.e., lying between minimum wage and 1.15 times minimum wage) had a probability of losing their jobs equal to 10%, whereas this probability rose to 16% for young people whose wage lay between the previous value of the minimum wage and the latest one. For the United States, this study looked at the outcomes of persons whose wage became *higher* than the minimum wage, as the latter gradually declined in real terms. They show that these individuals had a higher probability of keeping their jobs. To sum up, this study suggests that in France an increase of 1% in the minimum wage reduces the probability, among men receiving minimum wage, of keeping their jobs by 1.3%, while for women the figure is 1%. In the United States a reduction of 1% in the minimum wage increases the probability that workers paid at this level will keep their jobs by 0.4% for men and 1.6% for women.

The study of the French case by Kramarz and Philippon (2001) supplies further interesting results. It uses the same methodology but takes the *cost of labor* as the pertinent variable in trying to assess the impact of the minimum wage on employment. It estimates that an increase of 1% in the cost of jobs compensated at minimum wage entails a rise of 1.5% in the probability of job loss for workers who are being paid minimum wage.

Portugal and Cardoso (2006) arrive at different results using the same type of methodology. They exploit changes made in 1987 to Portuguese legislation regarding the minimum wage of young people 19 and under. The minimum wage was raised by 50% for youths aged 17 and by 33% for youths aged 18 and 19. Portugal and Cardoso find that these minimum wage hikes had a depressant effect on the hiring of this category of workers. But they also highlight a "supply effect": after the reform of 1987, young people 19 and under had a greater tendency to keep their jobs. Portugal and Cardoso observed fewer separations, which counteracted the fall in hires. This result, coherent with the prediction of the monopsony model and of the search model, probably reveals a greater attachment of youth to their jobs when wages improve. Overall, this research shows that the minimum wage can have significant effects on the probabilities of being hired and of losing a job. However, it does not invariably exert a positive effect on the probability of job loss among the populations whose livelihoods are directly dependent on this level of compensation.

What Is the Overall Employment Impact of the Minimum Wage?

On the whole, empirical analyses of the impact of the minimum wage on employment find results coherent with the predictions of the monopsony model and the equilibrium job search model. As Neumark and Wascher (2008) state in their review of empirical research, although the wide range of estimates is striking, out of 102 studies nearly two thirds give a relatively consistent (although by no means always statistically significant) indication that minimum wage exerts a negative effect on employment, while eight give a relatively consistent indication that its effect on employment is positive. Hence the

impact of a hike in minimum wage on employment is not univocal. The empirical studies available suggest that this impact may be positive if the minimum wage is low with respect to the median wage, but that it becomes negative when the minimum wage is high.

2.4 The Quality of Jobs

The minimum wage affects not just employment but also the kinds of jobs offered. That being the case, it may improve the allocation of resources by favoring the creation of more productive jobs.

The monopsony model and the search and matching model both reveal how complex the effects of the minimum wage are, as well as the idiosyncrasy of the competitive equilibrium model, with its conclusion that the minimum wage has a systematically negative impact on employment. Models built on different premises confirm this view. Jones (1987) looked at the impact of the minimum wage on a labor market in which "good" jobs requiring the accomplishment of complex tasks coexist with "bad" jobs, the results of which are perfectly verifiable. The workers with the good jobs, whose effort at work can only be observed imperfectly, receive an efficiency wage, while the ones with the bad jobs are paid at a lower rate, equal to their reservation wage. When a minimum wage lying somewhere between the reservation wage and the efficiency wage is introduced into this model, it reduces the efficiency wage and increases the number of good jobs opened up. In some circumstances, the increase in the number of good jobs even exceeds the decline in the number of bad ones, and that makes for an overall reduction in unemployment.

Substitution effects among different skill levels may also help to bring about a rising relation between the minimum wage and employment when compensations lying above minimum wage are bargained over. From this perspective, Cahuc et al. (2001) consider a model with skilled workers who bargain over their wage collectively and unskilled workers paid at the minimum wage. The impact of the minimum wage on the employment of the unskilled workers then depends on the elasticity of substitution between the two categories of worker. It results that an increase in the minimum wage can lead to increased global employment, including increased employment among the unskilled, for plausible values of the parameters of the model.

The minimum wage can improve global efficiency in other settings. Drazen (1986) assumes that workers and employers know the productivity of jobs imperfectly before hiring takes place. He also assumes that there is a positive linkage between the productivity of a worker and the compensation that he can obtain outside the labor market. In consequence, the payment of high wages makes it possible to attract good workers. If it is not possible for workers to look for a job while simultaneously receiving compensation outside the labor market, then an individual only decides to take part in the labor market if he will receive an expected gain that exceeds the compensation available outside the market. Obviously this expected gain increases with the average wage observed in the labor market. In this setting, the equilibrium is suboptimal, for single employers have no market power and therefore no capacity to affect the average wage: each has an individual interest in offering low wages. That being so, the introduction of a minimum wage makes it possible to attract high-productivity workers into the market and improve efficiency.

The effect of the minimum wage on the structure of employment has also been analyzed by Acemoglu (2001) in a matching model with good and bad jobs. The good

jobs have higher productivity, and cost more to create, than the bad ones. Wages, which firms and employees bargain over, are therefore higher for the good jobs. Acemoglu shows that decentralized equilibrium systematically leads to too few good jobs, and that introducing a minimum wage slightly higher than the lower bound of the distribution of wages makes it possible to improve welfare thanks to an increase in the number of good jobs. Cahuc and Michel (1996) obtain the same type of result in a model of endogenous growth in which the introduction of the minimum wage improves welfare by giving individuals an incentive to accumulate human capital, which favors growth.

2.5 The Minimum Wage and Inequality

A rise in the minimum wage has opposite effects on income inequality; the latter is generally measured by the standard deviation of the logarithm of incomes, or by the ratios between the average values of different deciles of the overall income distribution. On one hand, the minimum wage allows some people to receive a higher wage, and this favors the reduction of inequality among employees. But on the other, it can also destroy jobs, which leads to reduced incomes for those who would have been able to find a job in the absence of the minimum wage.

Empirical research generally concludes that the minimum wage does make it possible to reduce wage inequality (Brown, 1999). The contributions of DiNardo et al. (1996) and Lee (1999) suggest that the fall in the real value of the minimum wage contributed strongly to increasing wage inequality in the United States in the 1980s. DiNardo et al. (1996) look at the evolution of the distribution of men's and women's wages between 1979 and 1988, finding that the fall in the minimum wage in real terms explains one quarter of the rise in the standard deviation of the distribution of men's wages, and 30% of that for women. Lee (1999) for his part estimates that the shrinking minimum wage over this period explains 70% of the increase in the ratio of average fifth-decile wages to average first-decile wages. So changes in the minimum wage have had a significant impact on wage inequality in the United States.

In theory, increases in minimum wage have an ambiguous impact on the poverty rate, which is measured by the proportion of individuals whose *income* is less than a threshold value; this value is defined in absolute terms in most U.S. studies and in relative terms, generally half the median income, in most European studies. Moving from the distribution of wages to the distribution of incomes of households is complicated because some families have several wage-earning members while others have few or no labor earnings. A poor individual employed at minimum wage sees her income rise if her job is not destroyed, and this will tend to bring the poverty rate down if this individual belongs to a family with few or no labor earnings. But if the increase in minimum wage destroys jobs, some individuals see their incomes diminish, and this tends to push the poverty rate up, especially if these individuals belong to households with few labor earnings (see Brown, 1999).

The empirical literature, which relies mainly on U.S. data, finds little evidence that minimum wage hikes reduce poverty. In general, studies find that some low-wage workers living in poor families who keep their jobs do benefit from the rise in income and do move out of poverty when the minimum wage increases. However, other low-wage workers suffer from income losses because they lose their jobs or have their hours substantially reduced as a result of minimum wage increases. The study of Addison and Blackburn (1999) suggests that the rises in minimum wage that occurred

in the United States in the 1990s have contributed to reducing the poverty rate among youth aged 24 and under and among those over 24 who left school early. However, Neumark et al. (2005) find that minimum wage increases over the period 1986–1995 did not decrease the number of families in poverty and may even have increased this number slightly. In the same vein, Sabia and Burkhauser (2010) find that state and federal minimum wage increases between 2003 and 2007 had no effect on state poverty rates.

The ambiguous impact of the minimum wage on the poverty rate is not surprising, to the extent that the minimum wage creates losers and winners, and it does not allow the government to target individuals living in poor families. From this point of view, it can be more efficient to use taxes and transfers, like the EITC studied above, to reduce income inequality and poverty.

2.6 Is the Minimum Wage an Efficient Way to Redistribute Income?

The fact that the minimum wage can have beneficial effects in certain circumstances does not constitute sufficient justification for its use as a policy lever: there may be other, more efficient ways to achieve the desired goals. In particular, it is possible to act on inequality, the structure of employment, and the accumulation of human capital with a system of taxes and transfers. A system of taxes and transfers presents an obvious advantage in comparison to minimum wage: it permits the authorities to target the redistribution of incomes toward the poorest households because the relief can be adjusted with precision to aspects of their situation, for example the number of children under the roof, to which the minimum wage is blind. The question even arises whether there is any point to the minimum wage at all. Would it not be preferable in all cases to adopt a system of taxes and transfers? Many policy advisers argue that a minimum wage is always useful. Yet, as we now observe, economic analysis supplies a less affirmative response, whether in situations of perfect competition or imperfect competition.

2.6.1 A Labor Market with Perfect Competition

The efficiency of the minimum wage when there are taxes has generally been considered in labor markets with perfect competition. Most of the literature has adopted the standard Mirrlees (1971) model of optimal taxation with intensive labor supply, where individuals choose the number of hours they work and where the government observes earnings but not hourly wages or hours of work. It turns out that the minimum wage can be welfare-improving when tax schemes are constrained or when there are specific assumptions made to allow the government to observe skills at the bottom of the income distribution (Allen, 1987; Guesnerie and Roberts, 1987; Boadway and Cuff, 2001). However, as stressed by Guesnerie and Roberts (1987) and more recently by Lee and Saez (2012), informational inconsistencies arise when a minimum wage is introduced in the Mirrlees model because the implementation of minimum wage requires observing hourly wages. But if hourly wages were directly observable (and so hours worked, since the government observes earnings equal to the hourly wage times hours worked), then the government could achieve the first best allocation by conditioning taxes and transfers on the hourly wage, and the minimum wage would obviously not be useful.

This informational inconsistency causes more recent research to focus on labor supply at the extensive margin, where the agents' decision is zero–one, to work or not

to work. Lee and Saez (2012) have studied the impact of the minimum wage in this type of model, assuming perfect competition on the labor market and two skill levels. They find that the minimum wage is useful under the assumption that workers who involuntarily lose their jobs because of the minimum wage are those with the highest opportunity cost of work. In that case, overall social welfare might be improved by the minimum wage because the net loss on the part of those who lose their jobs is limited in comparison with the gains from the minimum wage for those in work. They also show that the minimum wage cannot improve upon the optimal tax/transfer allocation if workers who lose their jobs because of the minimum wage have the same opportunity cost of work as those who keep their jobs. Since there is no particular reason to think that workers who lose their jobs because of minimum wage hikes are those with the highest opportunity cost of work, a reasonable interpretation of the result of Lee and Saez is that there is no room for the minimum wage when the labor market is competitive.

2.6.2 A Labor Market with Imperfect Competition

The fact that there is no room for the minimum wage when labor markets are competitive does not mean that the minimum wage is pointless, inasmuch as labor markets do not work in a perfectly competitive fashion. When employers have some monopsony power, downward wage flexibility may increase inactivity, unemployment, and low-pay traps. In light of this, the OECD argues that "by preventing wage levels at the bottom from falling, minimum wages prevent employers from 'pocketing' the value of in-work benefits by lowering wages," and moreover "higher wage levels at the bottom mean that the same in-work income can be attained with lower in-work benefits payments" (OECD, 2005b, p. 142). Cahuc and Laroque (2013) have studied this problem in the standard optimum tax environment of models of optimal taxation with labor supply at the extensive margin, where the agent's decision is zero–one, that is, to work or not to work. They show that there is no room for the minimum wage when there is a continuum of skills with no isolated mass point at the bottom of the wage distribution. Accordingly, in the empirically relevant situation, where there is a continuum of wages at the bottom of the distribution, the minimum wage is not helpful.

All in all, economic analysis tends to indicate that in the absence of restrictions on the set of available tax levers, the minimum wage is useless in empirically relevant situations when labor markets are perfectly competitive, or monopsonistic, or fraught with search frictions. Whether the minimum wage may somehow be useful in other circumstances is an open question. In particular, the minimum wage might be desirable if its administrative implementation costs are lower than the costs of collecting taxes and subsidies. More research is needed in this area.

3 SUMMARY AND CONCLUSION

- Mandatory contributions comprise taxes and social security contributions. In continental Europe, the rate of mandatory contributions is at least 10 points higher than it is in the Anglophone countries. A large portion of this gap can be accounted for by the divergent nature—public for the former, private for the latter—of the social insurance and welfare systems.

- The gap between the cost of labor and the purchasing power of wages is measured by the *wedge*. The contribution of taxes to the wedge is referred to as the *tax wedge*.
- Theory shows that variations in marginal and average tax rates have very different impacts on labor market outcomes. Increases in the average tax rate, with the marginal tax rate held constant, increase hours of work at the intensive margin (i.e., for those who are working before the tax increase) but decrease employment and labor market participation, and drive up unemployment. Increases in marginal tax rates reduce hours of work at the intensive and at the extensive margins but exert a negative pressure on unemployment.
- Empirical investigation confirms, to a certain extent, these predictions. Research shows that tax credits and in-work benefits that reduce the marginal tax rate can improve labor market participation. Conversely, social assistance schemes, if considered independently of active labor market policies, may be detrimental to participation, notably for those with low education.
- The level of the minimum wage is clearly higher in a number of European countries (where it exceeds 40% of the average wage) than it is in the United States (where it barely reaches 30% of the average wage). In France in 2011, 11% of workers were paid at minimum wage, as opposed to 5% in the United States.
- In the monopsony model, a rise in minimum wage from a low initial value leads to an increase in employment. This may also be the case in the search and matching model. However, a rise in minimum wage from a high initial value unambiguously decreases employment in both models.
- Macroeconomic studies which attempt to establish correlations between employment and minimum wage generally conclude that the effect of this policy lever is negligible, except perhaps when it comes to youth employment. Empirical studies relying on natural experiments suggest that the impact of a hike in minimum wage on employment may be positive if the minimum wage is low with respect to the median wage, but that it becomes negative when the minimum wage is high. These results conform to the predictions of the monopsony and job search models.
- The minimum wage may exert a negative effect on inequality among those in work, but its impact on poverty is ambiguous because of its impact on job creation and destruction. In theory, a system of taxes and transfers would in most cases be a better policy lever than the minimum wage for the purposes of both reducing inequality and increasing labor market participation.

4 RELATED TOPICS IN THE BOOK

- Chapter 1, section 2: The neoclassical theory of labor supply
- Chapter 1, section 3: Empirical aspects of labor supply
- Chapter 2, section 3: Dynamic labor demand
- Chapter 3, section 1.2: The question of tax incidence

- Chapter 5, section 4.2: The equilibrium search model
- Chapter 9, section 3: The matching model
- Chapter 10, section 2.4.2: Does the minimum wage reduce inequality?
- Chapter 13, section 2.2: The effects of employment protection
- Chapter 13, section 3: The interplay between employment protection and unemployment benefits
- Chapter 14, section 2.3: Employment subsidies and the creation of public-sector jobs

5 FURTHER READINGS

Card, D., & Krueger, A. (1995). *Myth and measurement: The new economics of minimum wage*. Princeton, NJ: Princeton University Press.

Chetty, R., Manoli, D., Guren, A., & Weber, A. (2013). Does indivisible labor explain the difference between micro and macro elasticities? A meta-analysis of extensive margin elasticities. *NBER Macroeconomics Annual 2012*, *27*, National Bureau of Economic Research.

Neumark, D., & Wascher, W. (2008). *Minimum wages*. Cambridge, MA: MIT Press.

Prescott, E. (2004). Why do Americans work so much more than Europeans? *Federal Reserve Bank of Minneapolis Quarterly Review*, *28*(1), 2–13.

6 APPENDIX: SOLUTION TO THE ROGERSON AND WALLENIUS MODEL

The aim of this appendix is to derive the solution of the model of Rogerson and Wallenius (2009) presented in section 1.3.3. The maximization problem of the individuals choosing their paths for consumption $\{c(a)\}$ and hours worked $\{h(a)\}$ can be written:

$$\max_{\{c(a),h(a)\}} \int_0^1 \{\ln c(a) + \alpha \ln [1 - h(a)]\} \, \mathrm{d}a$$

subject to:

$$\int_0^1 c(a) \mathrm{d}a = \int_0^1 (1 - \tau) g(a) \max \left[h(a) - h_f, 0\right] \mathrm{d}a + T$$

where $g(a) = \left(\frac{1}{2} - \left|\frac{1}{2} - a\right|\right)$.

This problem can be rewritten in a more convenient form as:

$$\max_{\{c(a),h(a)\}} \int_0^1 \ln c(a) + \alpha \ln [1 - h(a)] \, \mathrm{d}a$$

This last constraint can be rewritten in a more convenient form as:

$$\int_0^1 c(a)\mathrm{d}a = \int_0^1 (1-\tau)g(a)\left[h(a)-h_f\right]\mathrm{d}a + T \tag{12.41}$$

$$h(a) \geq h_f \tag{12.42}$$

Let us denote by λ and μ the Lagrange multipliers associated with contraints (12.41) and (12.42) respectively. The first-order conditions are:

$$\frac{1}{c(a)} = \lambda \text{ for all } a$$

$$\frac{\alpha}{1-h(a)} = (1-\tau)g(a)\lambda + \mu \text{ for all } a$$

and the complementary slackness condition is:

$$\mu\left[h(a)-h_f\right] = 0$$

Let us consider an interior solution, such that $h(a) > h_f$ with $\mu = 0$. In that case, the first-order conditions yield:

$$h(a) = 1 - \frac{\alpha c}{(1-\tau)g(a)} \geq h_f$$

where $c = 1/\lambda = c(a)$. The budget constraint (12.41) of a household implies, together with the budget constraint of the government, which reads, $T = \tau\int_0^1 g(a)\left[h(a)-h_f\right]\mathrm{d}a$, that $c = \int_0^1 g(a)\left[h(a)-h_f\right]\mathrm{d}a$. Therefore, the solution is:

$$h(a) = \begin{cases} 1 - \frac{\alpha c}{(1-\tau)g(a)} & \text{if } a \in [\frac{\alpha c}{\left(1-h_f\right)(1-\tau)} 1 - \frac{\alpha c}{\left(1-h_f\right)(1-\tau)}] \\ 0 & \text{otherwise} \end{cases}$$

where $c = \int_0^1 g(a)\left[h(a)-h_f\right]\mathrm{d}a$.

REFERENCES

Abowd, J., Kramarz, F., Lemieux, T., & Margolis, D. (2000). Minimum wages and youth employment in France and the United States. In D. Blanchflower & R. Freeman (Eds.), *Youth employment and joblessness in advanced countries* (pp. 427–472). Chicago, IL: University of Chicago Press.

Acemoglu, D. (2001). Good jobs versus bad jobs. *Journal of Labor Economics*, *19*(1), 1–21.

Acemoglu, D. (2009). *Introduction to modern economic growth*. Princeton, NJ: Princeton University Press.

Addison, J., & Blackburn, M. (1999). Minimum wages and poverty. *Industrial and Labor Relations Review*, *52*(3), 393–409.

Aghion, P., Algan, Y., & Cahuc, P. (2011). Civil society and the state: The interplay between cooperation and minimum wage regulation. *Journal of the European Economic Association*, *9*(1), 3–42.

Ai, C., & Norton, E. (2003). Interaction terms in logit and probit models. *Economics Letters*, *80*(1), 123–129.

Alesina, A., Glaeser, E., & Sacerdote, B. (2005). Work and leisure in the US and Europe: Why so different? In *NBER Macroeconomics Annual*, *20*, National Bureau of Economic Research.

Alesina, A., & Perroti, R. (1997). The welfare state and competitiveness. *American Economic Review*, *87*, 921–939.

Allegretto, S., Dube, A., & Reich, M. (2011). Do minimum wages really reduce teen employment? Accounting for heterogeneity and selectivity in state panel data. *Industrial Relations*, *50*(2), 205–240.

Allen, S. (1987). Taxes, redistribution and the minimum wage: A theoretical analysis. *Quarterly Journal of Economics*, *101*, 477–489.

Athey, S., & Imbens, G. (2007). Discrete choice models with multiple unobserved choice characterstics. *International Economic Review*, *48*(4), 1159–1192.

Autor, D. (2011). The unsustainable rise of the disability rolls in the United States: Causes, consequences, and policy options (MIT Working Paper).

Bargain, O., & Doorley, K. (2011). Caught in the trap? Welfare's disincentive and the labor supply of single men. *Journal of Public Economics*, *95*(9–10), 1096–1110.

Bassanini, A., & Duval, R. (2006). Employment patterns in OECD countries: Reassessing the role of policies and institutions (OECD Social, Employment and Migration Working Paper No. 35).

Bazen, S., & Marimoutou, V. (2002). Looking for a needle in a haystack? A re-examination of the time series relationship between teenage employment and minimum wages in the United States. *Oxford Bulletin of Economics and Statistics*, *64*, Supplement, 699–725.

Blundell, R., Bozio, A., & Laroque, G. (2013). Extensive and intensive margins of labour supply. *Fiscal Studies*, *34*(1), 1–29.

Boadway, R., & Cuff, K. (2001). A minimum wage can be welfare improving and employment-enhancing. *European Economic Review*, *45*, 553–576.

Bourguignon, F. (2001). Redistribution and labor-supply incentives. In M. Buti, P. Sestito, & H. Wijkander (Eds.), *Taxation, welfare and the crisis of unemployment in Europe* (pp. 23–51). Cheltenham, U.K.: Edward Elgar.

Brewer, M., Duncan, A., Shephard, A., & Suárez, M. J. (2006). Did working families' tax credit work? The impact of in-work support on labour supply in Great Britain. *Labour Economics*, *13*(6), 699–772.

Brown, C. (1999). Minimum wages, employment, and the distribution of income. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 3B, chap. 32, pp. 2101–2163). Amsterdam: Elsevier Science.

Brown, C., Gilroy, C., & Kohen, A. (1982). The effect of the minimum wage on employment and unemployment. *Journal of Economic Literature*, *20*(2), 487–528.

Burdett, K., & Mortensen, D. (1998). Wage differentials, employer size and unemployment. *International Economic Review*, *39*, 257–273.

Cahuc, P., & Laroque, G. (2013). Optimal taxation and monopsonistic labor market: Does monopsony justify the minimum wage? *Journal of Public Economic Theory*, forthcoming.

Cahuc, P., & Michel, P. (1996). Minimum wage, unemployment and growth. *European Economic Review*, *40*, 1463–1482.

Cahuc, P., Saint-Martin, A., & Zylberberg, A. (2001). The consequences of the minimum wage when other wages are bargained over. *European Economic Review*, *45*, 337–352.

Card D., & Hyslop, D. (2005). Estimating the effect of time-limited earnings subsidy for welfare leavers. *Econometrica*, *73*(6), 1723–1770.

Card, D., & Hyslop, D. (2009). The dynamic effects of an earnings subsidy for long-term welfare recipients: Evidence from the self-sufficiency project applicant experiment. *Journal of Econometrics*, *153*(1), 1–20.

Card, D., & Krueger, A. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review*, *84*, 772–793.

Card, D., & Krueger, A. (1995). *Myth and measurement: The new economics of minimum wage*. Princeton, NJ: Princeton University Press.

Card, D., & Krueger, A. (2000). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania: Reply. *American Economic Review*, *90*, 1397–1420.

Card, D., & Robins, P. (2005). How important are entry effects in financial incentive programs for welfare recipients? Experimental evidence from the self-sufficiency project. *Journal of Econometrics*, *125*, 113–139.

Chay, K., & Hyslop, D. (2014). Identification and estimation of dynamic binary response panel data models: Empirical evidence using alternative approaches. In S. Carcillo, H. Immervoll, S. Jenkins, S. Königs, & K. Tatsiramos (Eds.), *Research in labor economics*, vol. 39: *Safety nets and benefit dependence*. Bingley, U.K.: Emerald Group Publishing.

Chemin, M., & Wasmer, E. (2012). Ex-ante and ex-post evaluation of the 1989 French welfare reform using a natural experiment: The 1908 social laws in Alsace-Moselle (LIEPP Working Paper, Labour Market Research Group, No. 3).

Chetty, R., Friedman, J., & Saez, E. (2013). Using differences in knowledge across neighborhoods to uncover the impacts of the EITC on earnings. *American Economic Review*, *103*(7), 2683–2721.

Chetty, R., Manoli, D., Guren, A., & Weber, A. (2013). Does indivisible labor explain the difference between micro and macro elasticities? A meta-analysis of extensive margin elasticities. *NBER Macroeconomics Annual 2012*, *27*, National Bureau of Economic Research.

Daveri, F., & Tabellini, G. (2000). Unemployment, growth and taxation in industrial countries. *Economic Policy*, April, 49–104.

Deere, D., Murphy, K., & Welch, F. (1995). Employment and the 1990–1991 minimum wage hike. *American Economic Review*, *Papers and Proceedings*, *85*(2), 232–237.

DiNardo, J., Fortin, N., & Lemieux, T. (1996). Labor market institutions and the distribution of wages, 1973–1992: A semi-parametric approach. *Econometrica*, *64*, 1001–1044.

Dolado, J., Kramarz, F., Machin, S., Manning, A., Margolis, D., & Teulings, C. (1996). The economic impact of minimum wages in Europe. *Economic Policy*, October, 319–372.

Drazen, A. (1986). Optimal minimum wage legislation. *Economic Journal*, *96*, 774–784.

Dube, A., Lester, S., & Reich, M. (2010). Minimum wage effects across state borders: Estimates using contiguous counties. *Review of Economics and Statistics*, *92*(4), 945–964.

Eissa, N., & Hoynes, H. (2004). Taxes and the labor market participation of married couples: The earned income tax credit. *Journal of Public Economics*, *88*, 1931–1958.

Eissa, N., & Liebman, J. (1996). Labour supply responses to the earned income tax credit. *Quarterly Journal of Economics*, *112*(2), 605–607.

Flinn, C. (2006). Minimum wage effects on labor market outcomes under search, bargaining and endogenous contact rates. *Econometrica*, *74*, 1013–1062.

Flinn, C. (2010). *The minimum wage and labor market outcomes*. Cambridge, MA: MIT Press.

Fortin, B., Lacroix, G., & Drolet, S. (2004). Welfare benefits and the duration of welfare spells: Evidence from a natural experiment in Canada. *Journal of Public Economics*, *88*, 1495–1520.

Gelber, A., & Mitchell, J. (2012). Taxes and time allocation: Evidence from single women and men. *Review of Economic Studies*, *79*, 863–897.

Grogger, J. (2002). The behavioral effects of welfare time limits. *American Economic Review*, *Papers and Proceedings*, *92*(2), 385–389.

Grogger, J. (2003). The effects of time limits and other policy changes on welfare use, work, and income among female-headed families. *Review of Economics and Statistics*, *85*(2), 394–408.

Guesnerie, R., & Roberts, R. (1987). Minimum wage legislation as a second-best policy. *European Economic Review*, *31*, 490–498.

Hansen, C., Pedersen, L., & Slok, T. (2000). Ambiguous effects of tax progressivity—Theory and Danish evidence. *Labour Economics*, *7*(3), 335–347.

Hansen, J., & Lofstrom, M. (2008). The dynamics of immigrant welfare and labor market behavior. *Journal of Population Economics*, *22*(4), 941–970.

Heckman, J. (1981). Heterogeneity and state dependence. In S. Rosen (Ed.), *Studies in labor markets* (pp. 91–140). Chicago, IL: University of Chicago Press.

Holmlund, B., & Kolm, A. (1995). Progressive taxation, wage setting and unemployment: Theory and Swedish evidence. *Swedish Economic Policy Review*, *2*, 423–460.

Hotz, J., Mullin, C., & Scholz, J. (2011). Examining the effect of the earned income tax credit on the labor market participation of families on welfare (Duke University Working Paper).

Hotz, J., & Scholz, J. (2006). Examining the effect of the earned income tax credit on the labor market participation of families on welfare (NBER Working Paper No. 11968).

Hyslop, D. (1999). State dependence, serial correlation and heterogeneity in intertemporal labor force participation of married women. *Econometrica*, *67*(6), 1255–1294.

Immervoll, H. (2010). Minimum income benefits in OECD countries: Policy design, effectiveness and challenges (OECD Social, Employment and Migration Working Paper No. 100).

Jones, S. (1987). Minimum wage legislation in a dual labor market. *European Economic Review*, *33*, 1229–1246.

Kennan, J. (1995). The elusive effects of minimum wage. *Journal of Economic Literature*, *33*, 1949–1965.

Kramarz, F., & Philippon, T. (2001). The impact of differential payroll tax subsidies on minimum wage employment. *Journal of Public Economics*, *82*, 115–146.

Kubik, J. (2004). The incidence of personal income taxation: Evidence from the Tax Reform Act of 1986. *Journal of Public Economics*, *88*, 1567–1588.

Layard, R., Nickell, S., & Jackman, R. (1991). *Unemployment*. Oxford, U.K.: Oxford University Press.

Lee, D. (1999). Wage inequality in the United States during the 1980s: Rising dispersion or falling minimum wage? *Quarterly Journal of Economics*, *114*, 977–1023.

Lee, D., & Saez, E. (2012). Optimal minimum wage policy in competitive labor markets. *Journal of Public Economics*, *96*, 739–749.

Lehmann, E., Lucifora, C., Moriconi, S., & van der Linden, B. (2013). Beyond the labour income tax wedge: The unemployment-reducing effect of tax progressivity (CESIFO Working Paper No. 4348).

Lemieux, T., and Milligan, K. (2008). Incentive effects of social assistance: A regression discontinuity approach. *Journal of Econometrics*, *142*(2), 807–828.

Ljungqvist, L., & Sargent, T. (2006). Do taxes explain European employment? Indivisible labor, human capital, lotteries, and savings. *NBER Macroeconomics Annual*, *21*, 181–246.

Lockwood, B., & Manning, A. (1993). Wage setting and the tax system, theory and evidence for the United Kingdom. *Journal of Public Economics*, *52*, 1–29.

Malcomson, J., & Sator, N. (1987). Tax push inflation in a unionized labour market. *European Economic Review*, *31*, 1581–1596.

Manning, A. (1995). How do we know that real wages are too high? *Quarterly Journal of Economics*, *110*, 1111–1125.

Manning, A. (2003). *Monopsony in motion: Imperfect competition in labor markets*. Princeton, NJ: Princeton University Press.

Masters, A. (1999). Wage posting in two-sided search and the minimum wage. *International Economic Review*, *40*, 809–826.

McDaniel, C. (2011). Forces shaping hours worked in the OECD, 1960–2004. *American Economic Journal: Macroeconomics*, *3*(4), 27–52.

Meghir, C., & Phillips, D. (2010). Labour supply and taxes. In J. Mirrlees, S. Adam, T. Besley, R. Blundell, S. Bond, R. Chote, M. Gammie, P. Johnson, G. Myles, & J. Poterba (Eds.), *Dimensions of tax design: The Mirrlees review*. Oxford, U.K.: Oxford University Press.

Meyer, B., & Rosenbaum, D. (2001). Welfare, the earned income tax credit, and the labor supply of single mothers. *Quarterly Journal of Economics*, *116*(3), 1063–1114.

Michalopoulos, C., Tattrie, D., Miller, C., Robins, P., Morris, P., Gyarmati, D., Redcross, C., Foley, K., & Ford, R. (2002). *Making work pay: Final report of the self sufficiency project for long term welfare recipients*. Social Research and Demonstration Corporation, Ottawa.

Mirrlees, J. (1971). An exploration in the theory of optimum income taxation. *Review of Economic Studies*, *38*, 175–208.

Mirrlees, J., Adam, S., Besley, T., Blundell, R., Bond, S., Chote, R., Gammie, M., Johnson, P., Myles, G., & Poterba, J. (2010). *Dimensions of tax design: The Mirrlees review*. Oxford, U.K.: Oxford University Press.

Musgrave, R., & Musgrave, P. (1989). *Public finance in theory and practice* (5th ed.). New York, NY: McGraw-Hill.

Neumark, D., Salas, I., & Wascher, W. (2013). Revisiting the minimum wage-employment debate: Throwing out the baby with the bathwater? (NBER Working Paper No. 18681). *Industrial and Labor Relations Review*, forthcoming.

Neumark, D., Schweitzer, D., & Wascher, W. (2005). The effects of minimum wages on the distribution of family incomes: A non-parametric analysis. *Journal of Human Resources*, *40*(4), 867–917.

Neumark, D., & Wascher, W. (2000). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania: Comment. *American Economic Review*, *90*, 1362–1396.

Neumark, D., & Wascher, W. (2008). *Minimum wages*. Cambridge, MA: MIT Press.

Nickell, S., Nunziata, L., & Ochel, W. (2005). Unemployment in the OECD since the 1960s. What do we know? *Economic Journal*, *115*, 1–27.

OECD. (1994). *The OECD jobs study*. Paris: OECD Publishing.

OECD. (1998). *Employment outlook*. Paris: OECD Publishing.

OECD. (2001). *Taxing wages: Income tax, social security contributions and cash family benefits, 1999–2000*. Paris: OECD Publishing.

OECD. (2005a). *Employment outlook*. Paris: OECD Publishing.

OECD. (2005b). *Taxing wedges*. Paris: OECD Publishing.

OECD. (2009). *Benefits and wages*. Paris: OECD Publishing.

OECD. (2011). *Divided we stand: Why inequality keeps rising*. Paris: OECD Publishing.

OECD. (2013). *Society at a glance*. Paris: OECD Publishing.

Piketty, T., & Saez, E. (2013). Optimal labor income taxation. In A. Auerbach, R. Chetty, M. Feldstein, & E. Saez (Eds.), *Handbook of public economics* (vol. 5, pp. 391–474). Amsterdam: North-Holland.

Portugal, P., & Cardoso, A.-R. (2006). Disentangling the minimum wage puzzle: An analysis of worker accessions and separation. *Journal of the European Economic Association*, *4*(5), 988–1013.

Prescott, E. (2004). Why do Americans work so much more than Europeans? *Federal Reserve Bank of Minneapolis Quarterly Review*, *28*(1), 2–13.

Rebitzer, J., & Taylor, L. (1995). The consequences of minimum wage laws: Some new theoretical Ideas. *Journal of Public Economics*, *56*, 245–255.

Robinson, J. (1933). *The economics of imperfect competition*. London: Macmillan.

Rogerson, R. (2007). Taxation and market work: Is Scandinavia an outlier? *Economic Theory*, *32*(1), 59–85.

Rogerson, R., & Wallenius, J. (2009). Micro and macro elasticities in a life cycle model with taxes. *Journal of Economic Theory*, *144*, 2277–2292.

Rothstein, J. (2010). Is the EITC as good as an NIT? Conditional cash transfers and tax incidence. *American Economic Journal: Economic Policy*, *2*(1), 177–208.

Sabia, J., & Burkhauser, R. (2010). Minimum wages and poverty: Will a $9.50 federal minimum wage really help the working poor? *Southern Economic Journal*, 76(3), 592–623.

Salverda, W. (2008). The Dutch minimum wage: Radical reduction shifts main focus to parttime jobs. In D. Vaughan-Whitehead (Ed.), *The minimum wage revisited in the enlarged EU: Issues and challenges* (pp. 291–330). Geneva: International Labour Organisation.

Shapiro, C., & Stiglitz, J. (1984). Equilibrium unemployment as a worker discipline device. *American Economic Review*, *74*(3), 433–444.

Stewart, M. (2007). The interrelated dynamics of unemployment and low-wage employment. *Journal of Applied Econometrics*, *22*(3), 511–531.

Stigler, G. (1946). The economics of minimum wage legislation. *American Economic Review*, *36*, 535–543.

van den Berg, G., & Ridder, G. (1998). An empirical equilibrium search model of the labor market. *Econometrica*, *66*(5), 1183–1223.

Wooldridge, J. (2005). Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics*, *20*(1), 39–54.

Wooldridge, J. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). Cambridge, MA: MIT Press.

CHAPTER 13

Insurance Policies

In this chapter we will:

- Gain an overview of the unemployment insurance systems in the OECD
- Learn the characteristics of optimal unemployment benefits
- Study the kinds of policy measures intended to protect jobs
- Observe the effects of these employment protection measures on wages, unemployment, and productivity
- Understand why employment protection and insurance benefits should be designed together

INTRODUCTION

Public unemployment insurance systems were created in many European countries at the beginning of the twentieth century. In this area, the purpose of state intervention is to insure workers against the risk of unemployment, a burden the state was forced to assume because imperfect information hinders the creation of private insurance systems providing compensation for job loss. Chiu and Karni (1998) have in fact shown that the imperfection of the information available about the preferences of agents for leisure, and about the effort they may be making to hunt for a job, can tend to an equilibrium in which no unemployment insurance is provided by market forces, even though the agents are risk averse. It is therefore necessary for collectivities such as trade unions, or the state itself, to step in and operate a system of unemployment insurance. Such systems are found today in all industrialized countries. For that matter, the state also intervenes to provide social assistance, redistributing income in favor of the most disadvantaged workers—those who are generally faced with more frequent and lengthier spells of unemployment than other workers.

Compensation for job loss has had to weather a long-standing and well-rehearsed critique. Essentially, benefit payments are said to reduce the incentive to look for a job, increase the reservation wage (see chapter 5 on job search), and exert upward pressure

on wages (see chapter 7 on bargaining). These mutually reinforcing effects are said to increase the duration of unemployment. Overall, then, we are led to expect that generous unemployment benefits do have a positive impact on the unemployment rate and do lead to a reduction in aggregate output. Yet this expectation needs to be put in context and clarified. In the first place, unemployment benefits give the unemployed the means to reduce their income loss and to better select from among the jobs that are offered to them. From this standpoint, benefits constitute a "subsidy" to the job search, and an increase in the amount paid out in benefits can improve the average quality of jobs and increase overall production. Any assessment of the "right" level of unemployment benefit ought to take into account the advantages just mentioned, along with the well-known disadvantages. This is the core problem addressed by all the research on optimality in unemployment insurance systems at which we will be looking in this chapter.

There exists another and more indirect way of insuring workers against the risk of job loss: governments may choose to protect existing jobs, for example by using binding administrative regulations to make layoffs more costly for firms. The literature adopts the term "employment protection" to denote all policy measures that hinder firms from altering at will the terms of their labor contracts with their employees. The stricter the employment protection measures in place, the more "rigid" labor markets are characterized as being. Comparison of the employment performance of the OECD countries and the various approaches they take to regulating their labor markets has attracted a great deal of attention. It is widely believed that "rigidity" in these markets is responsible for unemployment. But is that really the case? The search and matching model set out in chapter 9 proves particularly useful when it comes to addressing this question. It represents the dynamic functioning of an imperfectly competitive labor market and describes behaviors with enough precision to allow us to study the impact of employment protection measures.

Section 1 offers an overview of unemployment insurance systems and studies the question of optimality in unemployment insurance in theory and in practice. Section 2 begins with an overview of the range of employment protection measures and continues with an analysis of their impact, using a matching model where job destruction is endogenous. It concludes by laying out the main empirical results on the topic. Section 3 examines the complex interplay between unemployment protection and unemployment insurance. It shows that "optimal" employment protection should cause firms to internalize the social costs incurred when they destroy jobs—social costs that depend in turn on the generosity of the unemployment benefit.

1 UNEMPLOYMENT INSURANCE

In the area of unemployment insurance, the OECD countries have adopted widely varying regulations, especially regarding eligibility, the amount of benefit paid, and its duration. Such wide variation leads to the question of optimality in unemployment insurance: how long should a job seeker continue to receive it? Basically, the question of optimal unemployment insurance comes down to determining the amount and the time profile of unemployment benefit that will maximize the welfare of job seekers under the

budgetary constraints of the agency in charge of managing the unemployment insurance system. Most often the agency in charge cannot check thoroughly on whether its unemployed clients are making appropriate efforts to find a job. The agency is faced with a "moral hazard" problem in that perfect insurance, in other words complete replacement of the unemployed person's lost income, might also rob him of any incentive to look for a job energetically.

We study this question with the help of a simple static model, highlighting the parameters that have to be known in order to calculate the optimal level of unemployment benefit. We then examine this question utilizing the standard job search model set forth in chapter 5 and assuming that the amount of unemployment benefit remains constant over time. Finally, we relax this hypothesis in order to study the optimal time profile of unemployment benefit.

As a preliminary to these theoretical developments, we take a broad look at the unemployment insurance systems existing in the principal OECD countries.

1.1 An Overview of Unemployment Insurance Systems

This section gives an overview of unemployment benefit in several OECD countries. The main parameter of unemployment insurance is the replacement ratio, in other words the ratio between the amount of the benefit payment and the last wage earned. But a large proportion of those who lose their jobs do not receive benefits from the unemployment insurance system because they have not paid contributions for long enough. So we must always bear in mind the distinction between assistance payments, which are conditional upon the income of agents, and insurance payments, which depend on the contributions agents (and their employers) have paid into the unemployment insurance system while they were in work. We also discuss short-time work, which resembles a system of insurance against unemployment in that it allows employers to reduce the hours worked by employees rather than lay them off.

1.1.1 Insurance vs. Assistance

The income of a job seeker most often combines payments from an insurance system and ones from a social assistance fund. Unemployment insurance systems generally pay benefits for a limited period, from several months to several years, to persons who have already been employed and paid into the fund (Grubb, 2001; Venn, 2012). The amount of benefit is often linked to the wage earned in the most recent job. Payments made by the social assistance fund, on the other hand, are means-tested and are classified as unemployment assistance. Like unemployment insurance benefits, unemployment assistance benefits are conditional upon job search and availability for work (which sets them apart from most kinds of social assistance). But unlike insurance benefits, they are generally of unlimited duration and independent of past earnings. To social assistance payments made specifically to job seekers we must add the various allowances (family allowance, housing allowance, single-parent allowance, etc.) that may be paid to any member of the labor force when she meets certain means criteria. These allowances may top up unemployment benefits, depending on household composition and income level.

Tables 13.1 and 13.2 present the main characteristics of unemployment insurance and assistance benefits in 15 countries in 2010. The variety of rules makes systems difficult to compare at first glance.

In the OECD, the average maximum duration of unemployment insurance benefits is 15 months, excluding Belgium, which is the only country with unlimited duration. In the United States the relatively high duration shown in table 13.1 is due to the temporary extension of benefits introduced in 2009 after the beginning of the Great Recession and which ended at the end of 2013. The normal duration in the United States is 6 months. In most OECD countries there is a ceiling for benefits, which is set at about 70% of the average wage.

In most countries payments are determined as a percentage of the earnings base, but there are exceptions, such as the United Kingdom. The third and fourth columns of table 13.1 show that this percentage varies from about 50% to 90%. However, these

TABLE 13.1
The rules of unemployment insurance schemes in selected OECD countries, in 2010.

Country	Maximum duration (months)	Payment rate (% of earnings base)		Minimum benefit as a % of AW	Maximum benefit as a % of AW
		Initial rate	At end of legal entitlement period		
Belgium	Unlimited	60	54 (after 1 year)	23.4	36.6
Canada	11	55	55	—	53.1
Denmark	24	90	90	43	52
France	24	57–75	57–75	28.1	227.5
Germany	12	60	60	—	91.7
Italy	8	60	50 (after 6 months)	—	45.6
Japan	9	50–80	50–80	—	52.7
Korea	7	50	50	20.8	39
Netherlands	22	75	70 (after 2 months)	30.4	79.9
Norway	24	62	62	15	60
Poland	12	Fixed amount (23.2% of AW after 3 months)		—	—
Spain	24	70	60 (after 6 months)	24.1	52.6
Sweden	35	80	70 (after 9 months)	22.6	48
United Kingdom	6	Fixed amount (9.9% of AW)		—	—
United States	23	53	53	13.3	41.2

Note: AW = gross average wage. The minimum/maximum benefits are for a single, 40-year-old worker without children; benefits may differ depending on family situation. All benefit amounts are shown on an annualized basis. The minimum/maximum gross benefits are expressed as a percentage of the gross average wage in the economy. In Australia there is no unemployment insurance scheme (as in New Zealand). For Canada, the duration of Employment Insurance (EI) benefits depends on the unemployment rate in the relevant EI region. The 11 months' duration shown here relates to an unemployment rate of 9% in Ontario. For the United States, the information reflects the situation of the Michigan unemployment benefit scheme. The payment duration has been extended in the United States due to high unemployment rates, up to 23 months. Emergency Unemployment Compensation and Extended Benefits are paid after exhaustion of regular unemployment insurance, which is 26 weeks.

Source: OECD Taxes and Benefits calculator (www.oecd.org/els/social/workincentives).

TABLE 13.2
The rules of unemployment assistance schemes in selected OECD countries, in 2010.

Country	Duration (months)	Payment rate	Maximum benefit as a % of AW	Test on Assets	Test on Income
Australia	Unlimited	Fixed amount	18	Yes	Family
France	6 months (renewable)	Fixed amount	15.6		Family
Germany	Unlimited	Fixed amount	10.2	Yes	Family
Spain	30	Fixed amount	20.6		Family
Sweden	14	Fixed amount	22.6		Individual
United Kingdom	Unlimited	Fixed amount	9.9	Yes	Family

Note: AW = average wage. The maximum benefit is for a single, 40-year-old worker without children; benefits may differ depending on family situation. All benefit amounts are shown on an annualized basis. In Spain, benefits are only paid to people with dependents unless aged over 45. The maximum gross benefits are expressed as a percentage of the gross average wage in the economy.

Source: OECD Taxes and Benefits calculator (www.oecd.org/els/social/workincentives).

differences across countries do not necessarily reflect differences in net replacement income: most countries calculate benefits on the basis of gross earnings; some do so on the basis of net earnings (e.g., Germany); yet others use pretax but post-social-security-contributions earnings as a base (e.g., Denmark). Also, benefits may be taxed at different rates. Ceilings may reduce replacement rates at higher wage levels; for instance, the maximum gross benefit represents more than 200% of the average gross wage in France, but only 37% in Belgium. In some European countries, payment rates decrease over time (Belgium, Italy, Netherlands, Spain, and Sweden). This characteristic will be discussed below in models of optimal unemployment insurance.

Unemployment assistance schemes do not exist in all the countries included in table 13.1. Where they do exist, payments are usually at a fixed rate and the amounts are much less than unemployment insurance (for comparability purposes, the maximum benefit a single person can get on an annual basis is expressed as a percentage of the national average gross wage in the fourth column of the table). Unemployment assistance is also conditional upon means testing, based on either family income and/or assets (e.g., ownership of a dwelling; see the last columns of table 13.2). In some countries people must exhaust their insurance entitlement before becoming eligible for assistance (e.g., France) and in many countries the only requirement is to be jobless and actively looking for work. Where unemployment assistance does not exist, the unemployed with no, or no more, entitlement to insurance have only general minimum-income schemes to turn to, if these exist.

1.1.2 How to Measure the Generosity of an Unemployment Insurance System

The OECD has constructed a synthetic indicator of the generosity of unemployment benefit: it is based on the replacement ratio. But to determine the effective level of compensation, the researcher must not neglect social assistance benefits either. Last, in order to have complete knowledge of any system of unemployment benefit, it is necessary to take into account the eligibility conditions and the sanctions imposed for half-hearted job search.

The OECD Synthetic Indicator of the Replacement Ratio

The OECD synthetic indicator of the generosity of unemployment benefit is an average of the entitlements of single, unemployed persons and of those living in couples, with or without children, whose spell of joblessness has lasted from zero to five years. This indicator is either a gross replacement ratio, equal to the ratio of gross benefit payments to gross wages, or a net replacement ratio, which takes into account payroll deductions, taxes, and transfers for both benefits and past wages. Figure 13.1 gives an overview of the path of gross replacement ratios over time in several OECD countries. On average, this ratio was about 23% in the OECD countries in 2011. We see that the replacement ratio exhibits an increasing trend on average, notably in Europe over the period 1961–2011. Still, this average trend masks strong disparities. In general, the countries with low replacement ratios are also the ones where this ratio remains most stable over the last five decades of the twentieth century: Canada, Japan, and the United States (excepting the period since 2009, due to the temporary extension of benefits under crisis conditions). Conversely, Denmark, France, and the Northern European countries at the beginning of the 1970s, as well as Austria, France, and Spain at the beginning of the 1980s, increased their replacement ratios, though these leveled off in the 1990s. Germany remained stable at a high level until the beginning of the 2000s, but following the Hartz reforms the generosity there has decreased significantly, while the United Kingdom saw a significant decline in the synthetic indicator of entitlements for the jobless over the whole period.

Net replacement ratios are significantly higher than gross ones. They provide a more comparable assessment of the generosity of systems across countries because in many countries taxes on benefits are distinct from taxes on wages, due to the progressivity of taxes and income redistribution policies. The average net replacement ratio is around 50% higher than the average gross ratio for the OECD countries as a whole (about 34% in 2011 over a five-year spell of unemployment). Figure 13.2 shows the net replacement rate for singles, over a five-year spell (averages for three levels of past wages: 67%, 100%, and 150% of the average wage), for 16 countries. As for the gross ratio, general social assistance payments are excluded from the calculation, and only unemployment assistance schemes are taken into account. France has the highest ratio during the first year at 68%, and the second highest over the five years, behind Belgium (54%). Replacement ratios over five years are also high in the Nordic countries and Germany, but still about half of those of France or Belgium. In the United States and Canada, net replacement rates are comparable to those in most European countries in the first year of unemployment but decline quickly after the first year, resulting in low period averages. In the United Kingdom, replacement rates are flat over five years and appear to be low because housing, family benefits, and other social assistance supplements are excluded. No data are available to compare net replacement ratios over the long run. Nonetheless, given the strong correlation between net ratios and gross ratios, it is likely that the average net ratio has risen since the beginning of the 1960s in the OECD countries.

The synthetic indicator masks in part the linkage between the duration of unemployment and the amount of the benefit payment. In many countries, unemployment benefits taper off as the jobless spell lengthens. Taking the case of singles with no children, figure 13.2 shows that benefits fall off very steeply in the United States, and that the replacement ratio is relatively generous in Japan for the first year but then falls off sharply

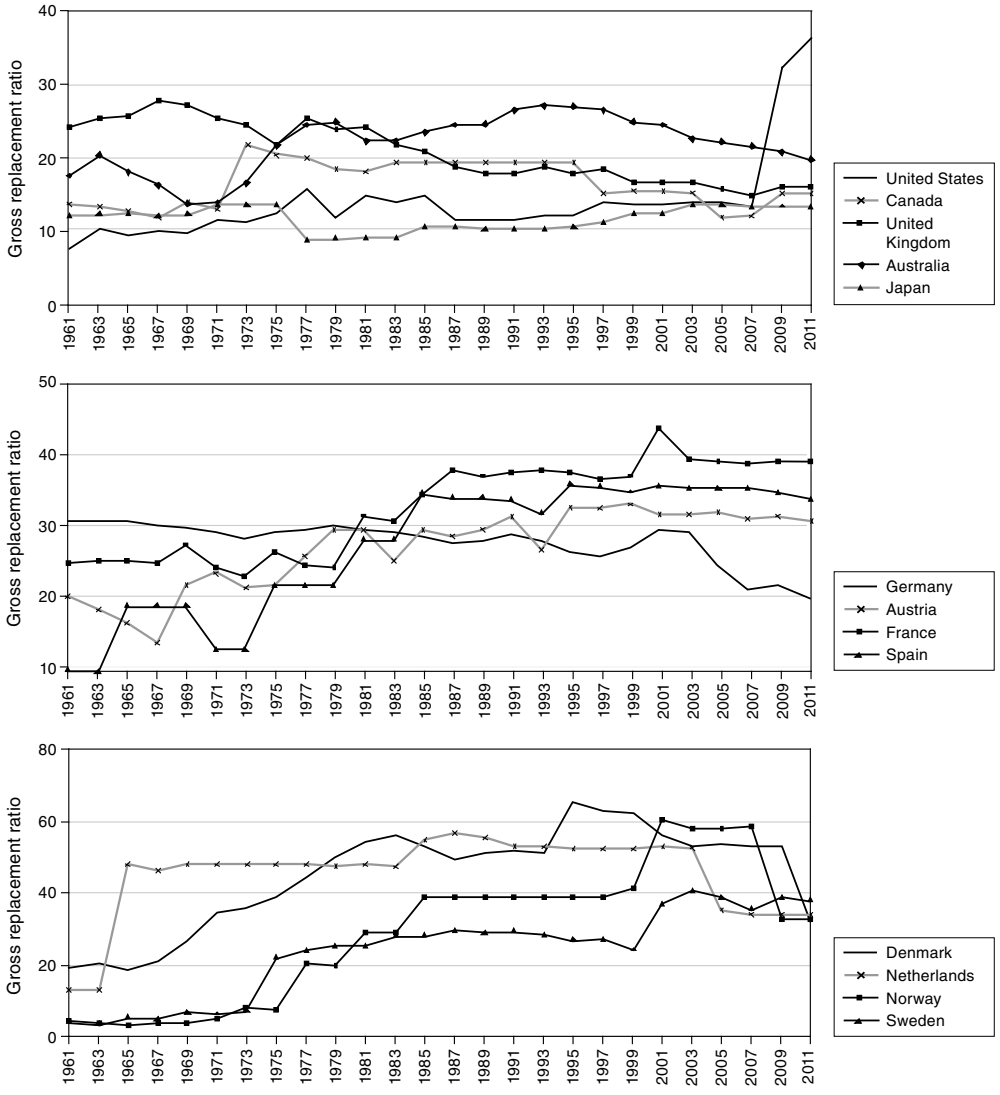


FIGURE 13.1

The synthetic indicator of entitlement to unemployment benefits (gross replacement ratio in percentage).

Note: The OECD summary measure is defined as the average of the gross unemployment benefit replacement rates over five years, for two earnings levels (67% and 100% of the average production wage), three family situations (single, one-earner couple, two-earner couple), and three durations of unemployment. It includes unemployment insurance and unemployment assistance benefits. For the years 2007–2011 the gross replacement ratio is based on the average production wage.

Source: OECD Tax-Benefit Models (www.oecd.org/els/social/workincentives). Gross replacement rates, uneven years from 1961 to 2011.

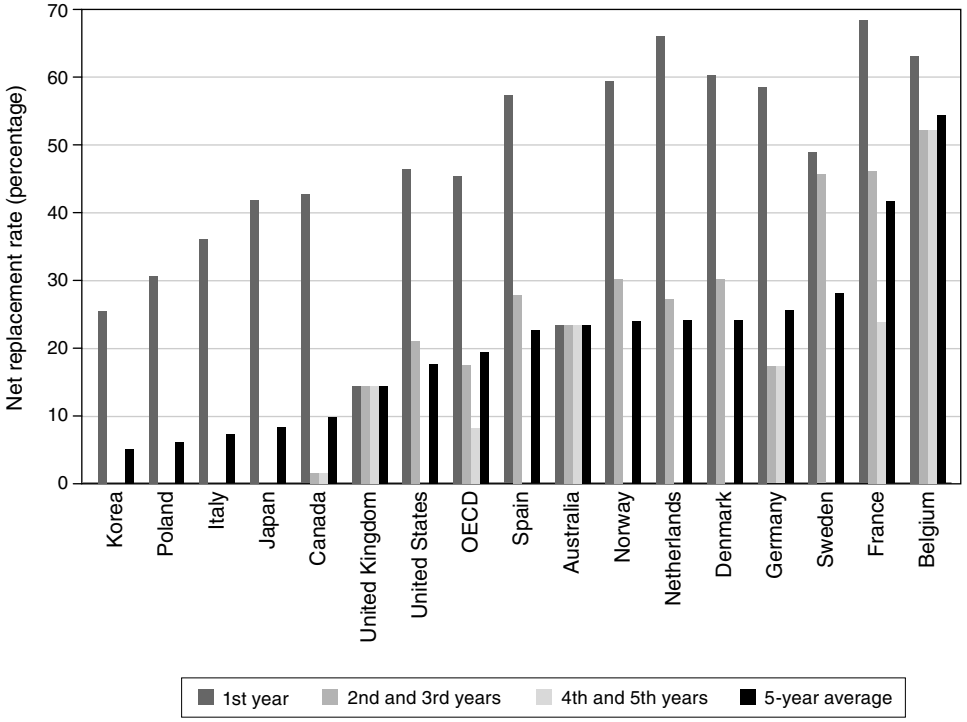


FIGURE 13.2
Net replacement rate for unemployment benefits only for single person in 2011 (in percentage of past earnings).

Note: Includes unemployment insurance and unemployment assistance benefits but not general social assistance or housing schemes. Rates are averages of replacement ratios at three levels of income (67%, 100%, and 150% of the average wage). Calculations consider cash incomes only, as well as income taxes and mandatory social security contributions paid by employees. The rate is equal to zero for some countries and some years, since we ignore assistance or housing benefits that could replace income at the end of unemployment benefit entitlements. OECD refers to the nonweighted average of rates for the OECD countries. Countries are ranked according to their 5-year average rate.

Source: OECD Tax-Benefit Models (www.oecd.org/els/social/workincentives).

at the beginning of the second year of unemployment. This tapering off in replacement ratios generally reflects a shift from unemployment insurance to social assistance.

The Effective Level of Compensation

Unemployment insurance and assistance benefits are often complemented by family benefit, as well as housing benefit and even some social assistance supplements in cases where unemployment benefits alone are not enough to reach the minimum income threshold set by social assistance programs. Assuming that households are eligible for these benefits, and adding them to unemployment benefits, can substantially change the level of compensation of the unemployed, as shown by the simulations in figure 13.3. These simulations reflect the replacement rate for four family types because the amount of the social assistance supplements often varies with the family composition. Whereas the five-year average net replacement rate is only 27% without housing and social assistance allowances (but still including family benefits for families with children), the rate goes up to 50% when other types of benefits are included (again for eligible recipients)!

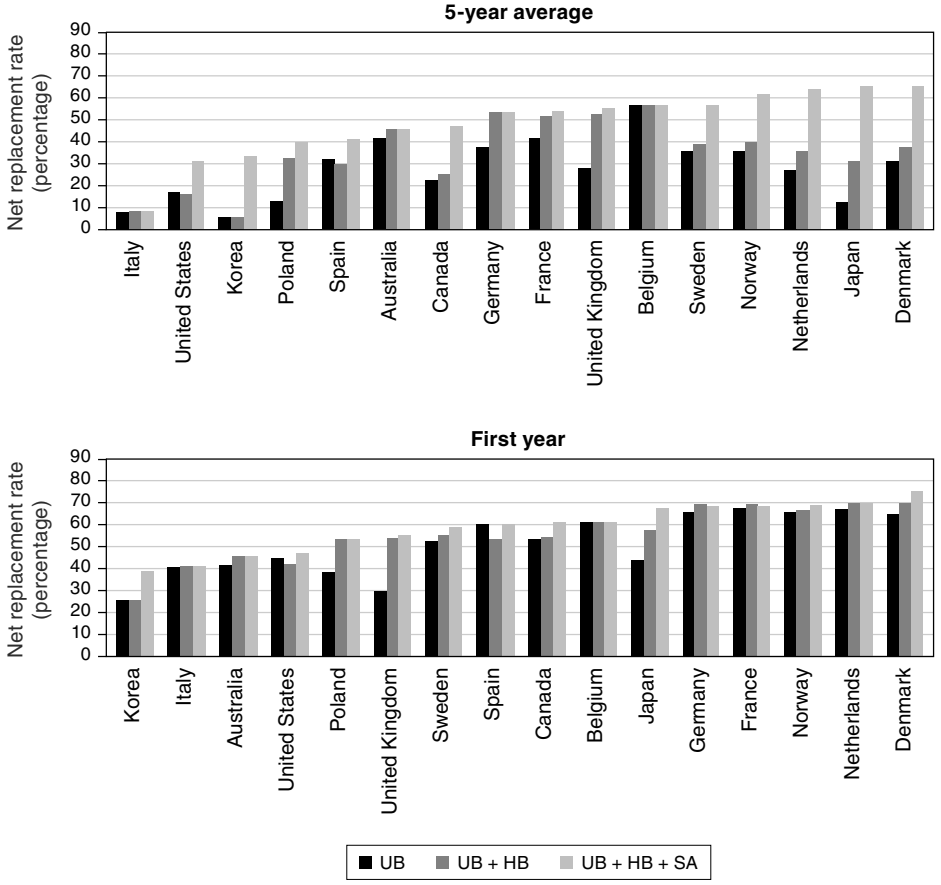


Figure 13.3

Net replacement rates for unemployment benefits, housing, and social assistance allowances for four family types in 2011 (in percentage of past earnings).

Note: UB = unemployment benefits and family benefits (when applicable); UB + HB = unemployment benefits, family benefits (when applicable), and housing benefits; UB + HB + SA = unemployment benefits, family benefits (when applicable), housing benefits, and social assistance. The rates are an average for four types of families (single, couple with one earner, lone parent with two children, couple with one earner with two children), with in-work earnings equal to 67%, 100%, and 150% of the average wage. Any income taxes payable on unemployment benefits are determined in relation to annualized benefit values (i.e., monthly values multiplied by 12) even if the maximum benefit duration is shorter than 12 months. Children are aged 4 and 6 and neither child care benefits nor child care costs are considered. Calculations consider cash incomes only, as well as income taxes and mandatory social security contributions paid by employees.

Source: OECD Tax-Benefit Models (www.oecd.org/els/social/workincentives).

The difference in replacement rates, with or without housing and social assistance, is stronger over a five-year period than in the first year because in most countries unemployment benefits do not last longer than one or two years. Figure 13.3 shows that these additional benefits play an important role for some households in Japan, Korea, Poland, the United States, and the United Kingdom, even during the first year of unemployment.

The Eligibility Conditions and the Rules for Sanctions

The synthetic indicator also masks factors having to do with the conditions under which unemployment benefit is paid. These eligibility conditions concern not only the duration of the contribution period (for insurance benefits) but also the reasons for the job loss. Many systems provide for sanctions when a person quits voluntarily or is fired for cause. Figure 13.4 gives an overview of the extent of such sanctions in some OECD countries. Australia and New Zealand require no contribution record and impose relatively light sanctions for voluntary unemployment (unemployment benefits are noncontributory in these countries). Less than a year of employment is required in Canada, France, Japan, Korea, the Netherlands, the United States, and the United Kingdom. Nordic countries have relatively relaxed entitlement conditions once sanctions for voluntary unemployment are taken into account. Workers voluntarily unemployed are not eligible for unemployment benefit in many countries, including Canada, Italy, Korea, the Netherlands, Spain, and the United States. In other countries benefits are either reduced or delayed for a certain period.

The eligibility conditions for unemployment benefit also include aspects of job search, with many systems specifying that beneficiaries must furnish proof that they are actively looking for work, must not actually be working, and must accept jobs offered to them that are judged to meet the criteria defined by the unemployment insurance system (see Venn, 2012, for more detail). Benefit recipients are subject to more or less strict

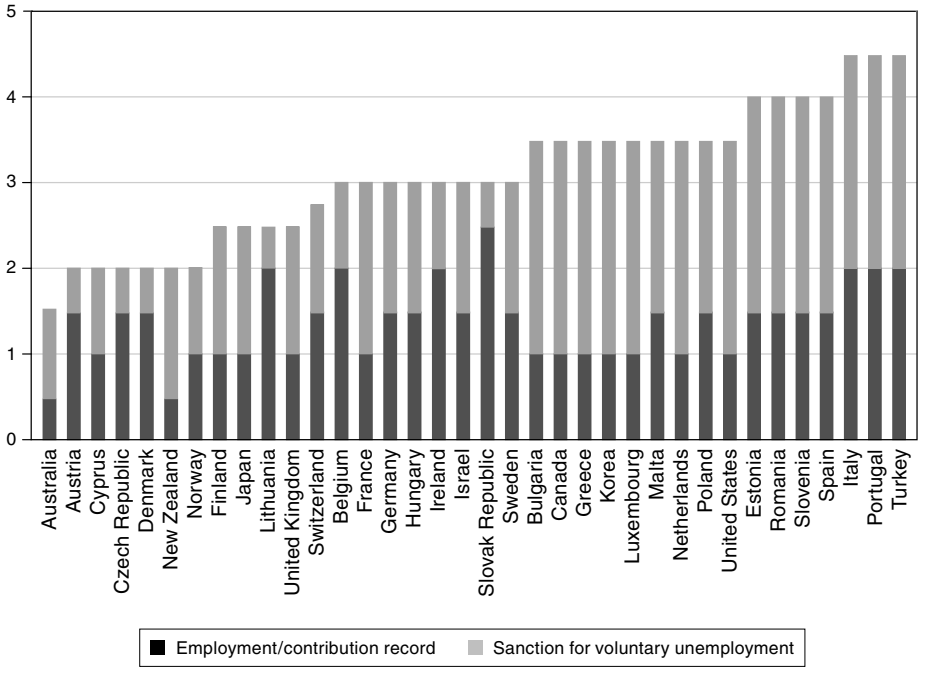


FIGURE 13.4

Strictness of entitlement to benefits in 2011. Indicator scored from 1 (least strict) to 5 (most strict).

Source: Venn (2012, figure 2, p. 15).

monitoring of their job search activities over the spell of compensated unemployment, and also to sanctions when they refuse offers or make half-hearted efforts to hunt for a job. As noted in chapter 5, section 2.2.5, the strictness of such sanctions varies widely in the OECD countries. In principle, the United States suspends benefits completely for an initial refusal of a job offer or refusal to participate in an active labor market program (ALMP), while suspensions are relatively short (one month or less) in Denmark, Germany, Japan, and Korea. In some countries, benefits are reduced for a fixed period (France, the Netherlands, Sweden) rather than suspended. Many countries have higher sanctions for benefit recipients who repeatedly refuse suitable job offers or participation in an ALMP without good reason: examples include Belgium, Denmark, Finland, and Sweden. The labor economist must bear in mind that it is difficult to assess the effectiveness of such sanctions. Enforcement is key in this matter. Observed low sanction rates in one country may mean either that sanctions are not seriously being applied or conversely that the threat exerted by sanctions is highly credible there. The impact of sanctions can only be assessed in experimental or quasi-experimental settings (see chapter 14). Overall, eligibility criteria and the rules governing sanctions are strictest in Portugal, Spain, and Italy and least strict in Canada, Denmark, Japan, and Sweden (Venn, 2012).

1.1.3 A High Proportion of Uninsured, Unemployed Persons

The OECD synthetic indicator is often used in international comparisons of unemployment benefit, but it is important to stress that it conceals wide heterogeneity. In particular, a large number of persons who are looking for work do not receive unemployment insurance benefit because they do not satisfy the eligibility conditions. As we have seen, though, they may receive transfers from the social assistance system, either in the form of unemployment assistance or through general social assistance schemes. Figure 13.5 gives an idea of the extent of this phenomenon, by representing the ratio of unemployment insurance recipients to the overall number of jobless persons (i.e., those not working, actively looking for a job, and available for work) over the period 2007–2010 for 33 OECD countries. This ratio is called the a "pseudo-coverage" rate.

Essentially, persons who do not benefit from unemployment insurance are new entrants into the labor market or have not paid into an unemployment insurance fund for a long time or have exhausted their entitlement to benefits after a long spell of joblessness. Scrutiny of figure 13.5 reveals that very few of those looking for work receive unemployment insurance benefit in the countries of Southern Europe. In Japan only 23% of the unemployed are compensated by insurance schemes, and in Germany only 33%. Overall, about 55% of the jobless in the OECD countries do not receive unemployment insurance benefit. Some countries feature high pseudo-coverage rates. For instance, in France about 85% of the jobless receive an insurance benefit, and the ratio is over 100% in Iceland because many unemployment insurance recipients do not consider themselves unemployed.[1] The pseudo-coverage rate is also above 50% in the United States, the United Kingdom, and the Northern European countries.

[1] The number of unemployment benefit recipients is not a subsample of the overall number of unemployed persons according to the ILO-OECD definition. Indeed, some benefit recipients may not declare themselves as unemployed in the Labor Force Surveys (i.e., not working, being available for work, and looking actively for a job). For this reason, the ratio of unemployment insurance recipients to the number of unemployed is called a pseudocoverage rate and may be over 100% in some cases.

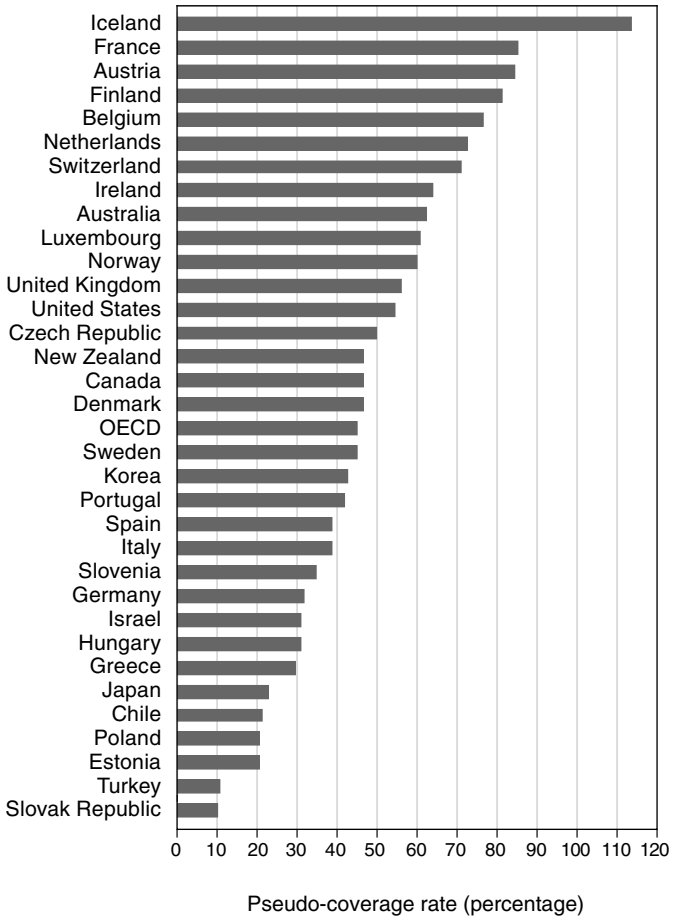


Figure 13.5
Average unemployment insurance benefits pseudo-coverage rate in 2007–2010 (percentage of the unemployed).

Note: The pseudo-coverage rate is the ratio of the number of unemployment insurance benefit recipients to the overall number of unemployed workers according to the ILO-OECD definition. For Australia and New Zealand, unemployment assistance schemes are considered as there are no insurance schemes in these countries. OECD refers to the nonweighted average of rates for the OECD countries.

Source: OECD Social Benefit Recipients database and Labor Force Survey database.

1.1.4 Short-Time Work

Short-time work (or short-time compensation) is an insurance scheme that aims at reducing layoffs by allowing employers to temporarily reduce hours worked while compensating workers for the forced drop in their income. The difference between this and unemployment insurance is that workers are not laid off. Either working hours are reduced or, in extremes cases, the labor contract is temporarily suspended. Compensation is usually delivered through the unemployment insurance scheme in the form of partial unemployment benefits, from special funds, directly by the state, or sometimes by a combination of these sources. Before the 2008–2009 crisis, short-time work

schemes were already in place in 18 OECD countries. They were implemented in even more countries during the crisis, and now they are in operation in 25 of the 34 OECD countries, including most of the continental European ones. Among the Nordic countries, Denmark, Finland, and Norway have short-time work schemes, as do Canada, Ireland, New Zealand, and the United States among the Anglophone countries. In good economic times, the number of workers in these schemes tends to be very small. But in hard times, participation in these schemes can balloon very quickly. For instance, at the end of 2009, 1.3% of employees in the OECD countries were taking part in such schemes.

However, figure 13.6 shows that there are large cross-country differences in take-up rates, which range from zero in some countries to 3% of employees in Germany and Italy, and even 5.5% in Belgium (on average in 2009). Countries where employment protection of regular contracts is stricter (see section 2 below), such as Belgium, Germany, Italy, and France, tend to resort more to these schemes as an alternative to layoffs in bad times. In the 2008–2009 recession, unemployment did not increase in some European countries featuring widespread and generous short-time compensation

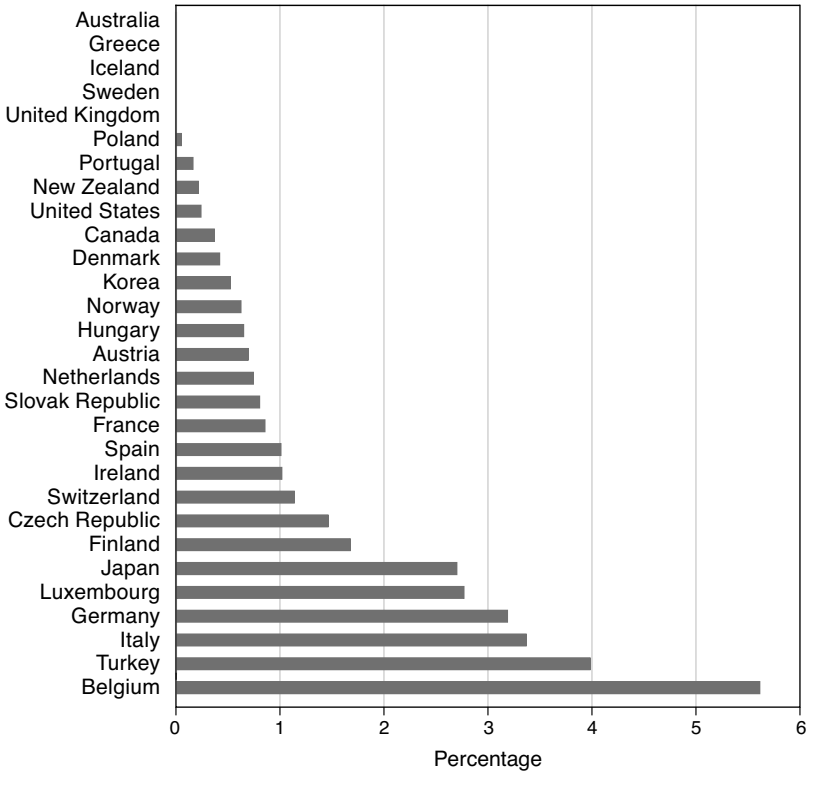


FIGURE 13.6
Participation rate of employees in short-time work schemes in 2009 (percentage of employees).

Note: In 2009 Australia, the United Kingdom, Greece, Iceland, and Sweden had no such schemes.

Source: Cahuc and Carcillo (2011, figure 1) based on OECD data.

programs as much as it did in other countries. The leading example is Germany, which makes particularly intensive use of a short-time work program (the *Kurzarbeit*). This success induced a renewal of interest in short-time work as a way to damp unemployment during recessions (Boeri and Bruecker, 2011; Cahuc and Carcillo, 2011; Hijzen and Venn, 2011; Brenke, Rinne, and Zimmermann, 2013), although the long-term consequences of these schemes for productivity are yet to be studied.

1.2 The Basic Analysis of Optimal Unemployment Insurance

In chapter 5 we looked at the impact of compensation in case of job loss on the duration of job search from both the theoretical and empirical standpoints. In that setting, compensation rules were taken as given. We now address the problem from a normative standpoint, seeking to illuminate the problem of determining these rules in an optimal manner. The canonical analysis of optimal unemployment insurance is that of Baily (1978). In his framework, there is an agency charged with setting the amount of unemployment benefit paid to job seekers and the amount of tax paid by wage earners. Knowing these two parameters, agents choose the intensity of their job search.

1.2.1 The Behavior of Agents

First we describe the behavior of suppliers of labor, then the behavior of the agency charged with managing the unemployment insurance system.

Workers

The model comprises two periods, with the length of each normalized to 1. Agents can neither save nor borrow (this hypothesis will be relaxed subsequently). During the first period, all individuals work, get an exogenous wage w, and pay a flat rate tax τ, which serves to finance the unemployment benefit b paid out to every job seeker. In the first period, every individual thus obtains a level of utility $v(w-\tau)$ with $v'>0$. It is assumed that all agents are risk averse, or in formal terms $v''<0$. At the end of the first period, individuals lose their jobs. They can then make a search effort that determines the duration of their job search during the second period. The more energetically they search for a job, the shorter this duration, denoted $D \in [0,1]$. Following Chetty (2006), we assume that search costs, the leisure value of unemployment, and the advantage of improved job matches flowing from additional search are captured by a concave, increasing function denoted $\phi(D)$. It is also assumed, for simplicity, that in the second period a job seeker pays no taxes. Knowing b and τ, every individual chooses a duration of unemployment D that maximizes his expected utility for the second period. On the assumption that the duration of unemployment is equivalent to the probability of being unemployed, the program of the second period of an agent's life is written:

$$\max_{D} (1-D)\, v(w) + Dv(b) + \phi(D)$$

Deriving with respect to D, we obtain the first-order condition:

$$\phi'(D) = v(w) - v(b) \tag{13.1}$$

This condition defines the duration of job search as a function of parameter b; we denote it $D(b)$ and we assume that it falls in the interval $[0, 1]$. It signifies that at the optimum, the agent selects a duration of unemployment that equalizes his marginal gain, $\phi'(D)$, and his marginal cost, $v(w) - v(b)$, expressed in terms of utility. As function ϕ is taken to be concave, ϕ'' is negative, and as v' is positive, the result is $D'(b) > 0$. A rise in unemployment benefit thus leads directly to an increase in the duration of unemployment.

The Agency

The objective of the agency charged with managing the system of unemployment compensation is to determine the amount of unemployment benefit and the flat-rate tax that will maximize the expected utility of an unemployed person under the constraint of a balanced budget. Assuming that there is no discounting, the program of the agency takes the following form:

$$\max_{b,\tau} v(w - \tau) + [1 - D(b)]\, v(w) + D(b)v(b) + \phi\,[D(b)] \tag{13.2}$$

subject to $\tau = D(b)b$.

The budget constraint is easily grasped. If we assume that the size of the active population is equal to 1, τ represents the sum total of the taxes levied on wage earners. With this hypothesis, D represents the number of job seekers. The budget constraint of the agency expresses the fact that the mass of taxes collected from wage earners in the first period is paid out in full to the job seekers. The agency's budget is thus balanced.

The budget constraint allows us to eliminate τ from the maximization criterion. The program of the agency consists of maximizing the expected utility of the agent with respect to the unemployment benefit, for the level of tax that balances the budget, and for the search effort the agent chooses to make. The expected utility of the agent is then written:

$$W(b) = v\,[w - bD(b)] + [1 - D(b)]\, v(w) + D(b)v(b) + \phi\,[D(b)]$$

In deriving this expected utility, and employing condition (13.1) that defines the duration of unemployment chosen by agents,[2] we can calculate the impact of a small hike in unemployment benefit on expected utility:

$$W'(b) = -\left(1 + \eta_b^D\right) D(b)v'\,[w - bD(b)] + D(b)v'(b) \quad \text{with } \eta_b^D = \frac{bD'(b)}{D(b)}$$

The term $\left(1 + \eta_b^D\right) D(b)v'\,[w - bD(b)]$ represents the marginal cost of a hike in the amount of benefit b, a cost borne by the wage earners who finance the unemployment insurance system. The marginal cost of raising the tax to finance an increase in b is given by the direct cost v' plus an added term $\eta_b^D v'$ arising from the increase in unemployment duration that follows the hike in b. This marginal cost increases with η_b^D, the elasticity of unemployment duration with respect to unemployment benefits. The term

[2] In other words, we make use here of the envelope theorem. See appendix A, section 1.5, at the end of the book.

$D(b)v'(b)$ represents the marginal gain flowing from a hike in the amount of benefit b, which increases the utility of the unemployed job seeker. At the optimum, the marginal cost of a hike in the amount of benefit b must be equal to the marginal gain, or:

$$\left(1+\eta_b^D\right)v'\left[w-bD(b)\right]=v'(b) \tag{13.3}$$

It is worth pointing out that if the duration of unemployment were inelastic ($\eta_b^D=0$), the optimal amount of benefit would offer perfect insurance to all agents, since we would have $w-\tau=b$. Conversely, if $\eta_b^D>0$, then $w-\tau>b$ and agents are imperfectly insured against the risk of job loss. Thus the elasticity η_b^D, which signals the existence of the moral hazard inherent in paying out unemployment benefits, limits the capacity of the agency to insure against the risk of job loss.

1.2.2 The Baily Formula

Let us first establish the theoretical formula describing optimal unemployment insurance. It will then be possible to show that this formula holds good in a model where agents can save and borrow.

Three Parameters Defining Optimal Unemployment Insurance

Let us denote by $c_e=w-\tau$ and $c_u=b$ respectively the consumption of a tax-paying wage earner and the consumption of a job seeker. With this notation, equation (13.3) characterizing optimal unemployment insurance becomes:

$$\frac{v'(c_u)-v'(c_e)}{v'(c_e)}=\eta_b^D \tag{13.4}$$

A Taylor expansion makes it possible to bring out the main ingredients of optimal unemployment insurance. We have:

$$v'(c_u)-v'(c_e)\simeq(c_u-c_e)v''(c_e)+\frac{1}{2}(c_u-c_e)^2v'''(c_e)$$

Let us denote by $\sigma(c)=-cv''(c)/v'(c)$ the coefficient of relative risk aversion and by $\rho(c)=-cv'''(c)/v''(c)$ the coefficient of relative prudence. Taylor's development can then be written:

$$\frac{v'(c_u)-v'(c_e)}{v'(c_e)}\simeq\frac{c_e-c_u}{c_e}\sigma(c_e)+\frac{1}{2}\left(\frac{c_e-c_u}{c_e}\right)^2\sigma(c_e)\rho(c_e)$$

By substituting this expression in equation (13.4), we find that optimal unemployment insurance is characterized by:

$$\frac{c_e-c_u}{c_e}\sigma(c_e)+\frac{1}{2}\left(\frac{c_e-c_u}{c_e}\right)^2\sigma(c_e)\rho(c_e)\simeq\eta_b^D \tag{13.5}$$

In the appendix to this chapter, we lay out in precise fashion the economic interpretation of the coefficients of relative risk aversion and relative prudence. The first

corresponds to its intuitive meaning, the shunning of risk, whereas relative prudence measures how risk aversion varies. Kimball (1990) has shown that the coefficient of relative prudence is also an indicator of an agent's inclination to save up, and for this reason $\rho(c)$ is sometimes referred to as measuring the "precaution motive."

Limiting ourselves to a first-order Taylor expansion, which amounts to the assumption that $\rho(c_e)$ and $v'''(c_e)$ are negligible quantities, we obtain the original formula of Baily (1978):

$$\frac{c_e - c_u}{c_e}\sigma(c_e) \simeq \eta_b^D \tag{13.6}$$

This formula defines the optimal relative difference between the consumption of a person in work and the consumption of a job seeker. This difference diminishes with risk aversion and increases with the elasticity of unemployment duration with respect to unemployment benefits. When aversion to risk is greater, drops in consumption do effectively cause more drastic losses of welfare. From the second perspective, the rise in consumption on the part of job seekers is more costly when the elasticity of unemployment duration with respect to unemployment benefit is greater. Hence optimality would suggest reducing the amount of unemployment benefit when this elasticity is strong.

More generally, formula (13.5) shows that three parameters determine the optimal difference between the consumption of wage earners and the consumption of job seekers: risk aversion, $\sigma(c_e)$, the precaution motive $\rho(c_e)$, and the elasticity of unemployment duration η_b^D. These three magnitudes are what we might call "sufficient statistics" for determining the optimal difference in consumption between a wage earner and a job seeker.

Optimal Unemployment Insurance with Savings

A priori, one might suppose that the possibility of saving changes the characterization of optimal unemployment insurance (on this point see Baily, 1978). We will now discover that this is not the case. Let us return to the basic model but assume that each agent has the possibility of building up savings s in the first period, which she consumes in the second period, knowing b and τ perfectly. Each agent then chooses $s(b,\tau)$ and $D(b,\tau)$ solutions of the program:

$$\max_{s,D} v(w - \tau - s) + (1 - D)\, v(w + s) + Dv(b + s) + \phi(D)$$

Let us designate by $V(b,\tau)$ the indirect utility associated with these solutions; the agency then chooses the pair (b,τ) that maximizes $V(b,\tau)$ under the budget constraint $\tau = D(b,\tau)b$. As before, this budget constraint allows us to view τ as a function of b, which we denote $\tau(b)$. The program of the agency then consists of selecting b in such a way as to maximize $V[b,\tau(b)]$. Employing the envelope theorem, we arrive at:

$$\left(1 + \eta_b^D\right) v'\left[w - s - bD(b)\right] = v'(b + s) \quad \text{with} \quad \eta_b^D = \frac{bD'(b)}{D(b)}$$

Let us again denote by $c_e = w - s - \tau$ and $c_u = b + s$ the consumptions of a tax-paying wage earner and a job seeker respectively: we return exactly to equation (13.4) characterizing optimal unemployment insurance. The Baily formula (13.6) and the augmented Baily formula (13.5) therefore both remain valid. Evidently, whenever savings

s are greater than zero, the optimal value of unemployment benefit is not the same as in the model without savings (i.e., on the assumption that $s = 0$). Still, characterizations of optimal unemployment insurance grounded on the consumption gap between wage earners and job seekers retain all their validity.

More generally, Chetty (2006a) has shown, with the help of a more elaborate dynamic model, that the Baily formula (13.5) augmented by the precautionary saving motive retains validity in very general environments. For example, it holds good when we bring into account the insurance that flows from having a working spouse or when we assume that the chances of finding a job grow with the duration of job search.

1.3 The Optimal Level of Unemployment Benefit in Practice

If we dispose of both quantitative data on the parameters that determine the optimal gap in consumption between wage earners and job seekers and data on the relation between this consumption gap and unemployment benefit, we can in principle form an idea about the optimal setting of unemployment benefit by using either the simple Baily formula or the augmented version. Nonetheless, utilization of the Baily formula remains challenging in practice, since it requires the availability of data on parameters such as risk aversion, and the relation between consumption and unemployment benefit, that are very tricky things to estimate.

Under these circumstances, Chetty (2008) and Shimer and Werning (2007) have elaborated alternative formulas, the purpose of which is to characterize optimal unemployment insurance with parameters that are easier to quantify. Chetty (2008) proposes a formula that relies uniquely on elasticities in the duration of unemployment. Shimer and Werning (2007) obtain a formula that depends especially on the elasticity of the reservation wage.

1.3.1 The Application of the Baily Formula

It is a practical possibility—if we are in a position to assign quantitative values to the elasticity of unemployment duration, to relative risk aversion, and to the relation between unemployment benefit and the consumption gap between wage earners and job seekers—to form an idea of the optimal value of unemployment benefit, by employing the Baily formula (13.6). Gruber (1997) undertook this exercise on American data for the period 1968–1987 issuing from the Panel Study of Income Dynamics (PSID), a survey yielding information about expenditure on food consumption in households. Looking at individuals who lost their jobs between the two boundary dates of the survey, Gruber estimates by a simple regression, while controlling for a battery of individual characteristics, the relation between $(c_e - c_u)/c_e$ for food consumption and the replacement ratio b. He finds $(c_e - c_u)/c_e = 0.22 - 0.27b$ (Gruber, 1997, table 1, column 2). For the elasticity of unemployment duration, Gruber selects the value 0.9 deriving from the paper by Meyer (1990), who uses administrative data from the Continuous Wage and Benefit History (CWBH); these data cover men who received unemployment benefits in 12 states from 1978 to 1983. This elasticity is multiplied by the take-up elasticity of unemployment benefit (i.e., by how much the take-up increases when the benefit amount increases), equal to 0.48, which finally yields a coefficient $\eta_b^D = 0.43$. The Baily formula (13.6) then resolves into a simple increasing relation between the replacement ratio

and the degree of relative risk aversion, or $b = (0.22/0.27) - (0.43/0.27\sigma)$, or again $b = 0.81 - (1.59/\sigma)$. With a degree of relative risk aversion of an empirical magnitude very likely inferior to 2 (Chetty, 2006b), the optimal replacement ratio would have to be very tiny indeed, inferior to 0.015! The fact is that according to the survey data used by Gruber, the average replacement ratio comes to 0.426. From this Gruber deduces that over the period in question, unemployment benefits in the United States were very likely too high with respect to the theoretically optimal setting yielded by the Baily formula.

But the Baily formula leaves out the reaction of precautionary savings, since it assumes $\rho(c_e) = 0$, and this may produce a strong underestimate of the optimal amount of benefit. For example, on the assumption that the utility function is of the CRRA (Constant Relative Risk Aversion) type, or $v(c) = \frac{c^{1-\sigma}}{1-\sigma}$, with σ lying in the interval $[1, 5]$ and $\eta_b^D = 0.5$, Chetty (2006a) calculates that the optimal unemployment benefit may sometimes be underestimated by more than 30% if the coefficient of relative prudence is left out of account. Hence it will be preferable to make use of relation (13.5) instead of the Baily formula (13.6).

Results obtained with this approach must be interpreted with care, inasmuch as there exists at present considerable reserve about the pertinent values of parameters issuing from some empirical studies based on samples that are not necessarily representative and not always well suited to the populations concerned.

1.3.2 Liquidity and Moral Hazard Elasticities

Chetty (2008) adopts an approach very close to that of Baily but which permits calculation of the optimal replacement ratio without bringing in the measurement of risk aversion or the relation between consumption and unemployment benefit.

Liquidity and Moral Hazard Effects

Let us consider a variant of Baily's basic model (see section 1.2 above) with a single period where it is assumed that agents dispose of an initial endowment $A \geq 0$. In this context, the choice of the optimal duration of unemployment is the solution of the program:

$$\max_D (1 - D)\, v(w - \tau + A) + Dv(b + A) + \phi(D)$$

Deriving with respect to D, we obtain the first-order condition:

$$v(w - \tau + A) - v(b + A) = \phi'(D) \tag{13.7}$$

Differentiating this formula with respect to w, we find:

$$\frac{\mathrm{d}D}{\mathrm{d}w} = \frac{v'(w - \tau + A)}{\phi''(D)} < 0 \tag{13.8}$$

We observe that a wage rise leads to a shortening of the duration of unemployment: the reason is that it increases the return to job search.

Next, differentiating (13.7) with respect to A brings us to:

$$\frac{\mathrm{d}D}{\mathrm{d}A} = \frac{v'(w-\tau+A) - v'(b+A)}{\phi''(D)} > 0 \tag{13.9}$$

We see that an increase in the endowment A increases the duration of unemployment, since it narrows the gap between the marginal utility of a wage earner and that of a job seeker. The existence of this gap is a measure of the inevitable imperfection of any attempt to insure against the risk of a drop in income caused by the loss of a job. At the limit, this gap is null when individuals can insure themselves perfectly. If they can, endowment A has no impact on the duration of unemployment.

Using the two previous formulas, the impact of unemployment benefit on the duration of unemployment can be written as follows:

$$\frac{\mathrm{d}D}{\mathrm{d}b} = -\frac{v'(b+A)}{\phi''(D)} = \frac{\mathrm{d}D}{\mathrm{d}A} - \frac{\mathrm{d}D}{\mathrm{d}w} \tag{13.10}$$

This decomposition shows that unemployment benefits raise unemployment duration by producing two different effects. The first term, $\mathrm{d}D/\mathrm{d}A$, corresponds to a "liquidity effect," for a higher benefit amount increases the agent's resources, allowing her to maintain a higher level of consumption while unemployed and reducing the pressure on her to find a job quickly. The second term, $-\mathrm{d}D/\mathrm{d}w$, is the "moral hazard effect," which reduces the wage (i.e., the gap between the wage and the unemployment benefit) and consequently the search effort. The liquidity effect implies that unemployment benefit reduces the need for agents to rush back to work because they cannot smooth out their consumption. The moral hazard effect implies that unemployment benefit subsidizes unproductive leisure.

Optimal unemployment insurance is always obtained by maximizing expected utility under the budget constraint of the agency and taking into account the relation—induced by job search behavior—between the duration of unemployment and unemployment benefit. This expected utility may be written:

$$W(b) = [1 - D(b)]\, v\,[w - \tau(b) + A] + D(b)v(b+A) + \phi\,[D(b)]$$

where $\tau(b) = D(b)b/[1 - D(b)]$ and $D(b)$ always designates the duration of unemployment flowing from the optimal search effort of the agent defined by condition (13.7). Note that the expression for $\tau(b)$ shows that the longer the unemployment duration, the larger the numbers of the jobless, and the higher the tax rate on workers needed to finance benefits.

Again using the envelope theorem and the two formulas (13.8) and (13.9), we can write the derivative of the marginal utility with respect to unemployment benefit in the form:

$$W'(b) = \left(\frac{\mathrm{d}D/\mathrm{d}A}{-\mathrm{d}D/\mathrm{d}w} - \eta_b^D\right) D(b)v'\,(c_e) \tag{13.11}$$

On the right-hand side of this equality there appears the relation between the liquidity effect and the moral hazard effect. When this (positive) term is multiplied

by factor $D(b)v'(c_e)$, it represents the marginal gain from an increase in unemployment benefit. This marginal gain swells with the liquidity effect and shrinks with the moral hazard effect. The term $\eta_b^D D(b)v'(c_e)$ represents the marginal cost: it increases with the elasticity of unemployment duration with respect to unemployment benefit. Formula (13.11) shows that it is necessary to increase unemployment benefit if the relation between the liquidity effect and the moral hazard effect dominates the elasticity of unemployment duration with respect to unemployment benefit. If this is not the case, it is necessary to reduce unemployment benefit. The optimal level of unemployment benefit is attained when the marginal gain is just equal to the marginal cost (i.e., $W'(b) = 0$).

The expression (13.11) has the merit of characterizing optimal unemployment benefit solely on the basis of different elasticities of the duration of unemployment, which may be easier to estimate than the parameters employed in the Baily formula (13.6).

An Evaluation of the Optimal Amount of Unemployment Benefit

Chetty (2008) estimates the liquidity and moral hazard effects by adopting a range of strategies. He estimates the elasticity of unemployment duration with respect to unemployment benefit, or in other words, the total benefit effect, for liquidity-constrained and liquidity-unconstrained individuals. For liquidity-unconstrained individuals, who are able to smooth out their consumption perfectly when they lose their jobs, the total benefit effect is equal to the moral hazard effect, as portrayed in equation (13.10). For liquidity-constrained individuals, the total benefit effect is equal to the sum of the liquidity effect and the moral hazard effect. Therefore, the difference between the total benefit effect on the unconstrained and the constrained individuals identifies the liquidity effect. Chetty creates groups of individuals differentiated according to their liquid wealth, net of unsecured debt, at the time of job loss. He also considers groups of individuals differentiated by their spousal work status because those with a second income source are more likely to be able to borrow with at least one working person in their household. Another strategy consists in comparing the behavior of job losers who got severance payments with those who did not. The latter are more liquidity constrained.

The results obtained using these empirical strategies suggest that the link between unemployment benefit and unemployment duration is driven by a subset of the population that has limited ability to smooth out consumption. This pattern is suggestive of a substantial liquidity effect, which might explain 60% of the marginal effect of unemployment benefit on unemployment duration at current benefit rates in the United States. Chetty estimates that the ratio of the liquidity effect over the moral hazard effect that shows up in equation (13.11) is about 0.6, whereas η_b^D is about 0.5 for the U.S. economy. This suggests that unemployment benefit in the United States is a little below, but not far below, its optimal level.

1.3.3 Optimal Unemployment Benefits and the Reservation Wage

Shimer and Werning (2007) cast a complementary light on optimality in unemployment insurance by exploiting the intertemporal dimension of the job search model presented in chapter 5. This allows them to define the optimal level of unemployment benefit as a function of parameters differing from those of Baily (1978) and Chetty (2008).

The Reservation Wage with Taxes and Unemployment Benefits

In essence the model of Shimer and Werning (2007) coincides with the basic model of job search set out in chapter 5, with the addition of a budget constraint for the agency in charge of managing unemployment insurance and with the assumption that agents present aversion to risk. At every date a job seeker confronts a stationary distribution $H(.)$ of possible wages. If he receives an offer of wage w and accepts it, he pays a constant flat-rate tax τ at every date for as long as he stays with the firm. It is always assumed that the wage earner obtains an instantaneous utility $v(w-\tau)$ with $v'>0$ and $v''<0$. The expected intertemporal utility, $V_e(w)$, procured by a job paying wage w, is thus written:

$$rV_e(w) = v(w-\tau) + q\,[V_u - V_e(w)]$$

In this expression, V_u designates the intertemporal utility of a job seeker, and the exogenous constant parameters, r and q, designate respectively the interest rate and the job destruction rate. The foregoing equation can again be written in the form:

$$V_e(w) - V_u = \frac{v(w-\tau) - rV_u}{r+q}$$

From it we deduce that a job seeker accepts every wage w such that $v(w-\tau) > rV_u$, but that he keeps on seeking if $v(w-\tau) < rV_u$. The reservation wage, denoted x, is thus defined by:

$$v(x-\tau) = rV_u$$

As in the basic job search model from chapter 5, we assume that a job seeker has an exogenous constant probability, denoted λ, of receiving a wage offer at every date. Leaving aside the costs entailed by job search and assuming that at every date a job seeker receives a constant unemployment benefit equal to b, his intertemporal utility takes the expression:

$$rV_u = v(b) + \lambda \int_x^{+\infty} [V_e(w) - V_u]\,\mathrm{d}H(w)$$

From that it follows that the reservation wage is defined by the equation:

$$v(x-\tau) = v(b) + \frac{\lambda}{r+q}\int_x^{+\infty} [v(w-\tau) - v(x-\tau)]\,\mathrm{d}H(w) \qquad \textbf{(13.12)}$$

This equation defines the reservation wage x as a function of parameters (b,τ) characterizing the unemployment insurance system. We may thus denote it $x(b,\tau)$. Readers can easily verify that the reservation wage is increasing with b by following the same procedure as in chapter 5. They can also verify that the after-tax reservation wage, that is, $x(b,\tau)-\tau$, is decreasing with τ, which is explained by the fact that a hike in the tax

diminishes the gain procured by accepting a job. On the other hand, the direction in which the reservation wage varies as a function of tax τ remains ambiguous.[3]

The Agency's Budget Constraint

The optimal level of unemployment benefit corresponds to a value of b that maximizes the intertemporal utility of the job seeker under the constraint that the net actualized cost of a job seeker be null. We begin by finding the expression of the costs generated by the unemployment insurance system. It will be convenient to define the net actualized cost of a job seeker, denoted C_u, and the net actualized cost of a wage earner, denoted C_e, in recursive fashion, with the system of equations:

$$rC_u = b + \lambda\,[1 - H(x)]\,(C_e - C_u) \tag{13.13}$$

$$rC_e = -\tau + q(C_u - C_e) \tag{13.14}$$

Equation (13.13), describing the time path of the net actualized cost of a job seeker, is to be understood as follows: at every instant a job seeker costs b but has a probability $\lambda\,[1 - H(x)]$ of finding a new job, in which case she becomes a wage earner and her net cost to the unemployment insurance system then becomes equal to C_e. Equation (13.14) has the identical explanation: at every instant a wage earner pays in τ to the unemployment insurance system, but she can lose her job with a probability q, in which case she becomes a job seeker and her actualized cost to the unemployment insurance system then amounts to C_u. The constraint that the net actualized cost of a job seeker be null is obtained by setting $C_u = 0$ in equations (13.13) and (13.14). We thus arrive at the budget constraint of the agency:

$$bD = \frac{\tau}{r+q} \quad \text{with} \quad D = \frac{1}{\lambda\,[1 - H(x)]} \tag{13.15}$$

This budget constraint includes the average duration D of an episode of unemployment. The quantity bD represents the average cost of an episode of unemployment, while the quantity $\tau/(r+q)$ represents the actualized average gain of an episode of work. When r goes to 0, the budget constraint (13.15) indicates that the average cost of an episode of unemployment is equal to the average gain of an episode of work.

Since the reservation wage x is a function of variables (b, τ) characterizing the system of unemployment insurance—see (13.12)—the average duration of an episode of unemployment is also a function of (b, τ). We may therefore denote it $D(b, \tau)$. The budget constraint (13.15) then defines a relation, denoted $\tau(b)$, between the tax τ and the benefit b which is written as follows:

$$\tau(b) = (r+q)bD\,[b, \tau(b)] \tag{13.16}$$

[3] Deriving (13.12) with respect to τ we get:

$$\left\{1 + \frac{\lambda\,[1 - H(x)]}{r+q}\right\}\left(\frac{\partial x}{\partial \tau} - 1\right)v'(x - \tau) = \frac{-\lambda}{r+q}\int_x^{+\infty} v'(w - \tau)\mathrm{d}H(w)$$

This proves that $\frac{\partial x}{\partial \tau} - 1 < 0$. The net wage $x - \tau$ is thus decreasing with τ, but we cannot draw any conclusion about the sign of $\frac{\partial x}{\partial \tau}$.

Let us denote D_b and D_τ the partial derivatives of function D with respect to its two arguments. Deriving the last relation with respect to b, we find:

$$\tau'(b) = \frac{bD_b + D}{\frac{1}{r+q} - bD_\tau} \tag{13.17}$$

The definition of D, given by (13.15), entails that $D_b > 0$ since D varies like the reservation wage. Conversely, we can say nothing about D_τ, since the direction in which the reservation wage varies is ambiguous with τ. Henceforth we assume $\tau'(b) > 0$, which signifies that every hike in the tax enlarges the stream of income from which job seekers are compensated. This hypothesis amounts to stating that we are situated on the "good side" of the Laffer curve.

To advance our analysis of the properties of optimal unemployment insurance, we will assume, like Shimer and Werning (2007), that the utility function exhibits constant absolute risk aversion (CARA). Thus we set $v(c) = -\gamma e^{-\gamma c}$, where the constant $\gamma > 0$ represents the absolute degree of risk aversion. With this hypothesis, relation (13.12) defining the reservation wage is written:

$$e^{-\gamma x} = e^{-\gamma(b+\tau)} + \frac{\lambda}{r+q}\int_x^{+\infty} (e^{-\gamma w} - e^{-\gamma x})\, \mathrm{d}H(w) \tag{13.18}$$

We see that with a utility function of the CARA type, the reservation wage depends only on the sum $(b + \tau)$. Thus we have $D_\tau = D_b$, since $D = 1/\lambda\,[1 - H(x)]$ depends exclusively on x. That being the case, equation (13.17) takes the form:

$$\tau'(b) = \frac{D(1+\eta)}{\frac{1}{r+q} - D\eta} \quad \text{with} \quad \eta = \frac{bD_b[b,\tau(b)]}{D[b,\tau(b)]} > 0 \tag{13.19}$$

A New Formula to Characterize Optimal Unemployment Benefit

The optimal setting of unemployment benefit corresponds to a value of b that maximizes the intertemporal utility of the job seeker under the constraint that the net actualized cost of a job seeker be null. Now, to any amount of benefit b there corresponds an amount of tax $\tau(b)$ given by (13.16), and since $v(x - \tau) = rV_u$, the optimal unemployment benefit simply maximizes the net reservation wage $x[b,\tau(b)] - \tau(b)$. Designating by x_b and x_τ the partial derivatives of function $x(b,\tau)$ with respect to b and τ, we arrive at the first-order condition:

$$\psi(b) = x_b[b,\tau(b)] + \tau'(b)x_\tau[b,\tau(b)] - \tau'(b) = 0$$

If the utility function $v(.)$ is of the CARA type, the reservation wage given by (13.18) depends exclusively on the sum $(b + \tau)$. We then have $x_b = x_\tau$ and the first-order condition becomes:

$$\psi(b) = x_b[b,\tau(b)]\left[1 + \tau'(b)\right] - \tau'(b) = 0$$

As $\tau'(b)$ is given by relation (13.19), the unemployment benefit verifies:

$$x_b = \frac{D}{\frac{1}{r+q} + D}(1+\eta) \tag{13.20}$$

The left-hand side of this equality represents the (gross) marginal gain from an increase in unemployment benefit to the intertemporal utility of a job seeker, while the right-hand side represents the marginal cost of a simultaneous tax hike. To better grasp the significance of this optimality condition, let us consider an increase in the amount of unemployment benefit. Higher benefits reduce the cost of remaining unemployed and therefore raise the pretax reservation wage x. Thus, if the pretax reservation wage is very responsive to unemployment benefits, raising unemployment benefits has a strong positive effect on workers' welfare. This effect is captured by the left-hand side of (13.20). However, this increase in benefits is financed by a hike in the tax τ. The more responsive the duration of unemployment D to the amount of benefit, the greater the need to increase the tax. Condition (13.20) signifies that at the optimum these two effects have the same magnitude.

Shimer and Werning have shown that formula (13.20) holds good for numerous extensions of the basic model. In particular, it remains valid if agents are able to save or borrow and if an unemployed worker's search effort affects the arrival rate of job offers.

Another Evaluation of the Optimal Amount of Unemployment Benefit

Shimer and Werning (2007) use equation (13.20) to make an assessment of whether, in practice, the amount of unemployment benefit in the United States ought to be increased or cut back. It will be optimal to increase the amount if the marginal gain, represented by the left-hand member of (13.20), is superior to the marginal cost, represented by the right-hand member of this expression.[4]

To implement this test, though, researchers have to be able to produce reliable estimates of the magnitudes appearing in the formula (13.20). In this light, we may begin by remarking that at stationary equilibrium, exits from employment are equal to entries into unemployment. Denoting by u the unemployment rate, we thus have:

$$q(1-u) = \lambda\,[1 - H(x)]\,u = \frac{u}{D} \tag{13.21}$$

Shimer and Werning observe that between 1948 and 2005, the average unemployment rate in the United States was 5.6%, and the average duration of an episode of unemployment was 13.4 weeks. Setting $u = 0.056$ and $D = 13.4$ in (13.21) yields $q = 0.00443$. For the weekly interest rate, Shimer and Werning set $r = 0.001$, which corresponds to an annual interest rate of 5.3%. They thus obtain $r + q = 0.00543$ and so $D/\{[1/(r+q)] + D\} = 0.068$.

[4] More formally, it may be noted that the second-order condition of the agency's program dictates that we have $\psi'(b^*) < 0$ for the optimal value of unemployment benefit, and so, by continuity, $\psi'(b) < 0$ in a neighborhood of b^*. As we have $\psi(b^*) = 0$, if we observe that the left-hand side of equation (13.20) is larger than the right-hand side—i.e., $\psi(b) > 0$—then an increase in unemployment benefit is welfare-improving.

For elasticity η Shimer and Werning use, like Gruber (1997), the paper by Meyer (1990) which furnished the estimate $\eta = 0.88$. The right-hand side of equality (13.20) thus comes to $0.067 \times 1.88 = 0.126$. It remains to obtain an estimate of the left-hand side of equality (13.20), which comes down to estimating the variation in the reservation wage when unemployment benefit is increased. Shimer and Werning make use of the paper of Feldstein and Poterba (1984), who looked at how self-reported reservation wages respond to unemployment benefits. They study a supplement to the May 1976 Current Population Survey (CPS), which asked 2,228 unemployment insurance recipients "What is the lowest wage or salary you would accept (before deductions)?" Feldstein and Poterba estimate that x_b lies in the interval $[0.13, 0.42]$. Since the right-hand side of equality (13.20) is equal to 0.126, Shimer and Werning conclude that in the actual U.S. system, the marginal cost of unemployment benefit is lower than the marginal gain and that the amount of unemployment benefit ought to be increased. This result complements those presented previously. Overall, however, these results rely on estimates that are still very approximate and that would need to be made more precise in order to obtain more credible evaluations of the optimal level of unemployment benefit. As well, we have limited ourselves for now to a situation where the amount of benefit remains constant during the episode of unemployment—a configuration that is not necessarily optimal, as we will see in the next section.

1.4 Optimal Unemployment Insurance in a Dynamic Environment

A relevant analysis of unemployment insurance should focus on the time profile of the benefit payments, which can provide at least as much incentive as their amount. This is the reason most unemployment insurance systems limit the period during which the unemployed can receive benefits, and provide for such benefits to tail off, the longer that period lasts. Research in this area does suggest that a time profile in which the amount of benefit decreases with the duration of unemployment may be optimal, but not in every case. It also suggests that the gains procured by a tapering profile are limited rather than constant.

A related topic is the path of optimal unemployment insurance over the course of the economic cycle. Some countries, like Canada and the United States, tie the duration of benefit to the prevailing level of unemployment: benefits may be paid for a longer period when the unemployment rate is higher. Research, both theoretical and empirical, tells us something about the conditions under which this practice is efficient.

1.4.1 The Optimal Profile of Unemployment Benefits

The dynamic job models with moral hazard and job search of Shavell and Weiss (1979), Hopenhayn and Nicolini (1997, 2009), and Wang and Williamson (1996, 2002) do in fact prove that optimal unemployment benefit must necessarily decrease as the unemployment spell lengthens. However, Shimer and Werning (2008) have shown that this result does not always hold good when individuals can have free access to the borrowing and lending of a riskless asset in order to smooth out their consumption. Moreover, calibration exercises suggest that declining profiles provide only very small welfare gains when the unemployment insurance agency can tax and subsidize wages.

The Model of Optimal Unemployment Insurance

To establish the main dynamic properties of optimal unemployment benefit, we follow the model in discrete time of Hopenhayn and Nicolini (2009) where the optimal contract should minimize the average cost of a jobless person while at the same time offering him an exogenous level of expected utility $\bar{V}$.

At every period t, the effort an agent makes to find a job can take no more than two values: either the constant value $a > 0$, in which case the agent finds employment at rate p, where $p > 0$ is an exogenous constant; or the value 0, in which case the agent gets no job offers and remains unemployed. The "principal," in other words the agency charged with managing unemployment insurance, proposes a *contract* to every person entering unemployment (by convention, unemployment begins on date $t = 0$) specifying the values b_t of the unemployment benefit to be received if the person is still looking for a job at period $t > 0$, and the values g_t of the transfers to be received if employment resumes on date t. It should be noted that the benefit payments b_t, and the transfers g_t should employment resume, are both conditional on the length t of the unemployment spell. We may also point out that if $g_t < 0$, what we have is a tax; and if $g_t > 0$, a subsidy.

It is worth noting that the contract between the unemployment insurance agency and the job seeker is much like a relatively sophisticated experience rating system. The unemployment insurance contracts of the real world share some of the characteristics highlighted in our model. The tailing-off of benefit payments the longer the spell of unemployment lasts is a measure that is not unusual, even though the amount usually drops by just one level, from full to partial (see table 13.1). On the other hand, systems in which subsidies are received or taxes collected after a return to work, both of them varying with the length of the unemployment spell, are less common but do exist. Certain countries have put in place "return to work premiums" aimed precisely at encouraging the unemployed to find a job quickly. Premiums of this type exist in Japan, where the premium decreases as the spell of unemployment persists; it is paid to people who return to work with at least a third, or in some cases at least half, of their benefit entitlement period remaining (see Duell et al., 2010). In Australia and France it is possible, for a period, to retain a portion of one's unemployment benefit while working part-time. Finally, the United States has tried out similar systems locally, and it has been found that they do in fact encourage the jobless to find work more rapidly (a detailed study of these experiments can be found in Meyer, 1995).

As previously, we assume that effort is not verifiable. Suppliers of labor do not have access to financial markets and therefore they do not save or invest. All jobs offer the same exogenous constant wage w; there is no job seeking by persons already on the job; and jobs are never destroyed. This last assumption is not essential and is chosen to simplify the presentation. If a job seeker finds work after an unemployment spell of t periods, she receives a net wage of $(w + g_t)$ and keeps her new job indefinitely. Denoting $\beta \in [0, 1]$ the discount factor, the discounted expected utility of a person finding a job after t periods of unemployment, denoted V_e^t, is thus given by:

$$V_e^t = \frac{v(w + g_t)}{1 - \beta} \tag{13.22}$$

In this expression, $v(.)$ represents the utility of the agent during one period of employment. The function $v(.)$ is such that $v' > 0$ and $v'' < 0$, which signifies that the

agent is risk averse. To this level of utility there corresponds a cost to the principal defined by:

$$C_e^t = \frac{g_t}{1-\beta}$$

Eliminating g_t between the last two equations, we see that the cost C_e^t can be expressed as a function of V_e^t, or $C_e^t = C_e(V_e^t)$. This relation simply conveys the fact that to every level of utility there corresponds a cost borne by the principal. It is easily verified that:

$$C_e'(V_e^t) = \frac{1}{v'(w+g_t)} > 0 \tag{13.23}$$

For what follows, it will also be helpful to note that $C_e''(V_e^t) > 0$, hence the cost function $C_e(.)$ is strictly convex.

The evolution of the expected utility of a job seeker making search effort a during period t, denoted V_u^t, is described by the following equation:

$$V_u^t = v(b_t) - a + \beta\left[pV_e^{t+1} + (1-p)V_u^{t+1}\right] \tag{13.24}$$

Equation (13.24) indicates that a job seeker making effort a during period t attains, over that period, the utility level $v(b_t) - a$. With probability p, he can then find a job that starts at period $t+1$ and procures an expected utility equal to V_e^{t+1}. With the complementary probability $(1-p)$, he remains unemployed and his discounted expected utility then amounts to V_u^{t+1}.

The Incentive Constraint

When the search effort is not directly checked on by the agency, the unemployed person has the opportunity to "cheat" by making no effort while continuing to receive unemployment insurance benefits. At each date, a job seeker chooses to make search effort a only if she thus obtains an expected utility V_u^t superior to the utility denoted V_s^t that she obtains by "cheating." The latter is defined by:

$$V_s^t = v(b_t) + \beta V_u^{t+1} \tag{13.25}$$

This equation indicates that a job seeker who does not make search effort a during period t receives unemployment benefit payments during this period—precisely because her effort is not verifiable—and attains a utility level equal to $v(b_t)$. She therefore has no chance of finding a job at date $t+1$ and so obtains the discounted utility expected by an unemployed person at this date.

To incentivize the job seeker to make effort a at any period $t \geq 0$, the agency must offer her unemployment benefits and a transfer giving her an intertemporal utility V_u^t superior to intertemporal utility V_s^t. Making the difference between equations (13.25) and (13.24), we find that the incentive constraint, $V_u^t - V_s^t \geq 0$, for all t, is finally written:

$$\beta\left(V_e^{t+1} - V_u^{t+1}\right) \geq \frac{a}{p} \tag{13.26}$$

This inequality shows that the need to give the unemployed an incentive to look for work obliges the principal to pay a "rent" at least equal to a/p when they do find work.

The Dynamic Properties of the Optimal System of Taxes and Benefits

We assume that the agency in charge of the unemployment insurance system minimizes its own costs, while both guaranteeing a certain level of utility to the job seeker and incentivizing him to hunt for work. When a job seeker has already undergone t periods of joblessness, the agency anticipates that it will still have to bear a cost of C_u^t before this job seeker finds work. The path of cost C_u^t is given by the following equation, the writing and the understanding of which both conform to (13.24):

$$C_u^t = b_t + \beta \left[pC_e^{t+1} + (1-p)C_u^{t+1} \right]$$

Since by convention entry into unemployment commences at date $t = 0$, the total cost of a job seeker amounts to C_u^0. The goal of the agency is to pinpoint the sequence of $\{b_t, g_t\}$ that minimizes C_u^0 while respecting the incentive constraint (13.26) and the participation constraint $V_u^0 \geq \bar{V}$, which ensures the job seeker a level of expected utility at least equal to $\bar{V}$. This level of utility is exogenous.

Hopenhayn and Nicolini (1997, 2009) came up with a very elegant way of solving this problem by making use of the linkage that exists between the cost of a program and the level of utility that it procures. With the notation $C_u^t = C_u(V_u^t)$, function $C_u(.)$ gives the (future) cost of a job seeker who has already undergone t periods of joblessness when the agency decides to procure him a level of (future) utility V_u^t. The policy of the agency must be time consistent, meaning that the policy adopted from any date t forward must minimize the cost to the agency from that date forward. Let us assume that the policy announced at date $t = 0$ does procure the sequence of utilities $\{V_u^t\}$ for a job seeker. The optimal policy is time consistent if at every date $t > 0$ it minimizes the cost to the agency while in fact offering the job seeker the announced utility V_u^t. For a given V_u^t, the policy put in place from date t forward must therefore be the solution of the program:

$$C_u(V_u^t) = \min_{b_t, V_e^{t+1}, V_u^{t+1}} \left\{ b_t + \beta \left[pC_e(V_e^{t+1}) + (1-p)C_u(V_u^{t+1}) \right] \right\}$$

subject to (13.24) and (13.26).

In light of the forms of the criteria to be minimized and those of constraints (13.24) and (13.26), we observe that it is indifferent whether the agency chooses a sequence $\{b_t, g_t\}$ from date t forward or chooses instead a triplet $\{b_t, V_e^{t+1}, V_u^{t+1}\}$. Everything unfolds just as it would if at date t the policy of the agency consisted of announcing the amount b_t of current unemployment benefit and the promised expected utilities (V_e^{t+1}, V_u^{t+1}).

Let us designate by μ and $\delta \geq 0$ the multipliers associated with constraints (13.24) and (13.26); the Lagrangian of the agency's program is then written:

$$\begin{aligned} \mathcal{L} = & \left\{ b_t + \beta \left[pC_e(V_e^{t+1}) + (1-p)C_u(V_u^{t+1}) \right] \right\} \\ & + \mu \left\{ v(b_t) - a - V_u^t + \beta \left[pV_e^{t+1} + (1-p)V_u^{t+1} \right] \right\} + \delta \left[\beta p \left(V_e^{t+1} - V_u^{t+1} \right) - a \right] \end{aligned}$$

The first-order conditions are obtained by canceling out the derivatives of this Lagrangian with respect to b_t, V_u^{t+1}, and V_e^{t+1}. We thus arrive at:

$$\frac{\partial \mathcal{L}}{\partial b_t} = 1 + \mu v'(b_t) = 0 \Leftrightarrow \mu v'(b_t) = -1 \tag{13.27}$$

$$\frac{\partial \mathcal{L}}{\partial V_u^{t+1}} = \beta(1-p)C_u'(V_u^{t+1}) + \mu\beta(1-p) - \delta\beta p = 0 \Leftrightarrow C_u'(V_u^{t+1}) = \delta\frac{p}{1-p} - \mu \tag{13.28}$$

$$\frac{\partial \mathcal{L}}{\partial V_e^{t+1}} = \beta p C_e'(V_e^{t+1}) + \mu\beta p + \delta\beta p = 0 \Leftrightarrow C_e'(V_e^{t+1}) = -\mu - \delta$$

In addition, as V_u^t is considered as a parameter in this program, the envelope theorem entails:

$$\frac{\partial \mathcal{L}}{\partial V_u^t} = C_u'(V_u^t) = -\mu \tag{13.29}$$

Canceling out the multiplier μ in equations (13.27), (13.28), and (13.29), we arrive at:

$$C_u'(V_u^{t+1}) = \delta\frac{p}{1-p} + \frac{1}{v'(b_t)}, \quad C_u'(V_u^t) = \frac{1}{v'(b_t)} \quad \text{for all } t$$

The last equality appearing in this equation being true for all t, we have $C_u'(V_u^{t+1}) = \frac{1}{v'(b_{t+1})}$, and in consequence:

$$\frac{1}{v'(b_{t+1})} - \frac{1}{v'(b_t)} = \delta\frac{p}{1-p} \tag{13.30}$$

It can be shown that the multiplier δ associated with the incentive constraint (13.26) is strictly negative,[5] which signifies that this constraint is always binding. As $v'' < 0$, we deduce immediately that $b_{t+1} < b_t$. The sequence of optimal unemployment benefits must therefore strictly decrease with time.

In determining the properties of the sequence of transfers g_t flowing to those who do find a new job, the first thing to remember is that the cost function $C_u(V_u^t)$ is increasing and strictly convex.[6] As $\delta < 0$ and $C_u'(V_u^t) = \frac{1}{v'(b_t)}$ for all t, rule (13.30) describing the time path of optimal unemployment benefit entails $C_u'(V_u^t) > C_u'(V_u^{t+1})$. The cost function

[5] To assume that $\delta = 0$ would lead to a contradiction. If $\delta = 0$, we would then have $C_u'(V_u^{t+1}) = C_u'(V_u^t)$ and so $V_u^{t+1} = V_u^t$. We would also have $-C_e'(V_e^{t+1}) = \mu = -\frac{1}{v'(b_t)} = -\frac{1}{v'(b_{t+1})}$. Now we know following (13.23) that $C_e'(V_e^t) = \frac{1}{v'(w+g_t)}$, which entails $w + g_{t+1} = b_{t+1} = b_t$ and so $(1-\beta)V_e^{t+1} = v(w + g_{t+1}) = v(b_t)$. Additionally, (13.24) and (13.26) entail $V_u^t \geq v(b_t) + \beta V_u^{t+1}$, and since $V_u^{t+1} = V_u^t$, we would thus have $(1-\beta)V_u^{t+1} \geq v(b_t)$. As $(1-\beta)V_e^{t+1} = v(b_t)$, we arrive finally at $V_u^{t+1} \geq V_e^{t+1}$, which violates the incentive constraint (13.26). Therefore, we necessarily have $\delta < 0$.

[6] To show this, we can replace the control variable b by its utility index $v = v(b)$, which is a monotonous transformation. With this transformation, constraints (13.24) and (13.26) are linear. Since the inverse function $b(v)$ is convex, convexity of the value functions follows immediately from standard dynamic programming arguments; see Stokey et al. (1989).

being strictly convex, we have $C_u'' > 0$ and consequently $V_u^t > V_u^{t+1}$. The sequence of utilities expected over the course of an episode of joblessness is thus strictly decreasing. The same holds true of the sequence V_e^t of expected utilities from starting a fresh job, since, the incentive constraint (13.26) being an equality, V_u^t and V_e^t necessarily vary in the same direction. Finally, from (13.22) we see that g_t varies like V_e^t, so the sequence of transfers flowing to wage earners is likewise strictly decreasing, but we do not know if g_t is a tax or a subsidy.

These results show that the optimal profile of benefits ought to decrease with the duration of unemployment when individuals are consuming the whole of their current income. Shimer and Werning (2008) have shown, however, that the optimal profile is constant when agents can borrow freely and without limit or risk in order to smooth out their consumption, given preferences of the CARA type. Otherwise, when preferences are not CARA, the profile can be decreasing or increasing.

The Optimal Profile of Unemployment Insurance in Practice

The model utilized to this point shows that the amount of unemployment benefit ought to diminish as the jobless spell persists so as to manage the insurance system optimally while offering the unemployed a predetermined level of utility. Hopenhayn and Nicolini (1997) take the view, moreover, that job search effort is a continuous variable that can be assigned any positive value, a hypothesis that adds considerable complication to the analytic results, yet without changing their qualitative prediction. Hopenhayn and Nicolini have also calibrated their model by taking as their benchmark the unemployment insurance system in place in the United States over the period 1978–1983. In this system, the replacement rate is 66% and benefits are paid for a maximum period of 26 weeks. In their basic calibration, Hopenhayn and Nicolini posit a utility function $v(c) = c^{1-\sigma}/(1-\sigma)$, with $\sigma = 1/2$. They assume that the exit rate from unemployment depends on job search effort a, according to the formula $p(a) = 1 - e^{-\rho a}$, where ρ is selected in such a way as to reproduce the estimated unemployment benefit elasticity of the probability of exiting from unemployment. It thus becomes possible to calculate the value $\bar{V}$ promised at the moment of entering unemployment. Hopenhayn and Nicolini then compare two unemployment insurance systems that offer the same discounted expected utility $\bar{V}$. In the first system, which approximates reality more closely, the agency cannot make transfers (this hypothesis amounts to positing $g = 0$ at every date in the theoretical model). The second system reproduces the optimal solution of the theoretical model, in which the agency is able to give subsidies to or levy taxes on those who find a job.

Table 13.3 presents the results obtained by Hopenhayn and Nicolini. The last column of this table shows that unemployment benefits tail off sharply as unemployment persists when the insuring body cannot tax, or subsidize, wages. Conversely, if transfers to those who become employed are allowed, the rate at which benefit payments tail off becomes very weak, and the replacement rate very high: 94% after a spell of unemployment lasting 52 weeks (at that time horizon, the probability of being unemployed is close to zero according to the calibrations used in this model). The third column of table 13.3 also reveals that the transfers are subsidies when the unemployment spell does not exceed 6 weeks (the taxes appearing in this column are negative) and that they become deductions after 6 weeks of joblessness. Hopenhayn and Nicolini stress further that the

TABLE 13.3
The optimal profile of the replacement rate in the presence of moral hazard.

	System with tax on wages		System without tax
Weeks of unemployment	Replacement rate (%)	Tax on wages (%)	Replacement rate without tax on wages (%)
1	99.0	−0.5	85.8
2	98.9	−0.4	80.8
3	98.8	−0.3	76.3
4	98.7	−0.2	72.1
5	98.6	−0.1	68.2
6	98.5	0.0	64.7
7	98.4	0.1	61.4
8	98.3	0.2	58.4
12	97.9	0.6	48.2
16	97.5	1.0	40.5
26	96.5	2.0	27.7
52	94.0	4.5	13.4

Source: Hopenhayn and Nicolini (1997, p. 426).

optimization of the unemployment insurance system would make it possible to reduce overall costs substantially, compared to the system in place. According to their estimates, for the same promised expected utility at entry into unemployment, costs are reduced by 7% when transfers to wage earners are not authorized, and 28% when they are.

The contribution of Hopenhayn and Nicolini (1997) underlines the potential importance of the ways unemployment insurance systems are structured when moral hazard is present. Wang and Williamson (1996) have extended their model by assuming that the probability of job loss depends on the effort made on the job by employees. In this hypothesis, moral hazard extends not just to the search efforts of the unemployed but also to the assiduity at work of those who are employed, for they may be tempted to shirk in order to lose their jobs if unemployment insurance benefits are too high. It is therefore desirable to adopt an experience rating scheme in which wages can be taxed and where income received depends on the duration not just of spells of joblessness but also of spells of work.

The Profile of Unemployment Benefits and Wage Setting

It should be noted that all these results were obtained within a partial equilibrium framework in which the impact of unemployment insurance on wage setting is ignored. We have seen, especially in chapter 7 in relation to wage bargaining, that the income of the jobless exerts upward pressure on wages when employees and firms are engaged in wage bargaining. From this perspective, shortening the period over which benefits are paid reduces the discounted expected utility of the jobless and exerts downward pressure on the wage being bargained over, and this in turn reinforces the incentive effect of regressive unemployment benefit on search effort. The same thing does not necessarily apply, however, if we look at the effect of a different profile of benefit payments,

with a *given budget* or *given tax rate*, which consists of paying more to the short-term unemployed and less to the long-term unemployed. Such a change of profile leads to an increase in the discounted expected gains of the short-term unemployed who have just lost their jobs at the expense of the long-term unemployed. For the same discount rate, intertemporal utility at the onset of a spell of unemployment rises, which increases the bargaining power of employees and thus promotes a rise in wages. Regressive benefits thus exert upward pressure on the rate of unemployment. For a given budget, the total effect of regressive benefit on unemployment is thus ambiguous: regressivity exerts an upward pressure on wages fixed through bargaining, which is unfavorable to employment, but it also intensifies the search effort of job seekers and lowers their reservation wage, which conversely promotes employment (Cahuc and Lehmann, 2000; Fredriksson and Holmlund, 2001). Changes in the rules regarding unemployment insurance thus have important impacts, and it is apparent that stricter rules may in certain cases have unfavorable effects in terms of employment. The reality is that the impact of the profile of benefit payments depends on the relative importance of the two effects just mentioned. When calibrated equilibrium models with an endogenous search effort, analogous to the matching model presented in chapter 9, are run, they suggest that rules providing for a rapid tailing-off of unemployment insurance benefits produce a positive but small effect on employment (Cahuc and Lehmann, 2000; Fredriksson and Holmlund, 2001).

1.4.2 Unemployment Insurance and the Business Cycle

Following the Great Recession, the question of unemployment benefit over the course of the cycle became the focus of much analysis. Should these payments be procyclical, contracyclical, or acyclical? In practice, and with reference to the formulas of the Baily, Chetty, and Shimer-Werning type reviewed previously, that comes down to determining how the elasticity of unemployment duration with respect to the amount of benefit varies with the economic trend, in other words with the unemployment rate.

Kroft and Notowidigdo (2011) carry out such an exercise on the basis of data on the unemployment rates in different states of the United States between 1985 and 2000. They estimate a hazard model where the effect of the amount of unemployment benefit on unemployment durations depends on the state unemployment rate (see chapter 5 for the estimation of duration models). They find that the elasticity of unemployment duration with respect to the level of unemployment benefit is 0.563 at the average state unemployment rate. But they also show that this elasticity varies widely with local labor market conditions, or more precisely that duration elasticity proves to be weaker when the local unemployment rate is high. The effect is of considerable magnitude, for Kroft and Notowidigdo estimate that a one-standard-deviation increase in the unemployment rate (an increase of 1.3 percentage points from a base of 6.2%) reduces the magnitude of the duration elasticity from 0.563 to 0.304 (a decline in magnitude of 46%). On the other hand, they do not find that the consumption term, $(c_e - c_u)/c_e$, drops in the Baily formula (13.6). In sum, Kroft and Notowidigdo find that the moral hazard cost of unemployment benefit is procyclical while the consumption-smoothing term is acyclical. In light of this empirical analysis, we ought to conclude that optimal unemployment benefit should be contracyclical: it should increase when the unemployment rate increases. Kroft and Notowidigdo estimate that a one-standard-deviation (1.3 percentage point)

increase in the local unemployment rate leads to a roughly 14 to 27 percentage point increase in the optimal replacement rate, depending on the value of the coefficient of relative risk aversion. The policy decision in the United States to lengthen the duration of unemployment benefit during the Great Recession was thus "theoretically well-grounded" according to Kroft and Notowidigdo.

Landais (2013) approaches the question using administrative data from the Continuous Wage and Benefit History Project (CWBH) about unemployment spells in five U.S. states from the late 1970s to 1984. He takes advantage of the wide variation in labor market conditions across states and over time in the CWBH data to investigate how estimates of the elasticity of unemployment duration with respect to the unemployment benefit vary with indicators of state labor market conditions. The results suggest that increases in the state unemployment rate are associated with a slight decrease in this estimated elasticity. According to the specification adopted by Landais, the estimated elasticity varies between 0.38 when the state unemployment rate is at 4.5% (the minimum in the CWBH data) and 0.25 when the unemployment rate is at 11.8% (the maximum in the CWBH data). Landais (2013) thus concludes that the labor supply response to unemployment benefits is (weakly) procyclical. So this result is in line with that of Kroft and Notowidigdo (2011), although the cyclicality of the estimates is somewhat larger for the latter. These converging results reinforce the view that optimal unemployment benefit ought to be contracyclical (see also Landais et al., 2010).

The analyses of Kroft and Notowidigdo (2011) and Landais (2013) are analyses in partial equilibrium, inasmuch as they leave out of account the reaction of wages to variations in the amount of unemployment benefit. Jung and Kuester (2011) have integrated this dimension into a search and matching model with risk-averse workers, endogenous hiring and separation, and unobservable search effort. They show that the social optimum may be decentralized through a production tax, a vacancy subsidy, a layoff tax, or unemployment benefits. Using a calibration targeted to the U.S. economy, Jung and Kuester conclude that hiring subsidies, layoff taxes, and the replacement rate at which unemployment insurance is paid should all rise in recessions. In an analogous model, but in which the amount of unemployment benefit is taken as the sole variable of economic policy, Mitman and Rabinovich (2011) arrive at a more nuanced conclusion. They find that the response of benefits to a negative shock should be nonmonotonic: unemployment benefit should be raised in the short term (4–6 weeks after the shock) in order to provide short-run relief to the unemployed and stabilize wages, but subsequently it should be brought back down to below its prerecession level in order to speed up the subsequent recovery. Their conclusions rest on the hypothesis of rapid wage adjustment.

2 EMPLOYMENT PROTECTION

Employment protection legislation is a set of mandatory restrictions governing the dismissal of employees. Their stated purpose is to increase the stability of employment. Despite that, there is intense debate about their actual effects, which also influence the level of employment and labor productivity. Firing costs do indeed reduce job destruction, but they also exert a negative effect on job creation; so the effect on employment

is ambiguous. Furthermore, firing costs may increase the stability of the jobs directly shielded by these costs, but they can also heighten the instability of the unshielded ones, like temporary work for example. By reducing job destruction, firing costs can be detrimental to labor productivity because they may block the reallocation of jobs toward more productive activities.

This section starts with a presentation of employment protection legislation in the OECD countries. We will then illustrate the theoretical attempts that have been made to pinpoint the effects of employment protection measures in a dynamic setting. These analyses do indeed suggest that employment protection has large-scale effects on workers and job flows; but whether these effects push unemployment up or down remains ambiguous. It depends especially on the wage-setting process. Finally, we will see that empirical studies focused on the wage-setting process tend to confirm the conclusions flowing from theoretical analysis.

2.1 What Is Employment Protection?

Measures to protect employment include severance payments, administrative firing taxes, advance notice of dismissal, administrative authorization, and prior negotiation with trade unions. The movement of individuals between labor contracts having shorter or longer time horizons (for example, in many European countries, the shift from a temporary job to "permanent" or "regular," that is, open-ended employment protected by regulation) is also covered by employment protection.

2.1.1 The OECD Index of Global Employment Protection

The OECD has constructed employment protection indicators based on 21 items quantifying the costs and procedures involved in dismissing individuals or groups of workers or in hiring workers on fixed-term or temporary work agency contracts. The latest version of this index is based not only on the relevant legislation but also on collective agreements and case law. The overall summary indicator takes values from 0 to 6; a higher value indicates a more stringent regulation. It is made up of three subindicators quantifying different aspects of employment protection:

- Individual dismissal of workers with regular contracts. This subindicator incorporates three aspects of dismissal protection: (i) procedural inconveniences that employers face when starting the dismissal process, such as notification and consultation requirements; (ii) notice periods and severance pay, which typically vary by tenure of the employee; and (iii) difficulty of dismissal, as determined by the circumstances in which it is possible to dismiss workers, as well as the repercussions for the employer if a dismissal is found to be unfair (such as compensation and reinstatement).
- Additional costs for collective dismissals. Into this category fall extra delays, costs, or notification procedures that kick in when an employer dismisses a large number of workers at one time. It does not include regular costs for individual dismissal. Nor does it reflect the overall strictness of the regulation of collective dismissals, which is the sum of costs for individual dismissals and any additional cost of collective dismissals.

- Regulation of temporary contracts. This subindicator quantifies regulation of fixed-term and temporary work agency contracts with respect to the types of work for which these contracts are allowed and their duration. The group of measures falling into this category also include regulations governing the establishment and operation of temporary work agencies and requirements for agency workers to receive the same pay and/or conditions as equivalent workers in the user firm, which can increase the cost of using temporary agency workers relative to that of hiring workers on permanent contracts.

2.1.2 The Regulation of Permanent Jobs

Figure 13.7 shows the stringency of employment protection for permanent workers, that is, those working on regular contracts, who amounted to 88% of dependent employment in 2012 in the OECD. The index refers to the regulation as in force on 1 January 2013.

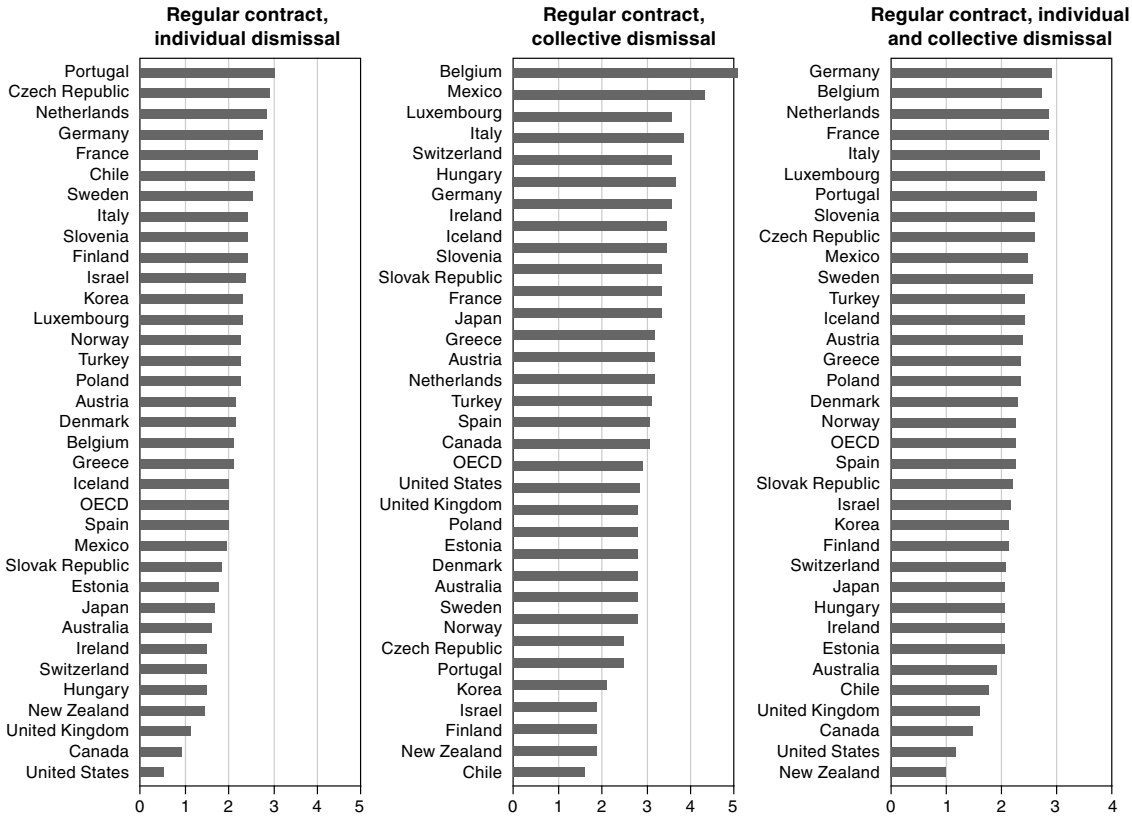


FIGURE 13.7
Protection of permanent workers on regular contract against individual and collective dismissal in 2013.

Note: The indicator goes from 0 for the weakest regulation to 6 for the strongest. For collective dismissal, the index correponds to additional regulations on top of those already pertaining to individual layoffs. OECD refers to the nonweighted average of rates for the OECD countries.

Source: OECD Employment Protection database.

The first panel of figure 13.7 shows that protection of permanent workers against individual dismissal is weak in the United States, the United Kingdom, Canada, and Japan, but it is rather stringent in continental Europe (Portugal, the Netherlands, Germany, and France). The second panel of figure 13.7 shows that the landscape is similar when we look at the hurdles facing collective dismissals (which include additional regulation on top of that pertaining to individual layoffs). Collective layoffs are more strictly regulated in continental Europe and much less so in the Anglophone and Nordic countries. In Denmark, for instance, most of the regulation of collective dismissals arises out of collective agreements. Overall, the picture we see is that the protection of permanent jobs by erecting barriers to individual *and* collective dismissal is strongest in Germany, Belgium, the Netherlands, and France, but it is weak in Anglophone countries like the United States, Canada, and the United Kingdom (third panel of figure 13.7). Of course, one of the challenges of this index is to properly take into account not only the legal text but the outcomes of the individual cases litigated under it, for in practice regulations may be applied more or less strictly. This is especially the case when, in a litigious context, judges have a lot of leeway in making their assessments of the situation. For instance, in many OECD countries a layoff for economic reasons is defined as a layoff for reasons independent of the person displaced. But in a few countries (e.g., France and Spain), the concept of economic layoff is a little different: an economic layoff is only legal if a firm is facing major economic difficulties, the situation being evaluated by a judge. This often renders the outcome of the layoff procedure uncertain for employers (see OECD, 2013, for more details on the content of regulation).

Figure 13.8 represents the overall index of protection for regular employment over the period 1985–2013. Most of the Anglophone countries have kept their regulation broadly stable over the last 30 years (except Australia, where regulation did expand, though still settling at low levels). In Southern Europe (Portugal, Spain, and Italy) there has been some deregulation, notably following the 2009 crisis. In the Nordic countries, some deregulation is observable in the 1990s (except in Norway).

2.1.3 The Regulation of Temporary Jobs

When we look at the cross-country stringency of regulation of temporary jobs, which comprises the regulation of both temporary work agency employment and fixed-term contracts, the picture is almost the same as for permanent jobs. As shown by figure 13.9, the Anglophone countries have weak regulation of temporary forms of employment (but so does Sweden), whereas Spain, Greece, and France have rather stringent regulation of such contracts. But the regulation of temporary contracts is particularly difficult to enforce, due to the very large number of contracts entered into and their short duration. Typically, in countries with rigid regulations on permanent contracts, the hiring of temporary and fixed-term workers contributes greatly to overall worker flows, even when such hiring is itself strictly regulated. For instance, in France 85% of hires were into temporary jobs in 2012. In Spain the share of temporary jobs was close to 25% of total employment in the same year (compared with 12% in the OECD). These figures are much higher among youth. One quarter of employees between 15 and 24 years old is on a temporary contract in the OECD area (over 60% in Spain).

In light of these elements of international comparison, we see that the Anglophone countries form a fairly homogeneous domain where weak employment protection

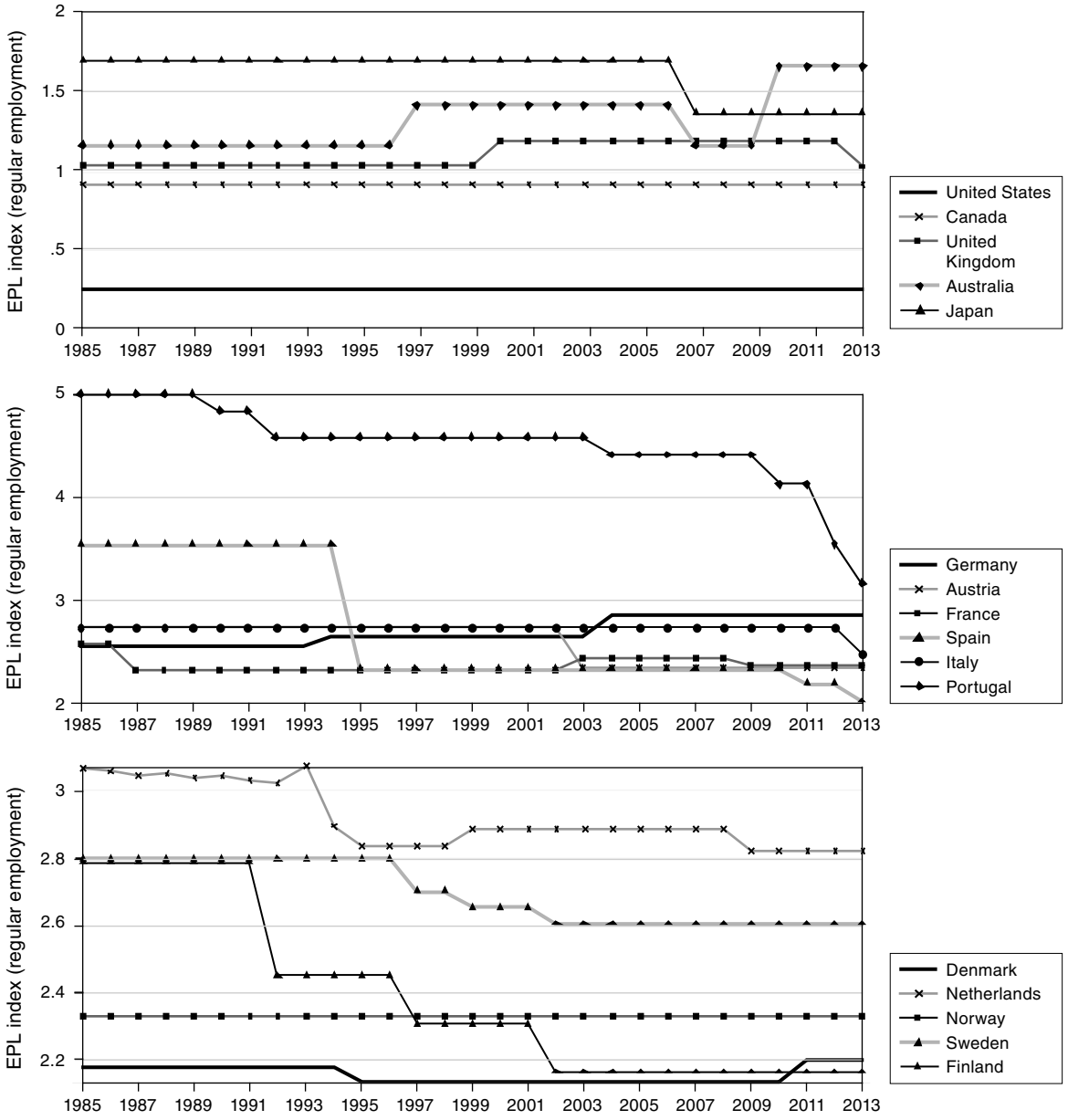


Figure 13.8

Protection of permanent workers on regular contracts against individual and collective dismissal in selected countries, 1985–2013.

Note: EPL = Employment Protection Legislation. The indicator goes from 0 for the weakest regulation to 6 for the strongest. For collective dismissal, the index correponds to additional regulations on top of those pertaining to individual layoffs.

Source: OECD Employment Protection database.

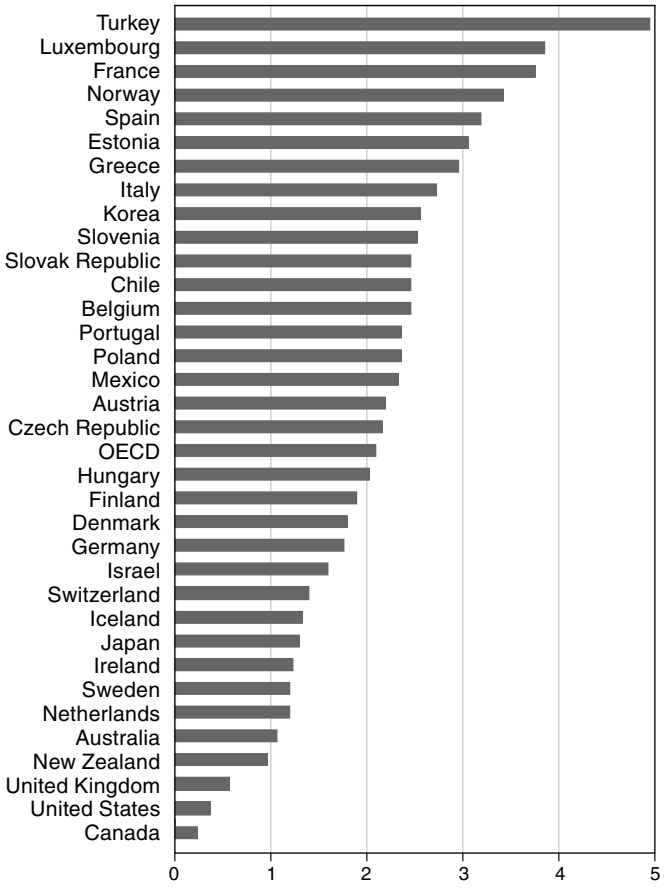


Figure 13.9

Regulation of temporary contracts in 2013.

Note: The indicator goes from 0 for the weakest regulation to 6 for the strongest. OECD refers to the nonweighted average of the indicator for the OECD countries.

Source: OECD Employment Protection database.

prevails. The countries of Southern Europe, including France, come close to forming an opposite pole, with strong employment protection. The countries of Nordic Europe show diversity, with Denmark for example being rather closer to the Anglophone model, whereas Sweden's ranking can shoot from low to high depending on which aspect of employment protection is measured.

In what follows, we analyze the impact of employment protection with the help of a matching model close to the one presented in chapter 9. We start with a model where the wage is exogenous, making it easier to grasp the main effects of employment protection. We then make the assumption that wages are bargained over with the goal of heightening the degree of employment protection; the effects of this might induce alterations in the level of wages.

2.2 The Effects of Employment Protection

In the versions of the matching model that we have used to this point, the exit rate from employment was most often considered an exogenous parameter—a hypothesis clearly ill suited to studying the effects of employment protection, which are intended to make the destruction of jobs and the firing of employees less frequent. It is therefore necessary to make decisions to destroy jobs endogenous. We can achieve that by adopting a model analogous to the one of Mortensen and Pissarides (1994, 1999), and within that setting we start by assuming that wages are exogenous. This assumption allows us to present decisions to destroy jobs and the impact of employment protection on unemployment and labor market flows in a very simple fashion. It also has the advantage of showing how the labor market functions in the presence of a minimum wage. Then we proceed to analyze the impact of job protection when wages are bargained over. We will see that the response of wages can counteract the impact of job protection on job creation, job destruction, and unemployment.

2.2.1 The Matching Model with Endogenous Job Destruction and Exogenous Wage

In what follows the firing of an employee occurs in the wake of a negative productivity shock of such magnitude that it costs the firm more to keep him on than it does to fire him. The basic matching model as formalized in chapter 9 will have to be modified somewhat to represent this scenario.

The Threshold of Job Destruction

We assume that the production of an individual, which has hitherto been a constant parameter denoted by y, is now a random variable ε with support[7] $(-\infty, \varepsilon_u]$. The cumulative distribution function of this random variable is designated by $G(\cdot)$. Another important element of the analysis is the *degree of persistence* of shocks, that is, the length of the period during which individual productivity keeps the same value. To grasp this notion, we assume that this productivity varies according to a Poisson process with parameter λ. Let us recall that this means that productivity changes with a probability $\lambda \mathrm{d}t$ over every small interval of time $\mathrm{d}t$. When a shock supervenes, the new value of productivity is found by a random draw from the distribution $G(.)$. Finally, individual productivities are independent of one another. Shocks are thus *idiosyncratic*: they affect every job independently.[8]

Employment protection gives rise to costs of two kinds: severance payments, which are transfers from the employer to the employee; and administrative costs to the firm with no transfer to the employee. It is worth noting that some rules include both kinds of costs. For instance the advance notice of dismissal and the obligation to try to find another position are both administrative costs and transfers to the employee. We will see below that when wages are endogenous, it is useful to distinguish these two

[7] The fact that the support has the upper limit ε_u is not essential. We follow the presentation of Mortensen and Pissarides (1994) here, which makes the exposition somewhat easier.

[8] For more on random variables and Poisson processes, see mathematical appendices C and D, respectively, at the end of the book.

kinds of cost, for they affect labor market equilibrium differently. Conversely, severance payments and administrative costs actually have exactly the same impact on employment when wages are exogenous. That is why the strictness of employment protection is identified by a single parameter, denoted f, which represents all the costs to the firm of firing an employee. It is thus a global measure of the rigor of employment protection, analogous to the synthetic OECD index by which countries are ranked in figure 13.7.

Let w be the wage. When current productivity takes the value ε, the expected profit $\Pi_e(\varepsilon)$ from a filled job at stationary equilibrium is written:

$$r\Pi_e(\varepsilon) = \varepsilon - w + \lambda\,[\Pi_\lambda - \Pi_e(\varepsilon)] \tag{13.31}$$

In this equality Π_λ designates the expected profit when a productivity change occurs; we give its exact expression below. Equation (13.31) is interpreted the same way as all the equations defining expected profits and utilities encountered thus far. For a given level ε of current productivity, the instantaneous profit is equal to $(\varepsilon - w)$, and the term $\lambda\,[\Pi_\lambda - \Pi_e(\varepsilon)]$ corresponds to the average gain linked to a possible change of state of the job. The only change of state envisaged here is a change in the level of individual productivity. This event comes about with probability $\lambda \mathrm{d}t$ over every small interval of time $\mathrm{d}t$.

When the employer fires a worker, she incurs fixed costs amounting to f and is left with a vacant job offering an expected profit equal to Π_v. In total, the expected profit flowing from the separation of an employee amounts to $-f + \Pi_v$. In consequence, the employer fires the employee when the discounted profit $\Pi_e(\varepsilon)$ from a filled job falls below the gain she gets by the firing. This situation comes about when the inequality $\Pi_e(\varepsilon) < -f + \Pi_v$ is satisfied. Now, relation (13.31) shows that profit $\Pi_e(\varepsilon)$ increases with individual productivity ε. In these conditions, the employer will fire the employee if $\varepsilon \leq \varepsilon_d$, where the *reservation productivity* ε_d is defined by the equality $\Pi_e(\varepsilon_d) = -f + \Pi_v$. Using equation (13.31), we immediately find that when the free entry condition $\Pi_v = 0$ is satisfied, the reservation productivity is given by:

$$\varepsilon_d = w - (r + \lambda)f - \lambda\Pi_\lambda \tag{13.32}$$

The Job Destruction Rate

In relation (13.32), Π_λ is endogenous. This variable must be known in order to describe labor market equilibrium completely. For that purpose, it will be helpful to note at the outset that the definition (13.31) of expected profit from a filled job entails $(r + \lambda)[\Pi_e(\varepsilon) - \Pi_e(\varepsilon_d)] = \varepsilon - \varepsilon_d$. Now, when the free entry condition $\Pi_v = 0$ is satisfied, we have $\Pi_e(\varepsilon_d) = -f$, and the expression of the expected profit from a filled job takes the following form:

$$\Pi_e(\varepsilon) = \frac{\varepsilon - \varepsilon_d}{r + \lambda} - f \tag{13.33}$$

When a shock alters productivity, two eventualities may ensue: if the new value of productivity is below the threshold ε_d, the employee is fired and the employer assumes the costs f arising from this firing; conversely, if productivity takes a new value ε above the threshold ε_d, the employer keeps the worker on, and her expected profit amounts to

$\Pi_e(\varepsilon)$. Using relation (13.33), the average profit Π_λ in the wake of a productivity shock is written thus:

$$\Pi_\lambda = \int_{-\infty}^{\varepsilon_d} -f \mathrm{d}G(\varepsilon) + \int_{\varepsilon_d}^{\varepsilon_u} \Pi_e(\varepsilon)\mathrm{d}G(\varepsilon) = -f + \frac{1}{r+\lambda}\int_{\varepsilon_d}^{\varepsilon_u}(\varepsilon - \varepsilon_d)\mathrm{d}G(\varepsilon) \tag{13.34}$$

If we bring this expression of Π_λ into definition (13.32) of the threshold value ε_d, it becomes:

$$\varepsilon_d = w - rf - \frac{\lambda}{r+\lambda}\int_{\varepsilon_d}^{\varepsilon_u}(\varepsilon - \varepsilon_d)\mathrm{d}G(\varepsilon) \tag{13.35}$$

This equation defines ε_d as a function of the parameters of the model. It shows that the reservation productivity ε_d is *inferior* to the wage w. In other words, for values of productivity lying close to the destruction threshold ε_d, the employer may suffer a loss in the current period. If she does not fire the employee when $\varepsilon < w$, it is because, for one thing, she must immediately pay costs f, and for another, she expects to be able, in the future, to make up for this loss through positive profits deriving from higher productivity. This possibility of future gain is represented by the term $\lambda\Pi_\lambda$ in equation (13.32), the equivalent of an "option value" of a filled job. The inequality $\varepsilon_d < w$ conveys the phenomenon of *labor hoarding*: the costs of firing give the firm an incentive to keep its workers during downturns because it anticipates future profits when the cycle turns back up.

The job destruction rate, which we will again denote by q, is easy to find if the value of the reservation productivity ε_d is known. For a job to be destroyed, the value of current productivity has to change—which happens at rate λ—and the new value of productivity has to lie below ε_d—which comes about with probability $G(\varepsilon_d)$. Hence, at every date, a filled job is destroyed at rate $\lambda G(\varepsilon_d)$. Therefore, if there is a large number of firms, the job destruction rate amounts to $q = \lambda G(\varepsilon_d)$. Differentiating equation (13.35) defining ε_d with respect to f and λ, we easily arrive at:

$$\frac{\partial \varepsilon_d}{\partial f} < 0\,, \quad \frac{\partial q}{\partial f} < 0 \quad \text{and} \quad \frac{\partial \varepsilon_d}{\partial \lambda} < 0$$

Hence an increase in firing costs lowers the reservation productivity ε_d and consequently lowers the rate of job destruction. This result is highly intuitive and corresponds to the stated goal of imposing firing costs, which is precisely to increase the rate of labor hoarding when unfavorable shocks occur. We see as well that a reduction in the degree of persistence of shocks (i.e., an increase in λ) will also tend to increase labor hoarding, so the effect on the job destruction rate is ambiguous. This result also implies that firing costs reduce labor productivity since they induce firms to keep workers at lower productivity levels. Lower labor productivity is the counterpart of labor hoarding.

The Impact of Firing Costs on Labor Market Equilibrium

To complete our description of the equilibrium at which the labor market arrives, we still have to specify the value of the labor market tightness θ which occurs in the

expression $\theta m(\theta)$ of the exit rate from unemployment. To accomplish that, we assume that the lifespan of a filled job always starts at the maximum value ε_u of productivity. This hypothesis is not at all essential in this context; it is introduced for the sake of simplicity and is justified when we introduce productivity growth (see chapter 10). It serves to convey the idea that newly created jobs most often have the benefit of the latest technological innovations and are thus the most productive. If h denotes, as it did above, the costs arising from the search for a worker, then the value of a vacant job is written:

$$r\Pi_v = -h + m(\theta)\left[\Pi_e(\varepsilon_u) - \Pi_v\right]$$

When the free entry condition $\Pi_v = 0$ is satisfied, this last relation entails $\Pi_e(\varepsilon_u) = h/m(\theta)$. We come back to the result that at free entry equilibrium, the average cost $h/m(\theta)$ of a vacant job is equal to the expected profit $\Pi_e(\varepsilon_u)$ of a job that has just been filled. Setting $\varepsilon = \varepsilon_u$ in (13.33) we get the expression of $\Pi_e(\varepsilon_u)$ as a function of ε_d, and if we make this expression equal to the average cost of a vacant job, we arrive at:

$$\frac{h}{m(\theta)} = \frac{\varepsilon_u - \varepsilon_d - (r+\lambda)f}{r+\lambda} \tag{13.36}$$

Knowing ε_d given by (13.35), this equation completely defines the labor market tightness θ. It is analogous to the "labor demand" equations that we obtained from different versions of the matching model when we assumed that the job destruction rate was an exogenous parameter. With the help of relation (13.35) giving the equilibrium value of the threshold ε_d, it is easy to verify that the expected profit $\Pi_e(\varepsilon_u)$ from a new job—which corresponds to the right-hand side of equality (13.36)—is reduced when firing costs increase. Firms then open up fewer vacant jobs (or, if one prefers, the period $1/m(\theta)$ during which a job remains vacant diminishes), and the labor market tightness θ and the exit rate from unemployment $\theta m(\theta)$ fall off. In sum, after several calculations, we arrive at the following results:

$$\frac{\partial\theta}{\partial\lambda} < 0, \quad \frac{\partial\theta}{\partial f} < 0 \quad \text{and} \quad \frac{\partial\theta m(\theta)}{\partial\lambda} < 0, \quad \frac{\partial\theta m(\theta)}{\partial f} < 0$$

Given that the job destruction rate q is here equal to $\lambda G(\varepsilon_d)$, relation (9.22) from chapter 9 giving the expression of the stationary unemployment rate u is now written:

$$u = \frac{q+n}{\theta m(\theta)+q+n} = \frac{\lambda G(\varepsilon_d)+n}{\theta m(\theta)+\lambda G(\varepsilon_d)+n} \tag{13.37}$$

Firing costs f thus have an *ambiguous* impact on the unemployment rate, since they combine two effects that work against each other. First, they favor labor hoarding and so reduce the job destruction rate, but at the same time they reduce job creation (the exit rate from unemployment falls) because higher firing costs have the effect of degrading the profit outlook of every new hire. From the standpoint of labor market equilibrium, these results confirm the ones already reached in chapter 2, where adjustment costs were introduced into models of labor demand. It is interesting to note that the

degree to which shocks persist conditions the impact of firing costs on job destruction and so on unemployment (see Cabrales and Hopenhayn, 1998). By way of example, let us imagine that after a shock, productivity falls irreversibly to zero. In that circumstance, the job destruction rate is necessarily equal to λ, so it is independent of firing costs. The upshot is that firing costs have the effect of decreasing labor market tightness without altering the job destruction rate, which entails a positive impact on unemployment.

All these results were obtained on the assumption that the wage was exogenous. But it is intuitive that wages are influenced by the rules in place regarding employment protection and will thus in turn affect labor market equilibrium.

2.2.2 Employment Protection and Wage Bargaining

The model just developed well illustrates the functioning of a labor market in the presence of a compulsory minimum wage. But if wages are open to bargaining, firing costs affect the level of compensation, and so, indirectly, employment. Thus, when wages are bargained over, it is easy to show that severance payments (i.e., transfers from employer to employee) have no impact on the exit rate from unemployment and the job destruction rate, for they simply make themselves felt in the form of a reduction in wages. Likewise, it will be evident that a portion of the administrative costs are in fact borne by the workers at the time of hiring, which has the effect of limiting their impact on job creation. In order to take these possibilities into account, we explicitly distinguish two components of firing costs by setting $f = f_a + f_e$. Parameter f_a designates the costs arising from various administrative hurdles (advance notice, prior obligations, possible legal proceedings, etc.), whereas parameter f_e represents an effective transfer from the firm to the employee. The two parameters f_a and f_e are here always taken to be exogenous (in the framework of the matching model, Pissarides (2001) endogenizes severance payments f_e by assuming that employees are risk averse and so wish to be insured against fluctuations in their future income). We will see that calibration exercises carried out on the model confirm the importance of the reaction of wages to employment protection. They suggest that firing costs may be favorable to employment when wages are flexible but that they may destroy a significant volume of jobs in the presence of a minimum wage.

Bargaining in the Presence of Firing Costs

We return to the previous model, but now we assume that wages are bargained over at the time of hiring and again every time a shock affects productivity. The existence of firing costs requires that we distinguish between wage bargaining at the start of the job, when these costs are still virtual, no contract having yet been signed, and wage renegotiations, which lead to firing costs if they fail.

We must also distinguish between the expected profit Π_0 from a new job, and the expected profit $\Pi_e(\varepsilon)$ from a filled job with current productivity ε. We thus have:

$$r\Pi_0 = \varepsilon_u - w_0 + \lambda\,(\Pi_\lambda - \Pi_0) \tag{13.38}$$

$$r\Pi_e(\varepsilon) = \varepsilon - w(\varepsilon) + \lambda\,[\Pi_\lambda - \Pi_e(\varepsilon)] \tag{13.39}$$

In these relations, w_0 and $w(\varepsilon)$ designate respectively the wage negotiated at hiring and the wage renegotiated when productivity takes the value ε. The term Π_λ is always

defined by equation (13.34). In similar fashion, the expected utility V_0 of a worker who has just been hired, and the expected utility $V_e(\varepsilon)$ of a worker who holds a job with current productivity ε, are defined by the formulas:

$$rV_0 = w_0 + \lambda\,(V_\lambda - V_0) \tag{13.40}$$

$$rV_e(\varepsilon) = w(\varepsilon) + \lambda\,[V_\lambda - V_e(\varepsilon)] \tag{13.41}$$

The term V_λ designates the expected utility of a worker when her job is affected by a productivity shock. With the reservation productivity (which, as we demonstrate below, is unique) again denoted by ε_d, this expected gain has the expression:

$$V_\lambda = \int_{-\infty}^{\varepsilon_d} (f_e + V_u)\mathrm{d}G(\varepsilon) + \int_{\varepsilon_d}^{\varepsilon_u} V_e(\varepsilon)\mathrm{d}G(\varepsilon) \tag{13.42}$$

where V_u is the expected utility of an unemployed person, defined by:

$$rV_u = z + \theta m(\theta)(V_0 - V_u) \tag{13.43}$$

These equations allow us to define the surplus S_0 of a new job and the surplus $S(\varepsilon)$ of a continuing job already hit by a shock with current productivity ε. It comes to:

$$S_0 = \Pi_0 - \Pi_v + V_0 - V_u, \quad S(\varepsilon) = \Pi_e(\varepsilon) - (\Pi_v - f_a) + V_e(\varepsilon) - V_u \tag{13.44}$$

These definitions are easily grasped. At the time of hiring, breaking off the bargaining entails neither the payment of a severance nor administrative costs. But during renegotiation, the various costs and transfers take effect if the bargaining fails, and the fallback profit of the firm amounts to $(\Pi_v - f_a - f_e)$, while the fallback utility of the worker takes the value $(V_u + f_e)$ since it is she who benefits from transfer f_e. The upshot is that the severance payments f_e do not come into the definition of the surplus. Moreover, for the same productivity, the surplus of a continuing job is greater than the one released by a new job. Noting that equations (13.38), (13.39), (13.40), and (13.41) entail $\Pi_0 + V_0 = \Pi_e(\varepsilon_u) + V_e(\varepsilon_u)$, the definitions (13.44) of the surpluses entail:

$$S_0 = S(\varepsilon_u) - f_a \tag{13.45}$$

The Impact of Firing Costs on Wages

As in the basic model of chapter 9, we assume that bargaining leads to a surplus-sharing rule dependent on the bargaining power of each of the agents. Let γ again be the relative power of a worker; for a new job this rule is written:

$$V_0 - V_u = \gamma S_0, \qquad \Pi_0 - \Pi_v = (1-\gamma)S_0 \tag{13.46}$$

On the other hand, since renegotiation gives rise to a severance payment in case of failure to agree, the surplus-sharing rule determining the renegotiated wage takes the form:

$$V_e(\varepsilon) - (V_u + f_e) = \gamma S(\varepsilon), \qquad \Pi_e(\varepsilon) - (\Pi_v - f) = (1-\gamma)S(\varepsilon) \tag{13.47}$$

Assuming that the free entry condition $\Pi_v = 0$ is satisfied, this rule entails that jobs are destroyed when the value of the surplus $S(\varepsilon)$ becomes negative. We see that the employer and the worker have an interest in separating for the *same* values of productivity, since equations (13.44) and (13.47) entail:

$$S(\varepsilon) < 0 \Leftrightarrow \Pi_e(\varepsilon) < -f \Leftrightarrow V_e(\varepsilon) < V_u + f_e$$

In other words, jobs are destroyed by common consent when they release a negative surplus. This result comes from the fact that the firm and the worker are capable of finding a mutually advantageous contract, one preferable to separation, if and only if the surplus obtained by keeping the job going is positive. It can be shown that there exists a unique threshold value of productivity, beneath which jobs are destroyed. Using relations (13.39), (13.41), and (13.44), the surplus $S(\varepsilon)$ is written as follows:

$$S(\varepsilon) = \frac{\varepsilon + \lambda(V_\lambda + \Pi_\lambda)}{r + \lambda} - (V_u - f)$$

As V_λ and Π_λ are independent of current productivity ε, this expression of the surplus entails $S'(\varepsilon) = 1/(r + \lambda) > 0$. The surplus is thus an increasing function of productivity. Consequently there exists a single value of ε, denoted ε_d, such that $S(\varepsilon_d) = 0$, and below which jobs are destroyed. Using relations (13.34) and (13.42) defining Π_λ and V_λ, we arrive at:

$$(r + \lambda)S(\varepsilon) = \varepsilon - rV_u + rf_a + \lambda \int_{\varepsilon_d}^{\varepsilon_u} S(x)\mathrm{d}G(x) \qquad \textbf{(13.48)}$$

With sharing rule (13.47), definition (13.39) of profit, and equation (13.34), this definition of the surplus allows us to write the renegotiated wage in the following manner:

$$w(\varepsilon) = rV_u + \gamma(\varepsilon - rV_u) + r(f_e + \gamma f_a) \qquad \textbf{(13.49)}$$

As for the wage negotiated at hiring, obtained from (13.38), (13.45), (13.46), and (13.48), it takes the form:

$$w_0 = rV_u + \gamma(\varepsilon_u - rV_u) - \lambda(f_e + \gamma f_a) \qquad \textbf{(13.50)}$$

These expressions of the hiring wage and the renegotiated wage well illustrate the effects of firing costs at the partial equilibrium of a decentralized negotiation (i.e., for given V_u). The hiring wage diminishes with firing costs, since firms anticipate that they will have to endure them in the future. The renegotiated wage, however, rises with firing costs, since the latter enhance the gains of workers if they do separate from their employer.

Labor Market Equilibrium

The equilibrium values of the reservation productivity ε_d and the labor market tightness θ are found, as they were when the wage was exogenous, using a job creation equation and a job destruction equation. The expected profit Π_v from a vacant job satisfies:

$$r\Pi_v = -h + m(\theta)(\Pi_0 - \Pi_v)$$

When the free entry condition $\Pi_v = 0$ is satisfied, we find the usual equality between the expected profit Π_0 of a job newly filled and the average cost $h/m(\theta)$ of a vacant job. The sharing rule (13.46) thus entails $(1-\gamma)S_0 = h/m(\theta)$. On the other hand, the definition (13.48) of the surplus allows us to write the latter as a function of the threshold ε_d in the form $S(\varepsilon) = (\varepsilon - \varepsilon_d)/(r+\lambda)$. Utilizing (13.45), it comes to:

$$\frac{h}{m(\theta)} = (1-\gamma)\left[\frac{\varepsilon_u - \varepsilon_d}{r+\lambda} - f_a\right] \tag{13.51}$$

This job creation equation defines a decreasing relation between labor market tightness and the reservation productivity. We can account for this result by noting that the average lifespan of a job, $1/\lambda G(\varepsilon_d)$, decreases with the reservation productivity ε_d. Consequently, when the reservation productivity rises, expected profit falls, and firms open up fewer vacant jobs.

Since $\Pi_0 = h/m(\theta)$, the job destruction equation is found by first noting that the expected utility (13.43) of an unemployed person is written, using sharing rule (13.46):

$$rV_u = z + \theta m(\theta)\gamma S_0 = z + \frac{\gamma h \theta}{1-\gamma} \tag{13.52}$$

If we substitute this value of rV_u in (13.48), the job destruction condition, $S(\varepsilon_d) = 0$, finally yields:

$$\varepsilon_d = z + \frac{\theta\gamma h}{1-\gamma} - rf_a - \frac{\lambda}{r+\lambda}\int_{\varepsilon_d}^{\varepsilon_u}(\varepsilon - \varepsilon_d)\mathrm{d}G(\varepsilon) \tag{13.53}$$

The job destruction equation defines an increasing relation between labor market tightness and the reservation productivity, for high tightness corresponds to a strong exit rate from unemployment and thus to high expected gains on the part of unemployed persons. Since the surplus diminishes with the expected utility of unemployed persons, a high value of labor market tightness signifies a small surplus, and that entails a high job destruction rate.

The equilibrium values of labor market tightness θ and the reservation productivity ε_d are defined by the system of equations (13.51) and (13.53). These values are independent of the severance payment f_e, which thus has the sole effect of altering the wage profile. Administrative costs, on the other hand, act simultaneously on the equations of job creation and job destruction. The impact of an increase in administrative costs is represented in figure 13.10. The curve of job creation shifts downward because an increase in these costs exerts downward pressure on job creation, and that

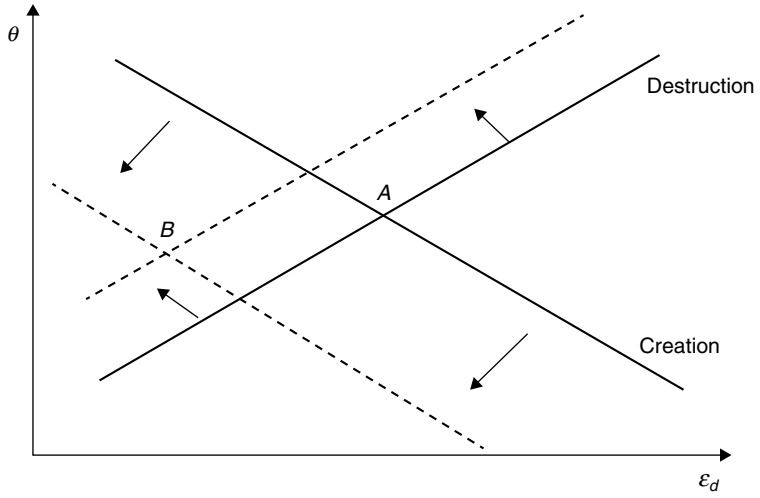


FIGURE 13.10
The impact of an increase in administrative firing costs.

has the effect of lowering the reservation productivity and labor market tightness. The job destruction curve shifts to the left because fewer jobs are destroyed when hiring costs are greater. Equilibrium thus moves from point A to point B. The threshold ε_d, and so the job destruction rate $\lambda G(\varepsilon_d)$, both decrease. The effect on labor market tightness is a priori ambiguous. It is possible to show, using equations (13.51) and (13.53), however, that labor market tightness falls with firing costs. The effect on the unemployment rate is thus indeterminate, since the new equilibrium is characterized by a lower exit rate from unemployment $\theta m(\theta)$ and a lower job destruction rate $\lambda G(\varepsilon_d)$.

2.2.3 The Importance of Wage Setting

Whether the wage is exogenous or negotiated, strengthened employment protection reduces manpower flows and has an ambiguous impact on unemployment. Negotiated wages, however, react to this strengthening. At equilibrium the hiring wage in particular falls. This result is established by substituting the expression (13.52) of rV_u in (13.50), which yields:

$$w_0 = (1-\gamma)z + \gamma(\theta h + \varepsilon_u - \lambda f_a) - \lambda f_e \qquad \textbf{(13.54)}$$

Since firing costs have a negative impact on labor market tightness θ, relation (13.54) shows that they also exert a downward pressure on hiring wages. The decline in the hiring wage thus makes it possible to lessen the negative effects of firing costs on profits, and thus on job creation. And on the contrary, a mandatory minimum wage, by preventing wages from declining, must amplify the impact of firing costs on job creation. The calibration exercises which follow confirm these intuitions.

Flexible Wage

The parameters used in these calibration exercises are described in table 13.4. The unit of time corresponds to one year. Annual production y has been normalized to one. The

TABLE 13.4
Parameter values of the model with endogenous job destruction.

γ	h	λ	r	z
0.5	0.3	0.15	0.05	0.46

matching process is represented by a Cobb-Douglas function, written $M(V, U) = V^{0.5}U^{0.5}$. The annual interest rate is set to 5%. We assume that γ is equal to the elasticity of the matching function with respect to the unemployment rate. We saw in chapter 9 that this hypothesis ensures the efficiency of the decentralized equilibrium. The cumulative distribution function $G(\cdot)$ is taken to be uniform over the interval $[0, 1]$ and the productivity shock is assumed to follow a Poisson process with parameter λ equal to 0.15 (for calibrations of the matching model using functional forms and similar numerical values, see Millard and Mortensen, 1997, and Mortensen and Pissarides, 1999). These values give plausible rates of job destruction, lying between 10% and 15% per annum. Finally, the values of h and z are chosen in such a way as to obtain unemployment rates and unemployment durations compatible with their average values in Europe.

Figure 13.11 presents the impact of an increase in administrative firing costs on the unemployment rate u, the exit rate from unemployment $\theta m(\theta)$, and the job destruction rate $\lambda G(\varepsilon_d)$ when wages are negotiated. We see that employment protection has little influence on the unemployment rate. An increase in firing costs by an amount equal to the average quarterly production of a worker provokes a fall in the unemployment rate of around 0.1 percentage point when wages are negotiated. It should also be noted that the exit rate from unemployment and the job destruction rate do not show much sensitivity either to this rise in administrative firing costs. It is important to emphasize that the negative relationship between firing costs and the unemployment rate is not robust to changes in the values of the parameters. The degree to which shocks persist does play an important part in this domain.

Figure 13.12 shows, on the other hand, that firing costs do exert a positive effect on unemployment if the gains of unemployed persons are relatively high. The extent of this effect is always very slight, however.

Other studies analyze the effects of firing costs by resorting to calibrated versions of matching models close to that of Mortensen and Pissarides (1994) on the constant

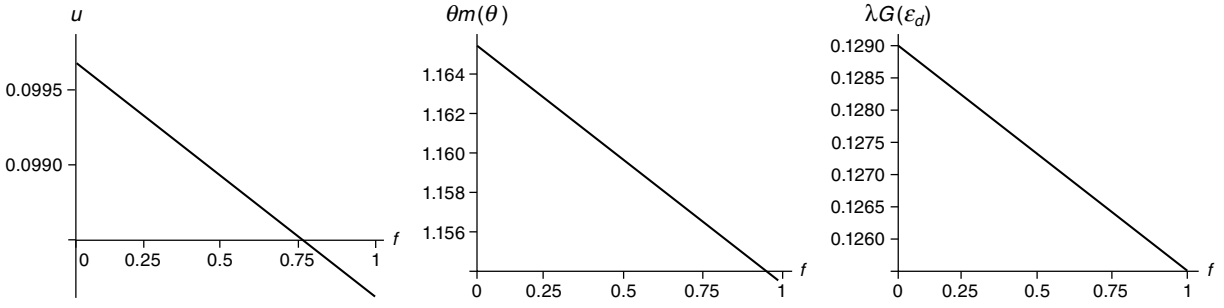


FIGURE 13.11
The impact of firing costs on the unemployment rate u, the exit rate from unemployment $\theta m(\theta)$; and the job destruction rate $\lambda G(\varepsilon_d)$ with negotiated wages and $z = 0.46$. f is expressed as a fraction of average quarterly production.

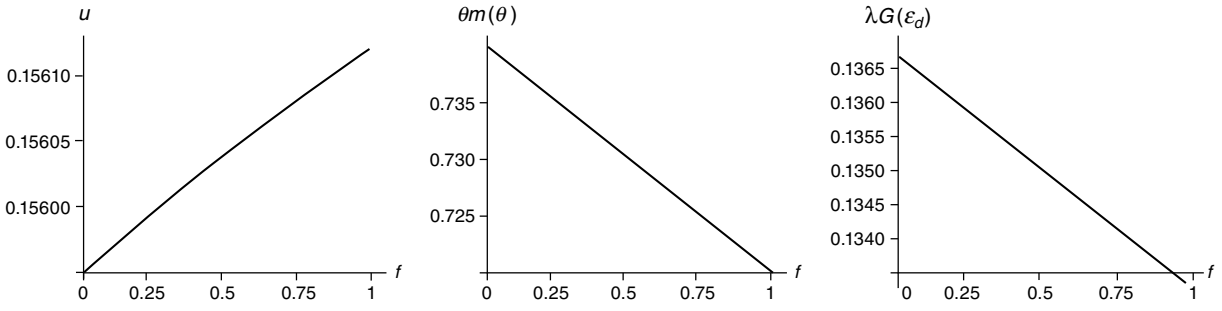


FIGURE 13.12
The impact of firing costs on the unemployment rate u, the exit rate from unemployment $\theta m(\theta)$, and the job destruction rate $\lambda G(\varepsilon_d)$ with negotiated wages and $z = 0.75$. f is expressed as a fraction of average quarterly production.

assumption that wages are flexible. Thus Mortensen and Pissarides (1999) find that the rise in firing costs reduces both labor market flows and the unemployment rate. Garibaldi (1998) arrives at an analogous result when the values taken by firing costs are not too high. On Spanish data, Cabrales and Hopenhayn (1997) estimate that higher firing costs are responsible for the reduced job turnover rate but that they explain no more than a small part of the equilibrium unemployment rate. Blanchard and Portugal (2001) develop a matching model in which the wage is negotiated once and for all at the outset of the matchup between employer and employee. A simulation of this model shows that the unemployment rate is an increasing, then a decreasing, function of firing costs. This result indicates that two countries—in this case the United States and Portugal, in the study of Blanchard and Portugal—may display identical unemployment rates while having very different legislation about employment protection (in the scale of strictness in employment protection reproduced in figure 13.7, the United States is one of the least strict countries and Portugal is one of the most strict). The simulations of Blanchard and Portugal do show, however, that the average duration of unemployment rises rapidly, and to a significant degree, when employment protection is strengthened.

Rigid Wages

The results are quite different when wages are rigid. Figure 13.13 represents the impact of administrative firing costs, on the assumption that there is a constant mandatory minimum wage and a corresponding unemployment rate of 10% in the absence of employment protection. In this situation, an increase in firing costs has a very marked impact on the unemployment rate. The latter rises by more than 10 points when firing costs increase by an amount corresponding to the average quarterly production of a worker. The exit rate from unemployment plummets, while job destruction is little changed. These results highlight the degree of interaction between the various institutions of the labor market. Employment protection leads to very different outcomes according to the nature of the other institutions that regulate the labor market. To be precise, the results obtained suggest that firing costs are probably unfavorable to the employment of low-skilled workers in certain European countries where a high proportion of

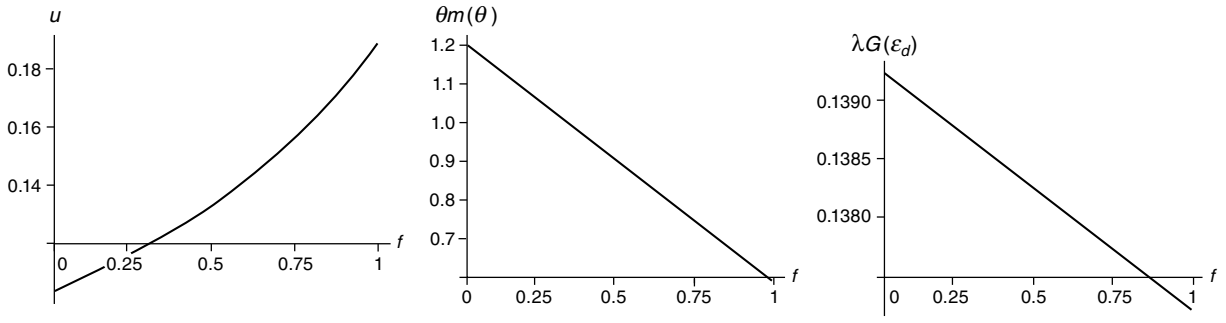


FIGURE 13.13
The impact of firing costs on the unemployment rate u, the exit rate from unemployment $\theta m(\theta)$, and the job destruction rate $\lambda G(\varepsilon_d)$ with exogenous wages. f is expressed as a fraction of average quarterly production.

them are paid at minimum wage. High firing costs would, however, have only negligible effects on employment if they were accompanied by high wage flexibility (Blanchard and Portugal, 2001).

It should be noted that the minimum wage on one hand and employment protection on the other exert directly opposite effects on the job destruction rate. Equation (13.35), which defines the reservation productivity when the wage is exogenous, shows that the minimum wage increases the job destruction rate, while firing costs reduce it. Bertola and Rogerson (1997) have pointed out that this type of effect might explain the similar rates of job destruction observed in different OECD countries with very different kinds of employment protection.

Also worthy of note is the fact that the effects of minimum wage and employment protection on the exit rate from unemployment have a tendency to mutually reinforce one another (see figure 13.13). The conjunction of a high minimum wage and rigorous employment protection ought thus to lead to relatively low exit rates from unemployment and, consequently, to a high proportion of long-term unemployed. The comparison of worker flows in France and the United States well illustrates this kind of effect, showing that the exit rate from unemployment is about seven times higher in the United States (see chapter 9, figure 9.13).

2.3 WHAT EMPIRICAL STUDIES SHOW

Recent research on the impact of employment protection has found that more stringent regulations do reduce job destruction. But along with that, stringent employment protection reduces productivity and the employment rate of some groups of workers (notably youth), and it increases unemployment duration and labor market segmentation.

2.3.1 THE IMPACTS ON THE LEVELS OF EMPLOYMENT AND UNEMPLOYMENT

The matching models used in this section have shown that employment protection has an ambiguous effect on the volume of employment. It certainly cuts back on job destruction, but it also diminishes job creation, since firms fear being unable, in future, to destroy unprofitable jobs protected by the legislation. So assessment of the impact

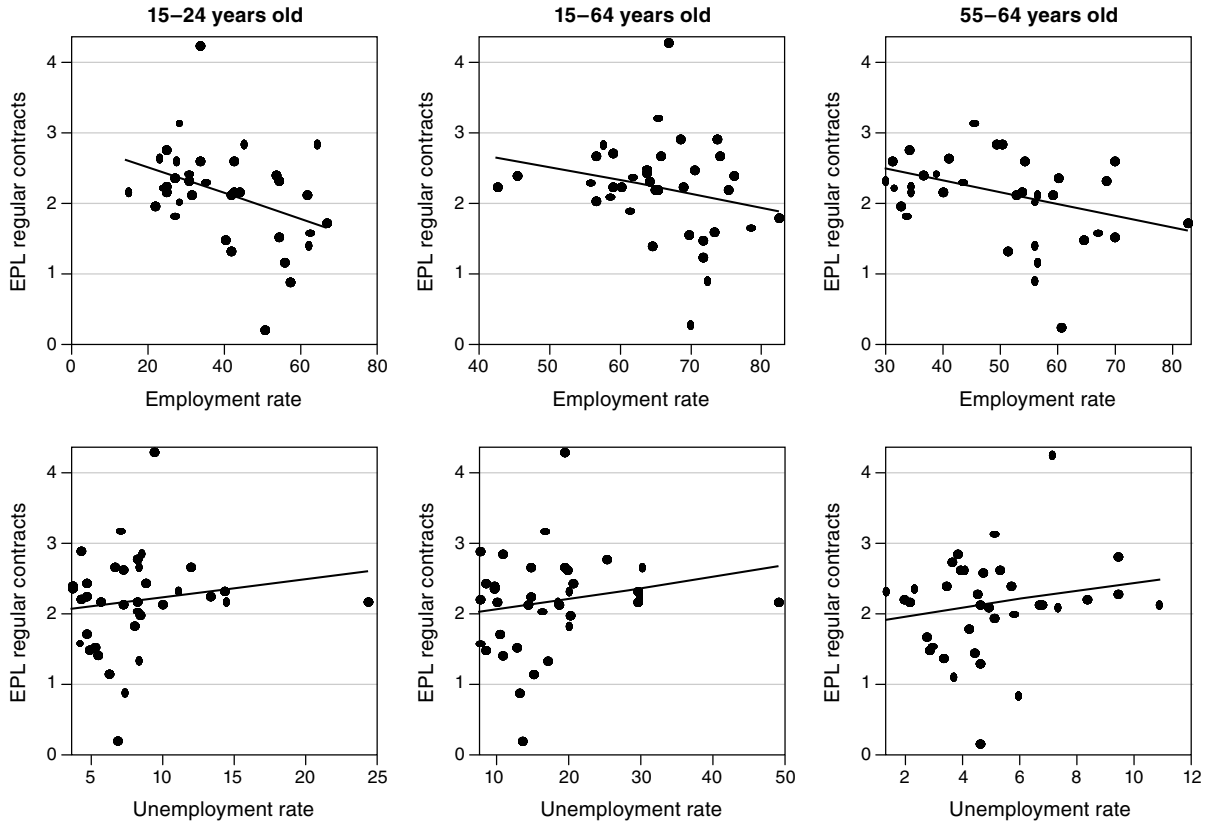


Figure 13.14

Employment protection on regular contracts, employment and unemployment rates in the 34 OECD countries, 2002–2012 averages. EPL = Employment Protection Legislation Index.

Source: OECD Employment Protection and Labor Force databases.

of employment protection remains primarily an empirical question. Much research has tackled this problem since the 1990s. As a general rule, researchers try to show a correlation, positive or negative, between the "rigor" of employment protection and the rate of unemployment, taking care to bracket all the other forces that might affect unemployment and employment. As shown on figure 13.14, the simplest cross-country correlations over a 10-year period show a negative relationship between employment protection and employment rates for different age groups, notably youth and older workers. Correlations are positive but weaker between employment protection and unemployment rates. These charts are purely descriptive and a number of confounding factors would need to be controlled for to make any inference about the causal effect of employment protection on labor market performances. Empirical studies of the impact of employment protection on unemployment and employment can be classified into two groups.

A first group of contributions analyze cross-country correlations of the type of those in figure 13.14 between unemployment and various indicators of employment protection legislation. The contributions of Lazear (1990), Nickell and Layard (1999), Blanchard and Wolfers (2000), Addison and Teixeira (2003), and Botero et al. (2004), among others, analyze this type of correlation. They generally find positive correlations between employment protection and unemployment. However, these results should be interpreted cautiously because changes in employment protection legislation and changes in unemployment can be codetermined by common factors. For instance, it is possible that negative macroeconomic shocks, which increase unemployment, also lead insiders to demand more job protection. Then, positive correlations between job protection and unemployment reflect not the positive impact of job protection on unemployment but the common impact of macroeconomic shocks on unemployment and employment protection legislation.

A second group of contributions at industry level, firm level, or the individual level allow for better identifications of the impact of labor market regulations on employment outcomes. In some cases, reforms of employment protection legislation were targeted at subgroups in the labor force, providing researchers with a natural experiment in which outcomes can be compared across subgroups. These studies find negative effects of job protection on employment and labor flows. For instance, Autor et al. (2006) estimate the effects on employment and wages of wrongful discharge protections adopted by U.S. state courts during the last three decades. Because state courts adopted the common-law wrongful discharge doctrines in different months and years during the 1980s and 1990s, the authors can compare, using a difference-in-differences approach, changes in employment and wages in states that adopted a given wrongful discharge doctrine in a given period with corresponding developments in states that did not adopt any doctrine during the same period. They base their analysis on the Current Population Survey (CPS) monthly files that provide data for approximately 100,000 adults over that period. They find that wrongful discharge protections reduced state employment rates by 0.8% to 1.7%, notably among the 41 states that adopted one specific type of wrongful discharge law—the implied-contract exception—which comes into force when an employer implicitly promises not to terminate a worker without good cause (see figure 13.15). The initial impact of protection is largest for female and less-educated workers, while the longer-term effect is greater for older and more-educated workers.

Using manufacturing data for India, Ahsan and Pagés (2009) study the economic effects of legal amendments on two types of labor laws: employment protection and labor dispute resolution legislation. They find that laws that increase employment protection or increase the cost of labor disputes substantially reduce registered sector employment and output. Almeida and Carneiro (2009) find that stricter enforcement of labor regulations constrains firm size and reduces the use of informal labor in Brazil. Micco and Pagés (2006) examine manufacturing data for a number of developed and developing countries and find that employment protection legislation constrains output and employment growth. The Spanish reforms of 1997, which reduced the dismissal costs of permanent jobs for workers under 30 and over 45 years old, but not for those aged 30 to 44 years, were associated with a relative increase in permanent employment for these groups (Kugler et al., 2005). Similarly, in Colombia in 1990, dismissal costs were lowered for jobs in the formal sector but not for the informal sector. This was associated with higher labor market turnover into and out of unemployment in the formal

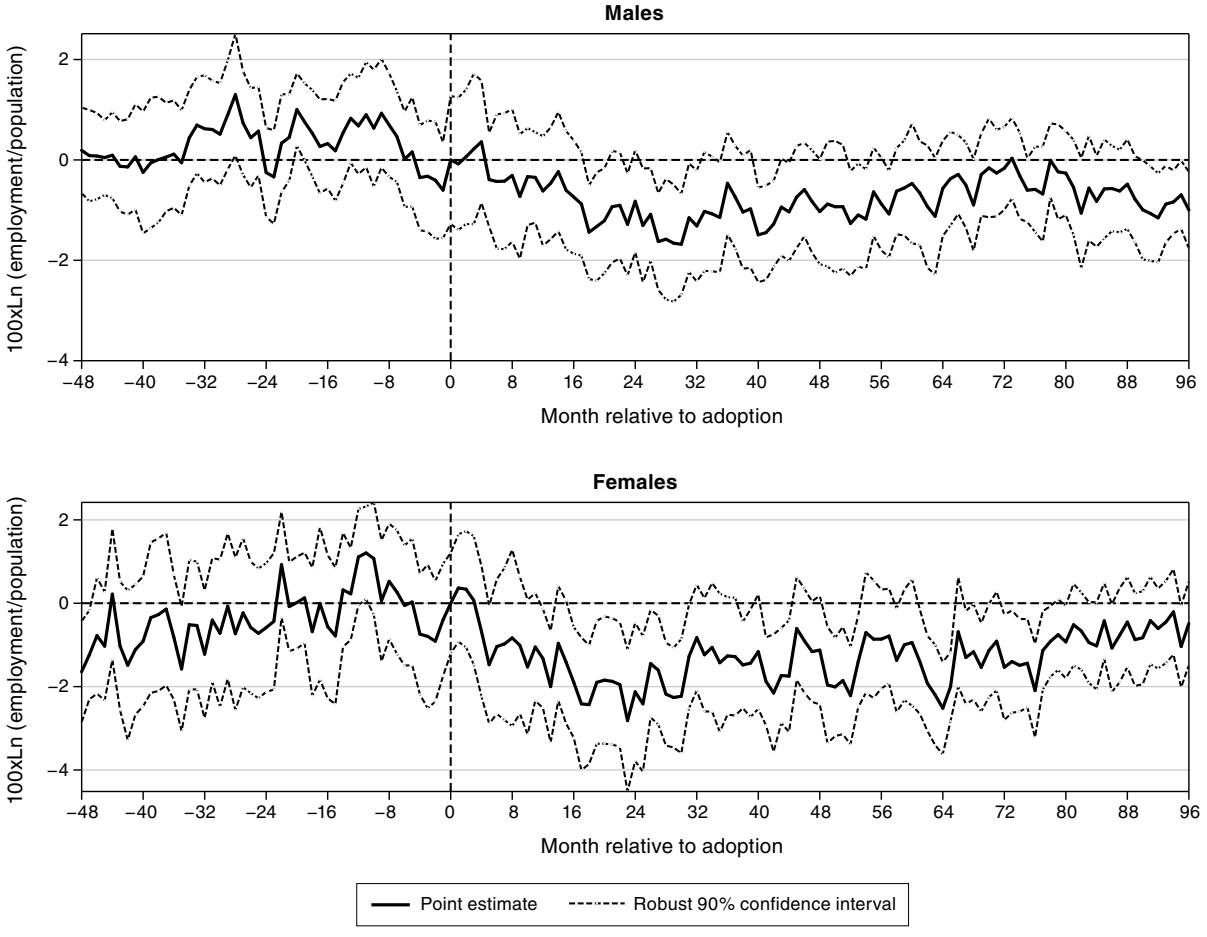


FIGURE 13.15
The impact of wrongful dismissal laws (of the "implied-contract exception" type) in the United States in the 1980s and 1990s.

Note: These figures plot estimated log employment-to-population ratios in adopting relative to nonadopting states at monthly intervals in the 4 years prior to through the 8 years following the adoption of the doctrine. The dashed lines in each figure represent robust 90% confidence intervals (allowing for arbitrary within-state error correlations) for each monthly point estimate. The month 0 corresponds to the date of adoption. The "implied-contract exception" is a type of wrongful discharge law that comes into force when an employer implicitly promises not to terminate a worker without good cause.

Source: Autor et al. (2006, figure 1, p. 216).

sector relative to the informal sector (Kugler, 1999). Increasing employment protection in the United Kingdom in 1999 lowered the probation period during which workers may not sue for unfair dismissal from two years to one year. This was associated with a decrease in the firing hazard for workers with up to two years of tenure relative to those with more tenure (Marinescu, 2007). The Italian reform of 1990 raising dismissal costs for firms with fewer than 15 workers was associated with reduced accessions and separations for these firms relative to larger firms (Kugler and Pica, 2008). Besley and

Burgess (2004) isolate the effect of a labor reform in a given state in India. They find labor regulations to have important adverse effects on output and employment, particularly in the registered manufacturing sector.

Overall, the two main conclusions we can draw from this set of studies are as follows.

1. The rigor of employment protection has no significant effect on the rate of unemployment. Hence more rigorous employment protection does not help to reduce the rate of unemployment.

2. Empirical studies, which rely on disaggregated data, find that more rigorous employment protection reduces employment.

2.3.2 Labor Market Segmentation

In practice, employment protection legislation induces labor market segmentation between unstable jobs with poor working conditions and stable jobs with better working conditions. This is because firms need to use more temporary jobs when protection of permanent jobs is stronger in order to adapt employment to changes in production. As shown by the top panel of figure 13.16, the share of temporary jobs tends to be higher in countries where protection of permanent jobs is more stringent.

More specifically, when there are substantial firing costs for permanent jobs, firms are relatively reluctant to hire new entrants into such jobs. Instead, new entrants are placed in temporary jobs where their productivity can be assessed before a permanent offer is made. New entrants disproportionately include the young, women, and, possibly, immigrants. The bottom panel of figure 13.16 shows that the share of young workers occupying temporary jobs grows with employment protection.

Available empirical work does indeed suggest that stringent regulation of permanent jobs increases labor market duality. Kahn (2007), using 1994–1998 International Adult Literacy Survey microdata, investigates the impact of employment protection laws on the incidence of temporary employment by demographic group. His study covers Canada, Finland, Italy, the Netherlands, Switzerland, the United Kingdom, and the United States—countries with widely differing levels of mandated employment protection. He finds that more stringent employment protection for permanent jobs (as measured by the OECD) increases the relative incidence of temporary employment for less experienced and less skilled workers and for young workers, native women, immigrant women, and those with low cognitive ability. This result is important, since temporary jobs tend to be lower paying and offer less training, other things being equal, than permanent jobs; moreover, workers in temporary jobs express lower levels of job satisfaction than comparable workers in permanent jobs (Booth et al., 2002). Thus, policies that lead to a substitution of temporary jobs for permanent jobs may actually worsen the welfare of the average worker, especially in the event that this policy does not lead to lower unemployment.

The labor market segmentation induced by stringent regulation of permanent jobs improves the security of permanent jobs but does so at the expense of an increasing instability of temporary jobs. Therefore, the impact of protection of permanent jobs

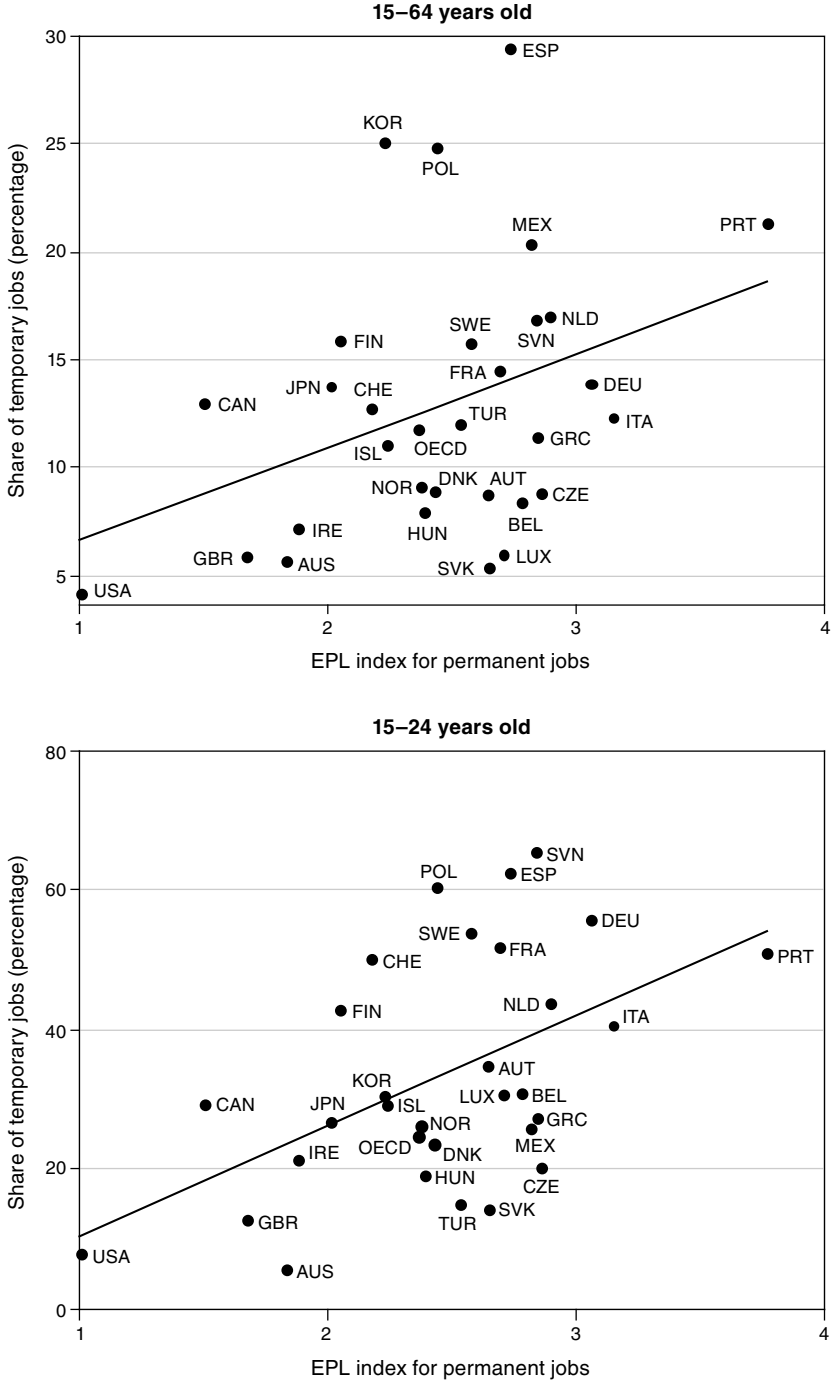


Figure 13.16
Share of 15–64- and 15–24-year-old employees in temporary jobs and protection of permanent jobs (average of protection against individual dismissal and specific requirements for collective dismissal), averages 2002–2012.

Source: OECD Labor Force and Employment Protection databases.

TABLE 13.5
The relations between employment protection, the generosity of unemployment benefits, and perceived job security in 12 European countries in the late 1990s.

Variable	Perceived job security		
	Permanent job, private sector	Permanent job, public sector	Temporary job
EPL index	−0.148 (.015)	0.007 (.016)	−0.134 (.019)
UI net replacement rate	0.883 (.079)	0.179 (.083)	1.400 (.099)

Note: All regressions control for age, education level, immigrant status, marital status, children, past unemployment experience, and local unemployment rate. The dependent variable is the perceived job security at individual level, after controlling in a first stage for macroeconomic and local labor market conditions. It is based on a question asked in the European Community Household Panel survey (ECHP): "How satisfied are you with your present job or business in terms of job security? Using the scale 1 to 6, please indicate your degree of satisfaction. Position 1 means that you are not satisfied at all, and 6 that you are fully satisfied." The EPL index and the unemployment benefit net replacement rates are from the OECD. Standard errors in parentheses.
Source: Clark and Postel-Vinay (2009, table 7, p. 231).

on overall job security is ambiguous. This property has been illustrated in search and matching models with temporary and permanent jobs (Blanchard and Landier, 2002; Cahuc and Postel-Vinay, 2002; Cahuc and Carcillo, 2006). Actually, more stringent regulation of permanent jobs can be associated with stronger feelings of job insecurity not only for temporary workers but also for permanent workers, as shown by Clark and Postel-Vinay (2009). They construct indicators of the perception of job security for various types of job in 12 European countries using individual data from the European Community Household Panel. Then, they consider the relation between reported job security and OECD summary measures of employment protection legislation strictness on one hand, and unemployment insurance benefit generosity on the other. Table 13.5 shows their main results. They find that after controlling for selection into job types as well as the state of local labor markets, workers feel most secure in permanent public-sector jobs and least secure in temporary jobs, with permanent private-sector jobs occupying an intermediate position. They also find that perceived job security in both permanent and temporary jobs is positively correlated with unemployment insurance generosity, while the relationship with employment regulation strictness is negative: workers feel less secure in countries where jobs are more protected! These correlations are absent for permanent public jobs, suggesting that such jobs are perceived to be, by and large, insulated from labor market fluctuations. While care needs to be taken in establishing the causality of these correlations, this result suggests that job protection is not the best response to the problem of job insecurity.

2.3.3 PRODUCTIVITY

Since job protection can have negative, but also positive, effects on productivity, the question of the actual impact of job protection on productivity is an empirical one.

Job Protection and Productivity: Theory

The search and matching model presented above shows that job protection reduces labor productivity because it creates labor hoarding. Indeed, job protection makes it more

difficult for firms to react quickly to rapid changes in technology or product demand that require reallocation of staff or downsizing, slowing the flow of labor resources into emerging high-productivity firms, industries, or activities (Hopenhayn and Rogerson, 1993). For instance, Saint-Paul (2002b) argues that stringent job protection may induce secondary innovations that improve existing products rather than introducing new products that may be more efficient but also riskier. Bartelsman et al. (2004) suggest that stringent employment protection legislation discourages firms from experimenting with new technologies that may exhibit higher mean returns but also higher variance. There is evidence that this mechanism may be at play. For instance, Pierre and Scarpetta (2005) show that innovative firms are the most negatively affected by stringent employment protection legislation.

But job protection can influence productivity through other channels.

First, job protection may induce workers to invest more in specific skills and to put more effort into cooperation within the firm because they anticipate that their long employment spell will allow them to get the returns to such investments (Wasmer, 2006b; Belot et al., 2007). This can improve labor productivity.

Second, job protection can lower the effort of workers because there is less threat of layoff in response to poor work performance or absenteeism. Observational data suggest that this channel might also play a role. Ichino and Riphahn (2005) show that the hike in job security at the end of the probation period induces a significant increase in absenteeism for white-collar workers in Italy. Similar findings are obtained by Riphahn (2004) using German data. Olsson (2009) analyzes the consequences of an exemption in the Swedish Employment Security Act (LAS) in 2001 which made it possible for employers with a maximum of 10 employees to exempt two workers from the seniority rule at times of redundancies. Using this within-country enforcement variation, the relationship between employment protection and sickness absence among employees is examined. The average treatment effect of the exemption is found to decrease sickness absence by more than 13% at those establishments that were treated relative to those that were not, and this was due to a behavioral rather than a compositional effect. The results suggest that the exemption had the largest impact on shorter spells and among establishments with a relatively low share of females or temporary contracts.

Again, the point to stress is that since job protection can have negative but also positive effects on productivity, the question of its actual impact on productivity is an empirical one.

Job Protection and Productivity: Empirical Results

The empirical literature dealing with the relationship between job protection and productivity can be classified into two types of contribution.

First, some contributions rely on aggregate cross-country data. These studies do not provide clear-cut conclusions. DeFreitas and Marshall (1998) find that stricter job protection has a negative impact on labor productivity growth in the manufacturing industries of a sample of Latin American and Asian countries. Nickell and Layard (1999) and Koeniger (2005) find weak positive relationships between the stringency of job protection, total factors productivity growth, and research and development intensity for OECD countries. These results are difficult to interpret because correlations observed with aggregate cross-country data do not allow us to pinpoint a causal impact of employment protection legislation on productivity.

A second set of contributions, using data at the industry, the firm, or the plant levels, provides more conclusive and more convincing results. Autor et al. (2007) study the impact of the adoption of wrongful-discharge protection norms by state courts in the United States using plant-level data from 1970 to 1999. They find that capital deepening is increased while employment flows, firm entry, and productivity are reduced. Similar findings are provided by Cingano et al. (2008) using Italian data to examine a 1990 reform that raised dismissal costs for firms with fewer than 15 employees only. In a study on job protection and job flows, Micco and Pagés (2006) also provide some weak evidence of a relationship between job protection and productivity, using a difference-in-differences estimator on a cross section of industry-level data for several OECD and non-OECD countries. They find a negative relationship between layoff costs and labor productivity. Bassanini et al. (2009) examine the impact of employment protection legislation on productivity in the OECD, using annual cross-country aggregate data on the degree of regulation and industry-level data on productivity from 1982 to 2003. They adopt a difference-in-differences framework, which exploits likely differences in the productivity effect of dismissal regulations in different industries. Their identifying assumption is that stricter employment protection influences worker or firm behavior, and thereby productivity, more in industries where the policy is likely to be binding than in other industries. The advantage of this approach is that, in contrast to standard cross-country analyses, it can control for unobserved factors that, on average, are likely to have the same effect on productivity in all industries. They find that mandatory dismissal regulations have a depressing impact on productivity growth in industries where layoff restrictions are more likely to be binding. Martins (2009) studies a quasi-natural experiment generated by a law introduced in Portugal in 1989: out of the twelve paragraphs in the law that dictated the costly procedure required for dismissals for cause, eight did not apply to firms employing 20 or fewer workers. Using detailed matched employer–employee longitudinal data and difference-in-difference matching methods, Martins examines the impact of that differentiated change in firing costs upon several variables, measured from 1991 to 1999. The results suggest that firing costs of the type studied here hurt firm performance, decrease workers' effort, and increase their bargaining power.

Overall, the empirical literature suggests that job protection has negative effects on productivity by lowering the involvement of workers in their job and by reducing the ability of employers to manage their manpower efficiently.

3 THE INTERPLAY BETWEEN EMPLOYMENT PROTECTION AND UNEMPLOYMENT INSURANCE

Unemployment insurance and legislation to protect employment both have the goal of damping down fluctuations in the income of wage earners over the course of their working lives. Hence it is important to analyze their joint impact and try to determine how an optimal insurance system should structure the combination of unemployment benefit and employment protection. Is it helpful to put in place employment protection

on top of unemployment benefit? If so, does there exist an optimal form of employment protection? How should it be linked with unemployment benefit?

This section supplies elements in response to these questions. The approach taken is normative. It will be shown that employment protection legislation can be justified by the need to protect workers from arbitrary dismissals and have firms internalize the social costs of labor turnover—costs which depend on the generosity of unemployment insurance benefit.

3.1 The Protection of Workers from Arbitrary Dismissals

The need to protect workers from arbitrary dismissal is met by the regulation of individual dismissals, according to which dismissal for reasons relating to the individual employee is justified only if the employee is guilty of breaking or failing to fulfill a contractual obligation. This is the just cause doctrine, adopted in European countries, which states that firms cannot dismiss employees without showing just cause.

This protection is not similarly granted in all countries. In the United States, the "employment at will" doctrine implies that either party can break the employment relationship with no liability, provided there was no express contract for a definite term governing the employment relationship and as long as the employer has not entered into a collective bargain (i.e., has not recognized a union; see chapter 7 for more detail on collective bargaining). Under this legal doctrine, any hiring is presumed to be "at will": the employer is free to discharge individuals for good cause or bad cause or no cause at all, and the employee is equally free to quit, strike, or otherwise cease work. There are several exceptions to the doctrine, especially if unlawful discrimination has played a part in the termination of an employee.[9] More generally, the Equal Employment Opportunity laws serve primarily to protect employees against violations of their work contract that do not respect the fundamental rights of the person. The basic argument put forward in defense of employment at will is that if just cause protection were worth more to employees than it costs firms, it would already have been implemented. If just cause protection were imposed from outside, wages would be lowered to compensate for the higher job security and the wage decrease would be worth less to the workers than the employment security. Thus, imposition of just cause policies will not help workers but will merely reduce the surplus from the worker–firm relationship (see Posner, 2003).

This reasoning is valid when labor markets are perfectly competitive. Under perfect competition, employers compete to attract workers and the competition among firms allows the workers to benefit from the best combinations of wages and working conditions available in the economy. In this context, the employment-at-will doctrine grants perfect protection to employees and there is no need for further employment protection legislation. However, labor markets are not perfectly competitive. Mobility

[9]During the 1970s and 1980s the majority of U.S. state courts adopted one or more common-law exceptions to the employment-at-will doctrine that limited employers' ability to fire. These are (1) the tort of wrongful discharge in violation of important public policy (public policy exception), (2) the implied covenant to terminate only in good faith and fair dealing (good-faith exception), and (3) the implied-in-fact contract not to terminate without good cause (implied-contract exception). See Autor et al. (2006) for a precise description of the employment-at-will doctrine and its exceptions.

costs, imperfect information, myopic behaviors, and contract incompleteness do not allow workers to fully benefit from competition among firms. Given that markets are not perfectly competitive, employment protection legislation can be useful to protect workers against the arbitrary decisions of employers. For instance, an employer who does not comply with health and safety regulations in the workplace may fire workers who complain. The employer may have an interest to do so if he has monopsony power, which allows him to replace those workers at low cost. Enacting a regulation that protects workers against such layoffs may improve efficiency.

However, even if there are good grounds for just cause protection, the legislation creating it should be drafted with caution because it can have perverse effects. For instance, Acemoglu and Angrist (2001) have studied the consequences of the Americans with Disabilities Act of 1990, which requires employers to accommodate disabled workers and outlaws discrimination against the disabled in hiring, firing, and pay. Although the Americans with Disabilities Act was meant to increase the employment of the disabled, the net theoretical effects are ambiguous because employers may find ways to avoid recruiting disabled employees. Actually, it would seem that the Americans with Disabilities Act has had an effect exactly opposite to its goal: for men of all working ages and women under 40, Acemoglu and Angrist find a sharp drop in the employment of disabled workers after the Americans with Disabilities Act went into effect.

Another example is given by Wasmer (2006b), who finds, using Canadian data including details on work-related stress and the consumption of various medications, that harassment of workers in order to induce a quit appears to be a substitute for greater freedom to simply dismiss. Wasmer finds positive links between individual employment protection and some dimensions of stress, and positive links between the stringency of employment protection, depression, and the consumption of various psychotropic drugs.

It should be noted that the benefits of job protection are generally unevenly distributed and can deteriorate the well-being of workers who do not enjoy them. In particular, temporary workers are generally disadvantaged by the protection afforded permanent jobs because employers are more reluctant to transform temporary jobs into permanent jobs when it is more costly to fire permanent workers. Therefore, stronger protection for permanent workers may indeed help permanent workers to keep their jobs, but at the expense of the unemployed and temporary workers whose opportunities to get stable jobs are reduced by this form of job protection. This mechanism explains why *insiders,* who occupy permanent jobs, can advocate stringent employment protection legislation at the expense of the *outsiders*, who do not occupy permanent jobs (see Saint-Paul, 2002a). Young people, less-skilled workers, and immigrants are those who are most frequently outsiders. Skilled prime-age males typically belong to the category of insiders.

3.2 The Internalization of the Social Costs of Labor Turnover

Modern economies are subjected to a permanent flux of technological innovations and changes in the preferences of individuals, necessitating the disappearance of some jobs and the creation of others (see chapter 9). This incessant process of job creation and destruction contributes to growth (see chapter 10). When a job vanishes for these reasons, it is thus not a loss for the collectivity, although it generally is, at least temporarily,

for the person who held that job. Legislation that prevented the destruction would by the same token have prevented a collective advantage from being realized. But conversely there are other reasons that weigh in favor of preserving certain jobs which firms might want to destroy. They spring from the difference between the *private value* and the *social value* of a job.

A worker is engaged by a firm to produce goods or services. This production represents the private value of the job and is split between a wage for the worker and profit for the firm. But the decision to destroy a job can have repercussions going well beyond the interests of the firm and the worker alone: it can give rise to externalities. In this case, the value of a job for the collectivity, its social value, does not coincide with its private value. The social value is measured by the sum of the private value and the value of the externalities.

One major reason for the gap between the social value and the private value of a job lies in the overall conception of the fiscal system. The largest portion by far of receipts to the tax authorities comes from persons who hold jobs. Unemployed and inactive persons contribute very little to the financing of collective goods and transfers. It follows that there is a gap between the social and the private value of a job, measured by the loss of compulsory payroll taxes and by extra costs in the form of social transfers that are triggered when someone moves from the status of wage earner to that of unemployed or inactive person. In most OECD countries this difference is considerable and justifies a form of employment protection.

The mode in which unemployment insurance and all forms of welfare are financed is another reason, perhaps more important than the previous one, for the divergence between the social and the private value of a job. In most industrialized countries, unemployment insurance is financed by a tax based on wages, which is paid in varying proportions by both employees and employers; it is one component of what are collectively called social security contributions. Under an efficient system of unemployment insurance, an employer who lets an employee go would have to take into account the externality arising from the financing of the unemployment insurance benefit then paid to that worker by other wage earners and other employers through their contributions to unemployment insurance. Under an efficient unemployment insurance system, the employer would also have to take into account the fact that the job she has destroyed will no longer contribute to financing the system. Absent such efficiency, every firm relies on all the other firms and wage earners to pay the unemployment benefits of the workers it lets go. The social value of a job exceeds its private value by an amount equal to what that person costs society during his spell of joblessness. In neglecting the externalities occasioned by their behavior when they let someone go, firms are reckoning only the private cost to themselves, not the real cost of this separation to society. In situations in which this real cost exceeds the individual cost to the firm, firms will have a tendency to destroy too many jobs.

The model that follows will show how externalities arising from employment protection can be internalized by integrating employment protection into the fiscal system.

3.2.1 A Model of Optimal Employment Protection and Unemployment Benefits

We will examine the optimal design of employment protection and its link with unemployment insurance using a static model taken from Blanchard and Tirole (2007). Cahuc

and Zylberberg (2005) have shown that the conclusions derived from this model hold good in a dynamic environment analogous to the matching model adopted in section 2.2 of this chapter. We begin by characterizing the social optimum. We then proceed to show how it can be implemented by resorting to a combination of firing taxes and unemployment benefits.

The Social Optimum

The economy comprises a continuum of identical workers and a continuum of identical entrepreneurs. The mass of each continuum is equal to 1. A firm is made up of an entrepreneur and a worker. To create a firm, an entrepreneur must hire a worker at the start of the period by paying a fixed cost denoted k. The productivity of the worker, denoted y, is not known at the start of the period and is only revealed at the end of the period. The set of possible values of productivity is described by a cumulative distribution function $G(y)$ defined on $[0, +\infty)$.

Once productivity is known, the firm can either keep the worker and produce or lay the worker off, who then becomes unemployed. When an entrepreneur hires a worker, he announces the wage w that she will receive if she is not laid off. By assumption, this wage is not renegotiable once productivity has been revealed (Blanchard and Tirole look at the case with wage renegotiation and show that the conclusions are not substantially different). Workers present aversion to risk. We denote by $v(w)$ the utility that wage w procures for the worker, with $v' > 0$ and $v'' < 0$. At the start of the period, each worker also knows that she will receive an unemployment benefit b paid by the state if she is laid off, in which case her level of utility amounts to $v(z + b)$, where z is an exogenous parameter representing all the gains (leisure or domestic production for example) that an individual can expect if she does not work.

Once the productivity y of a worker has been observed, the role of an omniscient planner is to decide whether she ought to continue to produce or be consigned to unemployment. The planner's task is thus to set a threshold y_d for productivity such that if $y > y_d$, the worker produces quantity y, and if $y < y_d$ she is consigned to unemployment and produces nothing. The rate at which jobs are destroyed—or what comes to the same thing in this model, the unemployment rate—is thus equal to $G(y_d)$. The planner must also decide the consumptions, again denoted w and b for simplicity, available to the worker if she is producing goods and if she is unemployed. The planner's program consists of maximizing the expected utility of a worker under a resource constraint. It is written thus:

$$\max_{y_d, w, b} G(y_d)v(z + b) + [1 - G(y_d)]\, v(w)$$

under the constraint:

$$\int_{y_d}^{+\infty} y \mathrm{d}G(y) \geq k + G(y_d)b + [1 - G(y_d)]\, w \qquad \textbf{(13.55)}$$

The resource constraint signifies that the average production of an individual, $\int_{y_d}^{+\infty} y \mathrm{d}G(y)$, must at least cover her average consumption $G(y_d)b + [1 - G(y_d)]\, w$, to which is added the fixed cost k of creating a job.

Let us denote by μ the multiplier associated with the resource constraint; the Lagrangian of the planner's problem takes the expression:

$$\mathcal{L} = G(y_d)v(z+b) + [1-G(y_d)]\,v(w) + \mu\left\{\int_{y_d}^{+\infty} y\mathrm{d}G(y) - k - G(y_d)b - [1-G(y_d)]\,w\right\}$$

The first-order conditions are obtained by setting the partial derivatives equal to 0 with respect to w, b, and y_d, which yields:

$$\frac{\partial \mathcal{L}}{\partial w} = [1-G(y_d)]\,v'(w) - \mu\,[1-G(y_d)] = 0 \Longrightarrow v'(w) = \mu \tag{13.56}$$

$$\frac{\partial \mathcal{L}}{\partial b} = G(y_d)v'(z+b) - \mu G(y_d) = 0 \Longrightarrow v'(z+b) = \mu \tag{13.57}$$

$$\frac{\partial \mathcal{L}}{\partial y_d} = G'(y_d)\,[v(z+b) - v(w)] + \mu G'(y_d)\,(-y_d - b + w) = 0 \tag{13.58}$$

Equations (13.56) and (13.57) immediately yield:

$$z + b = w \tag{13.59}$$

At the social optimum, the hypothesis that suppliers of labor are risk averse entails that they must be perfectly insured against the risk of job loss. Equation (13.58) then furnishes the threshold value y_d, or:

$$y_d = z \tag{13.60}$$

This equality conveys a *productive efficiency* condition. Let us take an individual for whom $y < z$: what this means is that her "market" production is inferior to her domestic production. By dispensing this person from work, the planner bears a cost $b + y$ equal to the sum of foregone production y and the consumption b of a jobless person, but he saves the consumption of a wage earner, equal to w. Under the condition of perfect insurance (13.59), we then have $w = z + b > y + b$. The planner thus realizes a net gain without the utility of the person in question being altered, since she is perfectly insured against the risk of joblessness. Consequently, from the standpoint of the social optimum, only individuals for whom $y > z$ ought to be in employment. That is what the equality (13.60) expresses.

Implementing the Social Optimum with Layoff Taxes

In an economy with decentralized and perfectly competitive markets, does the state dispose of adequate levers to implement the social optimum? To address this question, Blanchard and Tirole (2007) assume that in such an economy decisions are taken in the three following stages.

Stage 1. The state chooses and announces a payroll tax rate $\tau(w)$, a layoff tax rate $f(w)$, and unemployment benefit $b(w)$ that depends on wage w.

Stage 2. Entrepreneurs decide whether to start firms and pay the fixed cost k. They also announce the wage w that will be paid to employees.

Stage 3. The productivity of each job is revealed and firms decide whether to keep or dismiss workers. Those kept on receive wage w and those let go receive unemployment benefit $b(w)$.

To characterize the equilibrium of this economy, the sequence of decisions has to be taken backward. At stage 3, an entrepreneur employing a worker whose revealed productivity does reach level y realizes a profit equal to $y - w - \tau(w)$. Conversely, if he decides to dismiss, he must pay a cost equal to $f(w)$. An entrepreneur employs all workers whose productivity is such that $y - w - \tau(w) \geq -f(w)$. The threshold of job destruction is thus characterized by the equality:

$$y_d = w + \tau(w) - f(w) \tag{13.61}$$

At stage 2, if an entrepreneur does decide to enter the market, he must pay the fixed cost k and his expected profit will then amount to:

$$\mathbb{E}(\pi) = \int_{y_d}^{+\infty} [y - w - \tau(w)] \, \mathrm{d}G(y) - f(w)G(y_d) - k$$

If we assume that there is free entry into this market and that entrepreneurs compete on wages, the equilibrium level of wages verifies the equality $\mathbb{E}(\pi) = 0$. We thus have:

$$\int_{y_d}^{+\infty} y \mathrm{d}G(y) = k + f(w)G(y_d) + [1 - G(y_d)] \, [w + \tau(w)]$$

At stage 1, the state must choose $\tau(w)$, $f(w)$, and $b(w)$ while respecting the budget constraint:

$$f(w)G(y_d) + [1 - G(y_d)] \, \tau(w) = b(w)G(y_d) \tag{13.62}$$

The left-hand side of this equality represents the state's resources, which come from payroll taxes and layoff taxes; the right-hand side represents the state's expenditure on unemployed persons.

It is easy to show that the state can implement the social optimum while respecting its budget constraint. If workers are to be perfectly insured against the risk of joblessness, equation (13.59) shows that for every wage, the state must set an amount of unemployment benefit such that $b(w) = w - z$. Equation (13.61) defining the threshold of job destruction in a market economy, and condition (13.60) of productive efficiency, entail $z = w + \tau(w) - f(w)$ and since we have $b(w) = w - z$, we arrive at:

$$f(w) = b(w) + \tau(w) \tag{13.63}$$

This is the essential result we can draw from this analysis. It shows that for the social optimum to be implemented, the cost of a layoff $f(w)$ must exactly offset the burden of loss that this layoff places on the social security system both by creating another unemployed person costing $b(w)$ and by reducing the intake from payroll tax by an amount $\tau(w)$. Cahuc and Zylberberg (2008) have shown that formula (13.63) retains validity with hypotheses more general than those adopted in this basic model, on the assumption, for example, that workers may differ in their abilities or that it is possible for them to retire from the labor force. The generality of this result is easily grasped: when it is the state that is in charge of financing the unemployment insurance system, a layoff creates a negative externality for the collectivity. Relation (13.63) is classic in this context, signifying simply that the layoff tax must internalize the costs that a private decision forces the collectivity to bear. For that matter, it is this principle that guides the experience rating systems in place in various U.S. states.

Bringing the value of $f(w)$ derived from (13.63) into the budget constraint (13.62), we find $\tau(w) = 0$. The state can thus attain the social optimum by adopting an economic policy defined by:

$$f(w) = b(w) = w - z,\ \tau(w) = 0$$

To implement the social optimum, in this basic model, the state does not even need a payroll tax; it need only set the layoff tax at the same level as unemployment benefit. Nor is this an astonishing result: it corresponds to the Tinbergen rule that only as many levers are needed as there are policy goals. Here the state's only goal is to furnish unemployment benefit to those who lose their jobs, so it only requires one lever. If it is also necessary to ensure a subsistence income for those not in the labor force, then the state must make use of the payroll tax as well as the layoff tax (a case studied in Cahuc and Zylberberg, 2008).

To close this model, it is necessary to specify the value of the wage at equilibrium. The resource constraint (13.55) makes this possible. As $b(w) = w - z$ and $y_d = z$, this constraint is written:

$$w \leq -k + z + \int_z^{+\infty} (y - z)\, \mathrm{d}G(y)$$

If we assume that the state's goal is to furnish agents with the greatest possible utility, in this simple model that entails having the highest possible wage. At equilibrium the resource constraint will thus be an equality, which determines the amount of the wage.

3.2.2 Experience Rating

As the model just developed shows, experience rating is a way to require employers to help pay for the unemployment benefit payments they bring into existence through their firing decisions. The systematic use of experience rating is an original feature of the U.S. unemployment benefit system: in most states, unemployment benefits are financed by

taxing firms in proportion to their separations. In contrast, experience rating is absent from the unemployment compensation systems of other OECD countries, where benefits are usually financed by payroll taxes paid by employers or employees and by government contributions.

In the United States, employer contributions to unemployment insurance depend on the number of layoffs they have accumulated over the course of the previous three or five years. In practice the methods of calculation are highly complex and vary from one state to another. For that matter, employers do not bear the entire cost. Around 40% is mutualized. The empirical literature on the question speaks in this context of an "implicit subsidy" (to layoffs). Hence the American system is an imperfect experience rating system. If the experience rating were perfect, the implicit subsidy would be null.

Using data flowing from, among other sources, the Current Population Survey (CPS), Topel (1983) estimated the elasticity of the unemployment rate with respect to the implicit subsidy furnished by the experience rating system. He finds that going over to a perfect experience rating system would reduce the unemployment rate by around 30%. This result is confirmed by Anderson and Meyer (1994) using individual data from firms. They estimate that 23% of temporary layoffs and 21% of permanent separations can be attributed to the implicit subsidy. The conclusions of Card and Levine (1994), who examine fluctuations in layoffs using data flowing from the CPS between 1979 and 1987, are in line with those already cited. They suggest that moving over to a perfect experience rating system could reduce partial unemployment by around 20%.

The most probative piece of research is that of Anderson and Meyer (2000), which relies on a natural experiment. Between 1972 and 1984, the state of Washington was characterized by a flat rate of employer contributions. But from 1985 onward, the law changed and a system of experience rating was put in place. Anderson and Meyer (2000) exploit this shift in the legislation to assess the impact of the introduction of experience rating. They find that it reduced turnover in the labor force. They also find that wages have dropped in sectors that featured "a lot" of layoffs before the 1985 reforms, which suggests that firms may have compensated for the hike in firing costs by lowering their wages.

4 SUMMARY AND CONCLUSION

- The generosity of unemployment insurance systems varies widely from one country to another, with net replacement rates that go from 25% to 65% of net wage. For the poorest unemployed, housing assistance, family benefits, and social assistance schemes often top up unemployment insurance benefit. The fact is that only a minority of job seekers benefit from unemployment insurance in the OECD countries, for many of them do not meet the eligibility conditions.
- The optimal amount of unemployment benefit tails off with the elasticity of unemployment duration with respect to unemployment benefit. It increases with the liquidity effect and decreases with the moral hazard effect.

- Raising unemployment benefits is desirable whenever it raises the after-tax reservation wage.
- Optimal unemployment benefits should be decreasing with the duration of an unemployment spell. Empirical analysis suggests that they should be contracyclical, meaning they should rise when the overall level of unemployment rises.
- Employment protection legislation is a set of mandatory restrictions governing the dismissal of employees. According to the synthetic index of the strictness of employment protection established by the OECD, the strictest employment protection for regular contracts is found in Germany, Belgium, the Netherlands, and France. The countries of Southern Europe also feature strong employment protection. The weakest employment protection is found in the United States, the United Kingdom, Canada, and New Zealand. It is noteworthy that Nordic countries such as Denmark, Finland, and Norway are at or below OECD average, notably regarding collective dismissals.
- A priori, firing costs have an ambiguous effect on unemployment and reduce manpower mobility by reducing both job creation and job destruction at the same time. When wages are bargained over, an increase in firing costs entails lower wages, and this attenuates the negative effects on job creation. On the other hand, if wages are exogenous (as they are, for example, in the case of workers being paid minimum wage), this attenuating mechanism is absent. Calibration exercises confirm that if wages are bargained over, employment protection measures have little influence on job creation, job destruction, and the unemployment rate. If wages are rigid, the job destruction rate shows little sensitivity to firing costs, but exit rates from unemployment fall off sharply, and the unemployment rate soars.
- According to empirical research, the rigor of employment protection has no significant effect on the rate of unemployment. Hence more rigorous employment protection does not help to reduce the rate of unemployment but might reduce the rate of employment. Additionally, many studies bring out a positive relationship between the strictness of employment protection and the duration of unemployment.
- Employment protection legislation induces labor market segmentation between unstable jobs with poor working conditions and stable jobs with better working conditions.
- Empirical studies relying on microeconomic data generally find that employment protection is detrimental to employment and labor productivity.
- Employment protection legislation can be justified by the need to protect workers from arbitrary dismissals and have firms internalize at least some of the social costs of labor turnover.
- An efficient unemployment insurance system must structure the combination of employment protection and unemployment benefit in a coherent fashion. The American system of experience rating broadly conforms to this guideline. It consists of adjusting the contributions to unemployment insurance demanded from individual firms in light of the volume of their past layoffs, thus making it possible to internalize the social costs of labor turnover.

5 RELATED TOPICS IN THE BOOK

- Chapter 2, section 3: Dynamic labor demand
- Chapter 5: Job search
- Chapter 6, section 2: Risk-sharing
- Chapter 6, section 3: Incentives in the presence of verifiable results
- Chapter 6, section 4: Incentives in the absence of verifiable results
- Chapter 9, section 3: The matching model
- Chapter 14, section 2.4: Evaluation of labor market policies
- Chapter 14, section 4.2: Job search assistance and monitoring

6 FURTHER READINGS

Addison, J., & Teixeira, P. (2003). The economics of employment protection. *Journal of Labor Research*, *24*, 85–129.

Blanchard, O., & Tirole, J. (2007). The optimal design of unemployment insurance and employment protection: A first pass. *Journal of the European Economic Association*, *6*(1), 45–77.

Chetty, R. (2006). A general formula for the optimal level of social insurance. *Journal of Public Economics*, *90*, 1879–1901.

Chetty, R. (2008). Moral hazard vs. liquidity and optimal unemployment insurance. *Journal of Political Economy*, *116*(2), 173–234.

Hopenhayn, H., & Nicolini, J. (2009). Optimal unemployment insurance and employment history. *Review of Economic Studies*, *76*, 1049–1070.

Shimer, R., & Werning, I. (2007). Reservation wages and unemployment insurance. *Quarterly Journal of Economics*, *122*(3), 1145–1185.

7 APPENDIX: THE COEFFICIENT OF RELATIVE RISK AVERSION AND THE COEFFICIENT OF RELATIVE PRUDENCE

Let us suppose that the consumption (or the income) which an agent discounts in the future is a random variable C of the "proportional risk" kind, described by:

$$C = \bar{C}(1 + \varepsilon) \Longleftrightarrow \frac{C - \bar{C}}{\bar{C}} = \varepsilon \tag{13.64}$$

where ε represents a disturbance with zero mean and variance γ^2 (in other words, $\mathbb{E}(\varepsilon) = 0$ and $\mathbb{E}(\varepsilon^2) = \gamma^2$). $\bar{C}$ is the average, or the "certainty equivalent," of random consumption C.

Equation (13.64) signifies that the *relative* gap between consumption and its average is a white noise. An agent who presents risk aversion prefers the certain outlook $\bar{C}$ rather than awaiting a draw from the random "lottery" C. That being so, he would even be willing to deduct from $\bar{C}$ a "risk premium" so as to obtain immediately at least as much as he could expect on average from a lottery draw. If the preferences of the agent are summed up by a utility function $v(.)$, the risk premium, denoted P_r, is defined by:

$$\mathbb{E}v(C) = v\left(\bar{C} - P_r\right) \tag{13.65}$$

This equation simply means that the certain outlook $\left(\bar{C} - P_r\right)$ is equivalent to the random outlook C.

A second order Taylor development of $v(C)$ gives:

$$v(C) = v\left[\bar{C}(1+\varepsilon)\right] \simeq v(\bar{C}) + \varepsilon\bar{C}v'(\bar{C}) + \frac{1}{2}\left(\varepsilon\bar{C}\right)^2 v''(\bar{C})$$

From which it results that:

$$\mathbb{E}v(C) \simeq v(\bar{C}) + \frac{1}{2}\gamma^2\bar{C}^2v''(\bar{C}) \tag{13.66}$$

In addition, a second order Taylor development of $v\left(\bar{C} - P_r\right)$ gives:

$$v\left(\bar{C} - P_r\right) \simeq v(\bar{C}) - P_rv'(\bar{C}) + \frac{1}{2}P_r^2v''(\bar{C}) \tag{13.67}$$

If ε represents a small disturbance, the risk premium P_r is also a small quantity, and we can neglect the term P_r^2 in equation (13.67). Adopting definition (13.65) of the risk premium, (13.66) and (13.67) then entail:

$$\frac{P_r}{\bar{C}} = -\frac{1}{2}\gamma^2\frac{\bar{C}v''(\bar{C})}{v'(\bar{C})}$$

This expression shows that the prime rate $\left(P_r/\bar{C}\right)$ is a function of the characteristics of the random process governing consumption (here, the variance γ^2) and of another term, called the relative degree of risk aversion and defined by $\sigma(\bar{C}) = -\left[\bar{C}v''(\bar{C})/v'(\bar{C})\right]$, which depends only on the agent's preferences. The greater an agent's relative degree of risk aversion, the greater the prime rate must be.

The relative degree of *prudence* is a measure of the *intensity* of the agent's risk aversion. It is defined with reference to marginal utility. Let us henceforth consider a "prudence premium" Q_r such that we have:

$$\mathbb{E}v'(C) = v'\left(\bar{C} - Q_r\right)$$

This equality means that the certain outlook $\left(\bar{C} - Q_r\right)$ is equivalent to the random outlook C as far as the marginal utility of an agent is concerned. The Taylor developments applied to function $v'(.)$ rather than to function $v(.)$ bring us to:

$$\frac{Q_r}{\bar{C}} = -\frac{1}{2}\gamma^2\frac{\bar{C}v'''(\bar{C})}{v''(\bar{C})}$$

In this expression there appears the relative degree of prudence defined by $\rho(\bar{C}) = -\bar{C}v'''(\bar{C})/v''(\bar{C})$. For an individual presenting risk aversion (i.e., $v'' < 0$), we have $\rho(\bar{C}) > 0$ if $v'''(\bar{C}) > 0$. Kimball (1990) has shown that an agent whose income is random saves up more, the higher his degree of prudence.

REFERENCES

Acemoglu, D., & Angrist, J. (2001). Consequences of employment protection: The case of the Americans with Disabilities Act. *Journal of Political Economy*, *109*, 915–957.

Addison, J., & Teixeira, P. (2003). The economics of employment protection. *Journal of Labor Research*, *24*, 85–129.

Ahsan, A., & Pagés, C. (2009). Are all labor regulations equal? Evidence from Indian manufacturing. *Journal of Comparative Economics*, *37*, 62–75.

Almeida, R., & Carneiro, P. (2009). Enforcement of regulation, informal employment, firm size and firm performance. *Journal of Comparative Economics*, *37*(1), 28–46.

Anderson, P., & Meyer, B. (1994). The effects of unemployment insurance taxes and benefits on layoffs using firm and individual data (Working Paper No. 4960). NBER, Cambridge, MA.

Anderson, P., & Meyer, B. (2000). The effects of the unemployment insurance payroll tax on wages, employment, claims and denials. *Journal of Public Economics*, *78*, 81–106.

Autor, D., Donohue, J., & Schwab, S. (2006). The costs of wrongful-discharge laws. *Review of Economics and Statistics*, *88*(2), 211–231.

Autor, D., Kerr, W., & Kugler, A. (2007). Do employment protections reduce productivity? Evidence from US states. *Economic Journal*, *117*, F189–F217.

Baily, M. (1978). Some aspects of optimal unemployment insurance. *Journal of Public Economics, 10*, 379–402.

Bartelsman, E., Bassanini, A., Haltiwanger, J., Jarmin, R., Scarpetta, S., & Schank, R. (2004). The spread of ICT and productivity growth: Is Europe really lagging behind in the new economy? In D. Cohen, P. Garibaldi, & S. Scarpetta (Eds.), *The ICT revolution: Productivity differences and the digital divide*. Oxford, U.K.: Oxford University Press.

Bassanini, A., Nunziata, L., & Venn, D. (2009). Job protection legislation and productivity growth in OECD countries. *Economic Policy*, *24*, 349–402.

Belot, M., Boone, J., & van Ours, J. (2007). Welfare effects of employment protection. *Economica*, *74*, 381–396.

Bertola, G., & Rogerson, R. (1997). Institutions and labor reallocation. *European Economic Review*, *41*, 1147–1171.

Besley, T., & Burgess, R. (2004). Can labor regulation hinder economic performance? Evidence from India. *Quarterly Journal of Economics*, *119*, 91–134.

Blanchard, O., & Landier, A. (2002). The perverse effects of partial labour market reform: Fixed-term contracts in France. *Economic Journal*, *112*, F214–F244.

Blanchard, O., & Portugal, P. (2001). What hides behind an unemployment rate: Comparing Portuguese and US labor markets. *American Economic Review, 90*(1), 187–207.

Blanchard, O., & Tirole, J. (2007). The optimal design of unemployment insurance and employment protection: A first pass. *Journal of the European Economic Association*, *6*(1), 45–77.

Blanchard, O., & Wolfers, J. (2000). The role of shocks and institutions in the rise of European unemployment: The aggregate evidence. *Economic Journal*, *110* (supplement), 1–33.

Boeri, T., & Bruecker, H. (2011). Short-time work benefits revisited: Some lessons from the Great Recession. *Economic Policy*, *26*(68), 697–765.

Booth, A., Francesconi, M., & Frank, J. (2002). Temporary jobs: Stepping stones or dead ends. *Economic Journal*, *112*, F189–F213.

Botero, J., Djankov, S., La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2004). The regulation of labor. *Quarterly Journal of Economics*, *119*, 1339–1382.

Brenke, K., Rinne, U., & Zimmermann, K. (2013). Short-time work: The German answer to the Great Recession. *International Labour Review*, *152*(2), 287–305.

Cabrales, A., & Hopenhayn, H. (1997). Labor market flexibility and aggregate employment volatility. In B. McCallum (Ed.), *Carnegie-Rochester conference series on public policy* (vol. 46, pp. 189–228). Amsterdam: Elsevier.

Cabrales, A., & Hopenhayn, H. (1998). Job dynamics, correlated shocks and wage profiles (Working Paper 260). Universitat Pompeu Fabra, Barcelona, Spain.

Cahuc, P., & Carcillo, S. (2006). The shortcomings of a partial release of employment protection laws: The case of the 2005 French reform (IMF Working Paper No. 06/301).

Cahuc, P., & Carcillo, S. (2011). Is short-time work a good method to keep unemployment down? *Nordic Economic Policy Review*, *1*(1), 133–165.

Cahuc, P., & Lehmann, E. (2000). Should unemployment benefits decrease with unemployment spell? *Journal of Public Economics*, *77*(1), 135–153.

Cahuc, P., & Postel-Vinay, F. (2002). Temporary jobs, employment protection and labor market performance. *Labour Economics 9*, 63–91.

Cahuc, P., & Zylberberg, A. (2005). Optimum income taxation and layoff taxes (IZA Discussion Paper No. 1478). Institute for the Study of Labor, Bonn, Germany.

Cahuc, P., & Zylberberg, A. (2008). Optimum taxation and layoff taxes. *Journal of Public Economics*, *92*, 2003–2019.

Card, D., & Levine, P. (1994). Unemployment insurance taxes and the cyclical and seasonal properties of unemployment. *Journal of Public Economics*, *53*(1), 1–29.

Chetty, R. (2006a). A general formula for the optimal level of social insurance. *Journal of Public Economics*, *90*, 1879–1901.

Chetty, R. (2006b). A new method of estimating risk aversion. *American Economic Review*, *96*(5), 1821–1834.

Chetty, R. (2008). Moral hazard vs. liquidity and optimal unemployment insurance. *Journal of Political Economy*, *116*(2), 173–234.

Cingano, F., Leonardi, M., Messina, J., & Pica, G. (2008). Employment protection legislation, productivity and investment: Evidence from Italy. Mimeo, University of Salerno.

Chui, H., & Karni, E. (1998). Endogenous adverse selection and unemployment insurance. *Journal of Political Economy*, *106*, 806–827.

Clark, A., & Postel-Vinay, F. (2009). Job security and job protection. *Oxford Economic Papers*, *61*, 207–239.

DeFreitas, G., & Marshall, A. (1998). Labour surplus, worker rights and productivity growth: A comparative analysis of Asia and Latin America. *Labour*, *12*(3), 515–539.

Duell, N., Grubb, D., Singh, S., & Tergeist, P. (2010). Activation policies in Japan (Working Paper No. 113). OECD Social, Employment and Migration, Paris.

Feldstein, M., & Poterba, J. (1984). Unemployment insurance and reservation wages. *Journal of Public Economics*, *23*(1–2), 141–167.

Fredriksson, P., & Holmlund, B. (2001). Optimal unemployment insurance in search equilibrium. *Journal of Labor Economics*, *19*, 370–399.

Garibaldi, P. (1998). Job flow dynamics and firing restrictions. *European Economic Review*, *42*, 245–275.

Grubb, D. (2001). Eligibility criteria for unemployment benefits. In *Labour market policies and the public employment service* (pp. 205–237). Paris: OECD Publishing.

Gruber, J. (1997). The consumption smoothing benefits of unemployment insurance. *American Economic Review*, *87*, 192–205.

Hijzen, A., and Venn, D. (2011). The role of short-time work schemes during the 2008–09 recession (Working Paper No. 115). OECD Social, Employment and Migration, Paris.

Hopenhayn, H., & Nicolini, J. (1997). Optimal unemployment insurance. *Journal of Political Economy*, *105*, 412–438.

Hopenhayn, H., & Nicolini, J. (2009). Optimal unemployment insurance and employment history. *Review of Economic Studies*, *76*, 1049–1070.

Hopenhayn, H., & Rogerson, R. (1993). Job turnover and policy evaluation: A general equilibrium analysis. *Journal of Political Economy*, *101*(5), 915–938.

Ichino, A., & Riphahn, R. (2005). The effect of employment protection on worker effort: A comparison of absenteeism during and after probation. *Journal of the European Economic Association*, *3*(1), 120–143.

Jung, P., & Kuester, K. (2011). Optimal labor market policy in recessions (Working Paper No. 11-48). Federal Reserve Bank of Philadelphia.

Kahn, L. (2007). The impact of employment protection mandates on demographic temporary employment patterns: International microeconomic evidence. *Economic Journal*, *117*(521), F333–F356.

Kimball, M. (1990). Precautionary savings in the small and in the large. *Econometrica*, *58*, 53–73.

Koeniger, W. (2005). Dismissal costs and innovation. *Economics Letters*, *88*(1), 79–85.

Kroft, K., & Notowidigdo, M. (2011). Should unemployment insurance vary with the unemployment rate? Theory and evidence (Working Paper). University of Chicago.

Kugler, A. (1999). The impact of firing costs on turnover and unemployment: Evidence from the Colombian labour market reform. *International Tax and Public Finance Journal*, *6*, 389–410.

Kugler, A., Jimeno, J., & Hernanz, V. (2005). Employment consequences of restrictive permanent contracts: Evidence from Spanish labor market reforms (Working Paper). University of Houston.

Kugler, A., & Pica, G. (2008). Effects of employment protection on worker and job flows: Evidence from the 1990 Italian reform. *Labour Economics*, *15*, 78–95.

Landais, C. (2013). Assessing the welfare effects of unemployment benefits using the regression kink design (Working Paper). London School of Economics.

Landais, C., Michaillat, P., & Saez, E. (2010). Optimal unemployment insurance over the business cycle (NBER Working Paper No. 16526).

Lazear, E. (1990). Job security provisions and employment. *Quarterly Journal of Economics*, *105*, 699–725.

Marinescu, I. (2007). Shortening the tenure clock: The impact of strengthened UK job security legislation (Working Paper). University of Chicago.

Meyer, B. (1990). Unemployment insurance and unemployment spells. *Econometrica*, *58*, 757–782.

Meyer, B. (1995). Lessons from the U.S. unemployment insurance experiments. *Journal of Economic Literature*, *33*, 91–131.

Martins, P. (2009). Dismissals for cause: The difference that just eight paragraphs can make. *Journal of Labor Economics*, *27*(2), 257–279.

Micco, A., & Pagés, C. (2006). The economic effects of employment protection: Evidence from international industry-level data (IZA Discussion Paper No. 2433).

Millard, S., & Mortensen, D. (1997). The unemployment and welfare effects of labour market policy: A comparison of the USA and the UK. In D. Snower & G. de la Dehesa (Eds.), *Unemployment policy: Governement options for the labour market*. CEPR, Cambridge University Press.

Mitman, K., & Rabinovich, S. (2011). Pro-cyclical unemployment benefits? Optimal policy in an equilibrium business cycle model (PIER Working Paper No. 11-023).

Mortensen, D., & Pissarides, C. (1994). Job creation and job destruction in the theory of unemployment. *Review of Economic Studies, 61*, 397–415.

Mortensen, D., & Pissarides, C. (1999). Unemployment responses to skill-biased technology shocks: The role of labour market policy. *Economic Journal*, *109*, 242–265.

Nickell, S., & Layard, R. (1999). Labor market institutions and economic performance. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 3C, pp. 3029–3084). Amsterdam: Elsevier.

OECD. (1994). *The OECD jobs study*. Paris: OECD Publishing.

OECD. (1996). *Employment outlook*. Paris: OECD Publishing.

OECD. (1999). Employment protection and labor market performance. In *Employment outlook* (chap. 2). Paris: OECD Publishing.

OECD. (2013). Protecting jobs, enhancing flexibility: A new look at employment protection legislation. In *Employment outlook* (chap. 2). Paris: OECD Publishing.

Olsson, M. (2009). Employment protection and sickness absence. *Labour Economics*, *16*, 208–214.

Pierre, G., & Scarpetta, S. (2005). Employment protection: Do firms' perceptions match with legislation? *Economics Letters*, *90*, 328–334.

Pissarides, C. (2001). Employment protection. *Labour Economics*, *8*, 131–159.

Posner, R. (2003). *Economic analysis of law*. New York, NY: Aspen Publishers.

Riphahn, R. (2004). Employment protection and effort among German employees. *Economics Letters*, *85*, 353–357.

Saint-Paul, G. (2002a). The political economy of employment protection. *Journal of Political Economy*, *110*, 672–704.

Saint-Paul, G. (2002b). Employment protection, international specialization, and innovation. *European Economic Review*, *46*(2), 375–395.

Shavell, S., & Weiss, L. (1979). The optimal payment of unemployment insurance benefits over time. *Journal of Political Economy*, *87*, 1347–1362.

Shimer, R., & Werning, I. (2007). Reservation wages and unemployment insurance. *Quarterly Journal of Economics*, *122*(3), 1145–1185.

Shimer, R., & Werning, I. (2008). Liquidity and insurance for the unemployed. *American Economic Review*, *98*(5), 1922–1942.

Stokey, N., Lucas, R., & Prescott, R. (1989). *Recursive methods in economic dynamics*. Cambridge, MA: Harvard University Press.

Topel, R. (1983). On layoffs and unemployment insurance. *American Economic Review*, *73*(4), 541–559.

Venn, D. (2012). Eligibility criteria for unemployment benefits: Quantitative indicators for OECD and EU countries (OECD Social, Employment and Migration Working Paper No. 131). OECD, Paris.

Wang, C., & Williamson, S. (1996). Unemployment insurance with moral hazard in a dynamic economy. In B. McCallum (Ed.), *Carnegie-Rochester conference series on public policy* (vol. 44, pp. 1–41). Amsterdam: Elsevier.

Wang, C., & Williamson, S. (2002). Moral hazard, optimal unemployment insurance and experience rating. *Journal of Monetary Economics*, *49*, 1337–1371.

Wasmer, E. (2006a). General versus specific skills in labor markets with search frictions and firing costs. *American Economic Review*, *96*(3), 811–831.

Wasmer, E. (2006b). The economics of Prozac: Do employees really gain from strong employment protection? (IZA Discussion Paper No. 2460). Institute for the Study of Labor, Bonn, Germany.

CHAPTER 14

Active Labor Market Policies

In this chapter we will:

- Survey the variety of labor market policies that have been tried in the OECD countries to improve the labor market outlook of the unemployed
- Consider how efficient active labor market policies, such as job search assistance, training, job subsidies, and temporary public employment, are in an equilibrium framework
- Understand how externalities can reduce the impact of targeted measures
- Learn the methodological principles that guide the evaluation of labor market policies, depending on the nature of available data (controlled experiment vs. observational data)
- Identify the impact of placement programs for youth, based on a randomized experiment led in the 2000s by Crépon, Duflo, Gurgand, Rathelot, and Zamora (2013) (Data and programs are available at www.labor-economics.org)
- Find out what assessments of labor market policies reveal

INTRODUCTION

Intervention by the state in the labor market is generally viewed as taking two forms: active policies and passive policies. The goal of active policies is to increase employment and wages among persons who find insertion into the labor market difficult. Job search assistance, upgrades to professional training, employment subsidies, and even public-sector job creation are the commonest forms. Passive policies aim rather to increase the material welfare of disadvantaged populations without a priori attempting to improve their labor market performance. Unemployment insurance, already studied in chapter 13, and provisions for early retirement fall under this heading.

Active policies, while they are generally justified by the many sources of inefficiency in the functioning of the labor markets, do not make it possible systematically to

improve the performance of these markets. Theoretical study and empirical evaluation both show that they can even turn out to be counterproductive. For example, the creation of temporary public-sector jobs intended to facilitate the entry of youth into the labor market can, because of cost burdens and low efficiency, lead to a decline in the total number of jobs held by this category of the population. Similarly, subsidies to promote certain types of employment run the risk of displacing workers whose jobs do not benefit from these subsidies. These remarks show that it is misleading to prejudge the effect of public interventions without engaging in closer scrutiny and evaluating all of their effects quantitatively. The purpose of this chapter is to set forth the state of theoretical and empirical knowledge in this area.

The first section supplies the main facts regarding employment policies in the OECD countries, highlighting the fact that different countries have had different experiences in this regard. Section 2 is dedicated to theoretical analysis of active labor market policies and makes abundant use of the matching model set out in chapter 9. This model proves particularly useful because it allows us to study the effects of various policy measures on labor market efficiency. Section 3 presents the methods of evaluation. It explains especially how these methods succeed (or not) in identifying the impact of labor market policies. Finally, section 4 presents the main empirical results that have been obtained in the realm of active policy.

1 Labor Market Policies: An International Perspective

We see great diversity in the policies adopted and the amount of financing channeled into them from one country to another. Active labor market policies aim at improving the situation, in terms of employment and wages, of the unemployed and of disadvantaged populations generally. They are to be distinguished from passive policies that aim at increasing the well-being of these groups without automatically pursuing a particular outcome in terms of placement in the labor market. They are also to be distinguished from more general policies like those intended to protect employment or guarantee a minimum wage, for the latter affect the whole labor force, not just narrowly targeted groups.

1.1 What Are Active Labor Market Policies?

The OECD employs a standard typology of labor market policies that distinguishes between active and passive measures. This typology has the advantage of being widely adopted and thus allowing us to make international comparisons. But it has the drawback of excluding some large-scale programs that, like active measures in the strict sense, also aim at lower unemployment.

1.1.1 The OECD Classification

In the strict sense in which the OECD and other international databases such as at Eurostat use the term, "active labor market programs" (ALMPs) include only policy measures

that are targeted at particular groups in the labor market in order to help them find jobs:

1. Placement and related services, which comprise information services, referral to job opportunities, job search assistance, counseling and case management of job seekers, as well as financial help to cover the cost of job search or mobility to take up work, provided either by the public employment services or by other publicly financed bodies.

2. Training, institutional or on-the-job, including course costs for the unemployed, subsidies to employers who offer training in some cases (but excluding training programs available for all employees), remedial education for disadvantaged youth, and support for the recruitment of apprentices from specific groups, as well as subsistence allowances while in training.

3. Hiring incentives or subsidized employment in the private sector for some groups of workers (e.g., youth and older workers).

4. Direct job creation (in the public or nonprofit sector), either of temporary positions or, in some cases, of regular jobs in the public sector, for some groups of workers (e.g., youth and older workers).

5. Support for unemployed persons starting up firms (in the form of unemployment benefits or special grants).

6. Vocational rehabilitation and sheltered work for the disabled (preparing people for integration into the regular labor market, but excluding social and medical rehabilitation).

7. Job rotation and job sharing, that is, full or partial substitution of an unemployed person for an existing employee.

As opposed to active measures, passive measures in the labor market are limited to cash benefits, including unemployment benefit (i.e., any benefit, either insurance-based or assistance-based, conditional upon job search activities) and early retirement programs for reasons related to the labor markets (as opposed to early retirement for health reasons).

These examples of active measures exclude measures which are not targeted at the unemployed or specific groups of disadvantaged workers, even though such measures could be considered part of the set of labor market policies aimed at reducing unemployment:

- For instance, general hiring subsidies for small firms or across-the-board reductions in the social security contribution for low-wage jobs would not be considered active measures in these databases, even though they are indeed part

of what is usually conceived as employment policy and might improve the job prospects of the unemployed.

- Similarly, in-work benefits targeted at low-income households may be used as an incentive to facilitate the transition from welfare to work, but if these benefits are equally available to the unemployed and to persons already in low-wage employment, they will not be regarded as an active measure according to the strict definition of active labor market policies.
- Short-time work schemes aim to reduce layoffs by allowing employers to temporarily reduce hours worked while compensating workers for the induced loss of income. A short-time work scheme can be thought of as an active measure because it allows firms to maintain the employment relationship instead of laying off workers. But short-time work is not recorded in existing labor market programs databases because it is not targeted at disadvantaged groups.

1.1.2 The Purposes of Active Labor Market Policies

Active labor market policies are public interventions designed to improve the situation, in terms of employment and wages, of the unemployed and of disadvantaged populations. As such, they may affect employment in different ways. Public employment services have the goal of reducing job search costs. Training programs, and many of the measures in favor of youth, aim to increase the "employability" of the persons concerned and ought to lead to a rise in individual productivity. Other policies have the objective of reducing the cost of labor or creating public-sector jobs directly. Unemployment insurance is viewed as a passive policy when it is regarded as pure insurance against risk and is quantified as all the transfers that go to eligible unemployed persons. However, we must carefully distinguish between this strictly financial aspect of the unemployment insurance system and the other things it does, like checking on search effort and sanctioning those who search half-heartedly; these ought instead to be considered as belonging to active policy. In what follows, we merely set out the specific purposes of the various active policies.

Public Employment Services

One of the aims of public employment services is to promote matches between firms with vacant jobs and persons looking for work. In all industrialized countries, specialized public agencies like the U.S. Employment and Training Administration, the Bundesagentur für Arbeit in Germany, and the Pôle Emploi in France supply services of this kind. But certain countries, like Japan, the United Kingdom, and the United States, have authorized private organizations ("providers") to compete with the public agencies in the job placement "market" (see section 2.1 below for a theoretical analysis). Australia is unique among the OECD countries in that its mainstream employment services are delivered exclusively by competing providers coordinated by a central agency. Among the activities of these public agencies and private organizations, it is *Job Search Assistance* (*JSA*) that falls into the category of active labor market policy. This assistance takes various forms according to cases. Sometimes it simply comes down to offering a certain number of free telephone calls for jobs listed by the agency. But unemployed persons may also be given help in drafting their résumés, in defining personalized search strategies and then putting them into operation, or in finding appropriate training. Checking

on the effort being made by the unemployed, and applying sanctions if necessary, are also part of the role of public employment services (see OECD, 2013, chapter 3, for a complete description of this role).

Labor Market Training

In many countries, Denmark and Germany for example, labor market training represents the bulk of active policy. It is often endorsed by politicians as the best weapon against unemployment. The prevalent form of labor market training is *classroom training* (*CT*). It takes place not in firms but in courses or temporary placements created by specialized establishments. The duration is generally brief, on the order of 3 or 4 weeks in Denmark and 3 months on average in the United States. The training may be general or specific to an industry or a firm. It may serve to make up for a gap in the basic education of some individuals (those, for example, who failed to finish, or even start, secondary school) or to bring the knowledge of skilled employees up to date.

Apprenticeship represents a large part of training measures aimed specifically at the young in most countries. Apprenticeship typically includes classroom instruction and on-the-job training. There are also programs to help disadvantaged or unemployed youth, addressed primarily to young people who leave school with no job to go to and those who drop out of high school prematurely. The *Job Corps* program in the United States is an example. It is aimed at young people from difficult urban neighborhoods, who must take training that gets them out of their normal environment. Many programs to help youth are not so precisely targeted, and there is little that really distinguishes them from general training programs. Some other training measures are not, for the most part, aimed specifically at the young. Rather, they represent an alternative to traditional classroom instruction. The goal of such *on-the-job training* (*OJT*) programs is to give employers an incentive, by means of a subsidy, to give training to disadvantaged categories of workers. An on-the-job training placement generally lasts from 3 to 12 months, and at the end of that period the employer has the opportunity to hire the trainee on a permanent basis. According to Heckman et al. (1999), in the United States these programs primarily make it possible to insert, or reinsert, certain persons into a work environment, and there may be no real distinction between them and programs that simply subsidize hiring.

Subsidized Employment

Subsidized employment covers a wide gamut of measures. Subsidies for employment in the private sector generally take the form of transfers to firms that hire members of particular groups. The transfer may be temporary, to subsidize the hiring of disadvantaged workers, or permanent, to subsidize the wage of these workers or of certain types of jobs. *Public-service employment* as an active policy measure is the direct creation of jobs in the public sector and is addressed in principle to the young and to the long-term unemployed. The purpose is to allow people who find themselves in this situation to hold a temporary job in the public sector so that they can acquire minimal skills or seniority as a step towards finding a regular job (or simply to make them eligible for unemployment insurance). Programs of this kind form a large part of the spectrum of active policy measures in Europe but are practically nonexistent in the United States (see OECD, 2007, chapter 5, for a comparative study of several OECD countries). It is important, however, to distinguish *temporary* public jobs created as part of an active labor market policy

from general public-sector policy, which consists of creating *permanent* civil service jobs. The overall extent of employment in the public sector is an "institution" proper to each country. The creation of temporary jobs in the public sector or in nonprofit organizations is intended to give a semblance of training and work habits to persons with little or no work experience who belong to economically disadvantaged groups. Finally, unemployed persons are given help in launching new enterprises in a number of countries (including the United States). Most often this involves the use of unemployment benefits to subsidize unemployed persons willing to have a go at becoming self-employed. Observation tells us that in general this measure applies only to a limited number of unemployed persons.

Another point to make is that the same individual may benefit from several of these measures at the same time, for public policy is often structured around programs with several facets. For example, the *Job Corps* program in the United States combines job search assistance, classroom training, and apprenticeship. Many programs are similarly multifaceted, which makes it more difficult to assess the effects specific to each measure. We also need to be aware that the distinction between active and passive measures is useful for analysis but that in practice the line between them is not always easy to draw. Comparisons between countries may be influenced by the type of benefits received by those losing their jobs. In the Netherlands or Norway, for example, the proportion of those benefiting from disability benefits is much higher than in most other countries. In these specific cases, what we really have is more a disguised form of assistance for certain categories of the unemployed, or preretirement support, than a measure specifically intended to get disabled people back into the labor force. Now, even though specific active measures exist for the disabled, this category of benefit recipients is typically more difficult to activate. The same phenomenon is not unknown in the United States: Autor and Duggan (2003) estimate that if access to disability insurance had not been made easier there in the middle of the 1980s, the current unemployment rate would be 2/3 of a percentage point higher. Similarly, certain youth training placements serve only to "park" the participants without really improving their productive capacities.

1.2 Differences Between Countries

Public employment policies vary widely both in the amount of money earmarked for them and in the way that money is divided up among the various policy options. Countries that spend a lot on benefits also usually spend a lot on active programs, but there is a wide variety of strategies. As opposed to passive policies, which tend to react swiftly to changes in unemployment, the active ones, expressed as a percentage of GDP, increase only slightly in recessions. In countries where unemployment benefits are generous or where employment protection is strong, active policies tend to be used more intensively to boost flows out of unemployment.

1.2.1 The Amount of Public Expenditure on Labor Market Policy

The amount of public funding earmarked for labor market policy (including both passive and active measures) varies widely from one country to another. The first panel of figure 14.1 gives an overview of this diversity over the 2000s. Japan, the

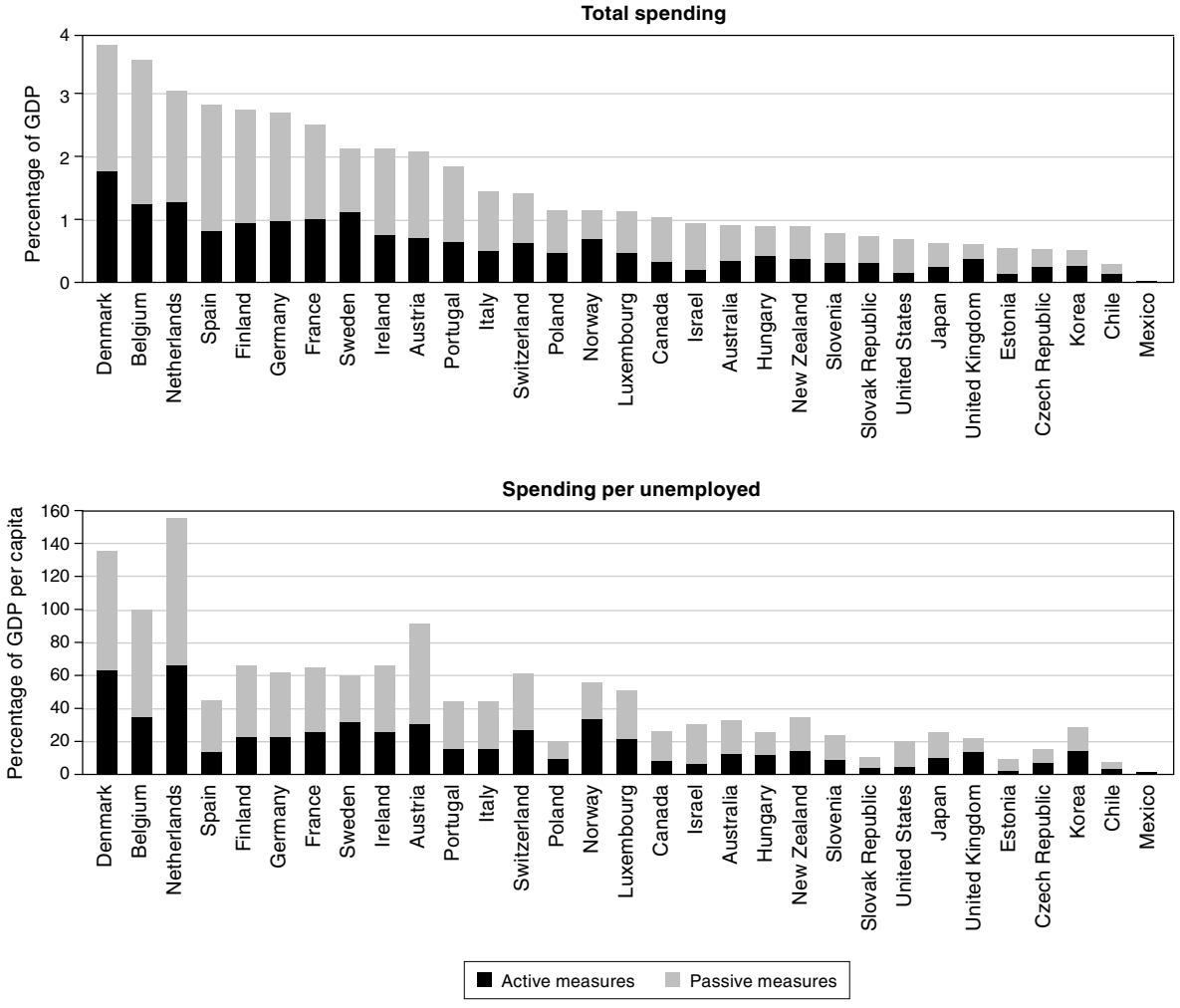


FIGURE 14.1

Spending on active and passive labor market programs in the OECD countries; average over the 2002–2011 period.

Note: In the bottom panel total spending in $PPP is divided by the number of unemployed, and the resulting spending per unemployed per person is then divided by the GDP per capita in $PPP.

Source: OECD Labor Market Programs database.

United Kingdom, and the United States are the countries that spend the least in this area (about 0.8% of GDP). The other Anglophone countries (Australia, Canada, and New Zealand) spend a larger share of their resources (between 1% and 1.5% of GDP). In contrast, other countries—mainly northern European ones—spend much more. In Denmark, for example, total public expenditure on labor market policy represents almost 3.7% of GDP; in the Netherlands this figure comes to around 3%; and in Sweden, 2%. Norway stands out among the Nordic countries on account of its relatively low outlay

on labor market policy: the order of magnitude is the same as in Canada. Germany and France occupy an intermediate position, spending about 2.5% of GDP.

The black portion of the bars in figure 14.1 represents the share of spending on active measures in each country. As a general rule the amount spent on passive policies clearly outstrips that spent on active ones. The Swedish, British, and Norwegian exceptions deserve notice. In Sweden, expenditure on labor market policy is divided in approximately equal parts between active measures and passive ones. Norway and the United Kingdom spend twice as much on active policy measures as they do on passive ones.

Now, of course the rate of unemployment in each country influences the level of spending. This is the case for instance in Spain, where the unemployment rate averages out to 13% in the 2000s. The second panel of figure 14.1 represents the average annual spending per unemployed person as a percentage of GDP per capita, and countries are ranked according to their total level of spending, as in the first panel. We see that Spain spends about as much as other Southern European countries, while France and Germany spend approximately as much as Sweden or Norway, about 60% of GDP per capita. However, Denmark, Belgium, and the Netherlands, which feature the highest levels of total spending, spend about twice as much per unemployed person as the countries previously mentioned.

1.2.2 How Public Expenditure on Active Employment Policy Is Divided Up

Figure 14.2 breaks down expenditure on active policy according to the first four OECD headings mentioned at the start of this section. Placement and administration of public employment services, plus training, are the two foremost categories of spending, followed by subsidies to the private sector and public employment. Independently of the volume spent on active employment policy, we note the wide range of choices of how to allocate it. Austria and Norway, for example, dedicate around 50% of their active policy expenditure to training, whereas the figures for Australia and the United Kingdom are around 6% for this item. The other countries fall in between, spending from 20% to 40%. France, Germany, and the United States allocate about a third of active spending to training. In many countries, training represents a greater outlay than placement and administration of the public employment services. Sweden, but also Italy and Spain, allocate more than a third of their spending to job subsidies in the private sector, while the Anglophone countries dedicate much smaller shares to the same item (between 2.5% and 6%). France, Belgium, Ireland, and Hungary feature high shares of spending on public job creation, between 20% and 40%. The rate even reaches 50% in Chile, while it is negligible or zero in Sweden and Denmark, and very low in Canada, the United Kingdom, and the United States. Australia allocates 17% of active spending to direct job creation, due to a workfare program in which some job seekers may be required to participate. Finally, it is interesting to note that the countries which, in global terms, spend little on active employment policy (Japan and the Anglophone countries) are also the ones which devote proportionally the most resources to public employment services. In these countries, between 30% and 50% of the money spent on active policies is dedicated exclusively to job search assistance. In the United Kingdom, this category represents almost all of spending. In Australia and Canada, placement represents more than half of total spending.

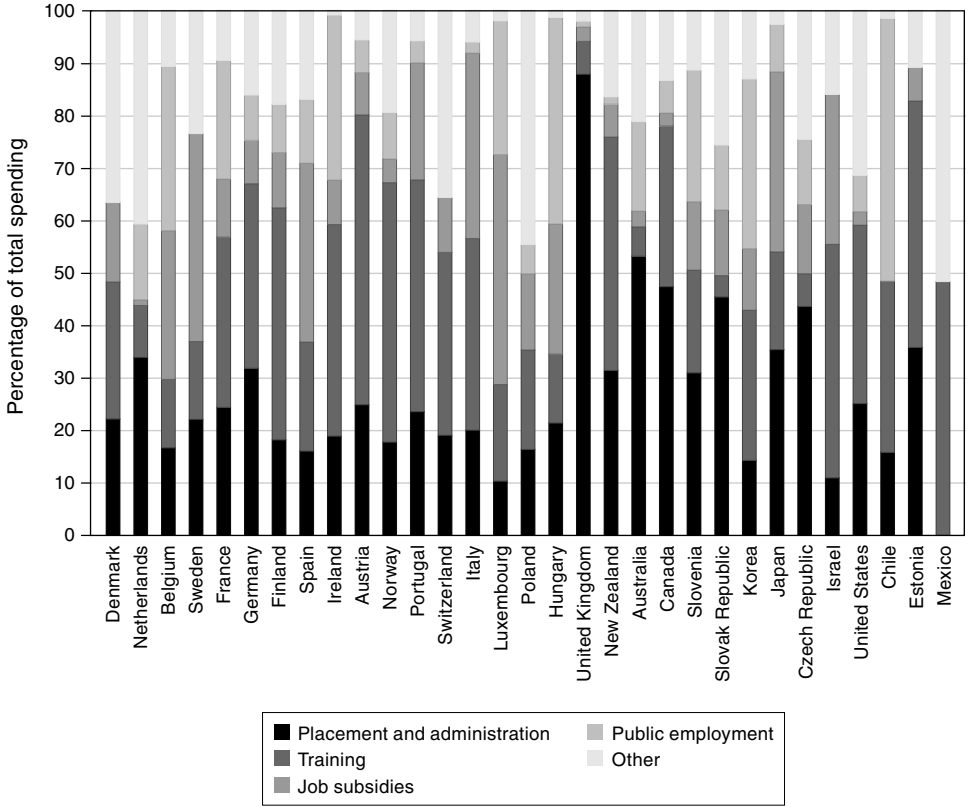


Figure 14.2
Breakdown of active spending by category in the OECD countries, average over the 2002–2011 period.

Note: Countries are ordered according to their level of total spending on active labor market programs (descending order).

Source: OECD Labor Market Programs database.

Figure 14.2 also shows that countries that spend more on active programs tend to use the whole range of interventions.

1.2.3 Changes Over the Business Cycle

There is no clear trend in spending on labor market measures over the last 30 years, but we do observe high volatility. Figure 14.3 plots active and passive spending as a percentage of GDP over time, as well as the rate of unemployment (right scale) in six countries. Within countries, passive measures, that is, benefits paid to the unemployed, are very much correlated with the rate of unemployment. The correlation is very high in the United States and the United Kingdom. In the case of the United States, the automatic variation of the maximum duration of unemployment benefits with the business cycle might reinforce this correlation. In France and Japan, the correlation is somewhat lower. The fact that spending on unemployment benefits augments almost instantly with unemployment is not surprising, since these benefits are an entitlement in all countries.

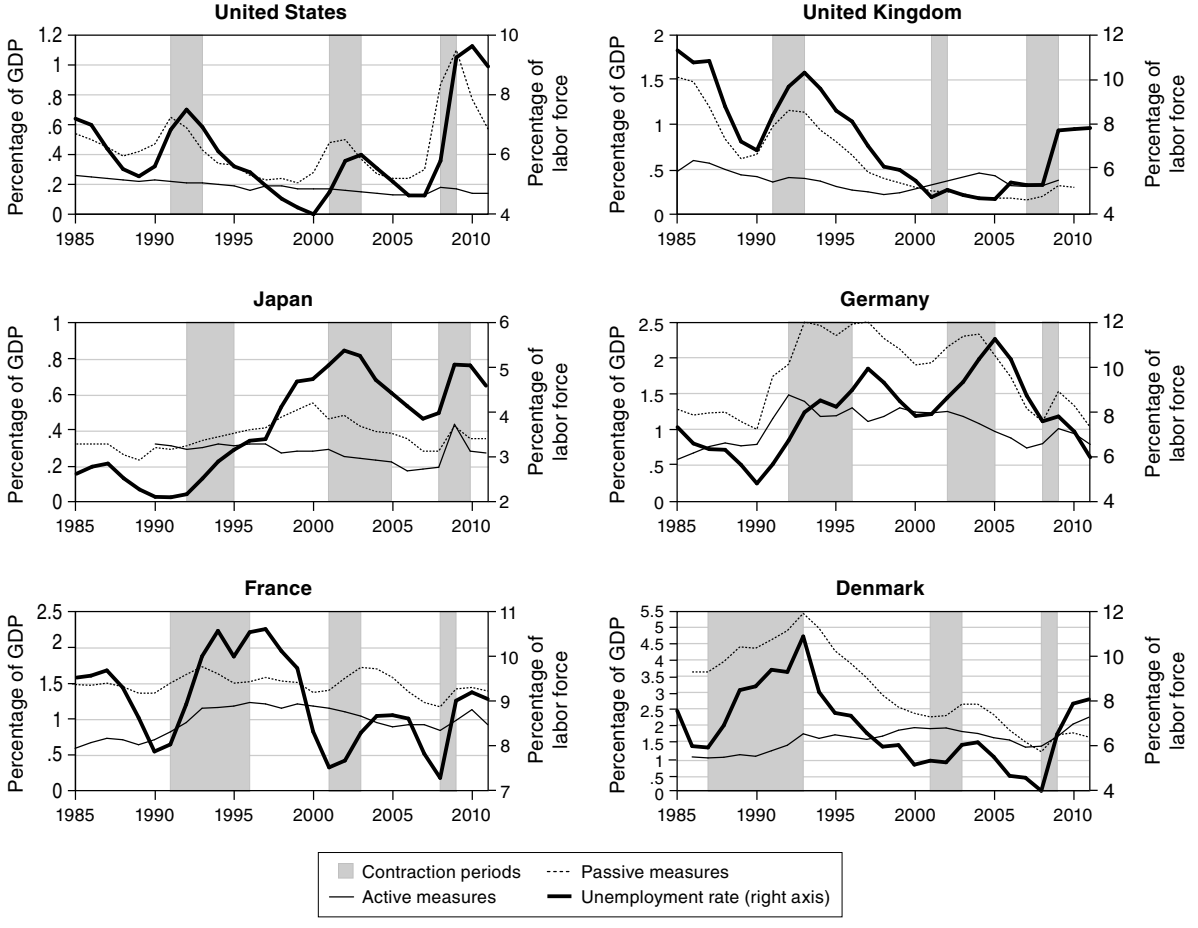


FIGURE 14.3
Spending on labor market programs and the economic cycle.

Source: OECD Labor Market Programs and Economic Outlook databases.

This is not the case for active measures. Benefiting from an active program most often depends on the decision of the caseworker at the public employment service, as well as on the availability of funds to finance the intervention. The corresponding budgets are discretionary and often have to be voted by local or national parliaments. As a result, total spending on active measures increases only very little with unemployment, even in Denmark, where spending reaches the highest level in the OECD countries. In the United States and Japan there is almost no correlation at all. As a consequence, the level of spending per unemployed person tends to decrease when unemployment rises, which is unlikely to help foster outflows from unemployment. It could be argued that when there are fewer jobs available, there is no point putting more means into placement. But, as we will see below, it is probably better to invest in some policies, such as training programs, when unemployment is high because this is the time when they are the most efficient.

1.2.4 What Are the Characteristics of Countries That Spend More on Active Labor Market Policies?

Figure 14.4 shows cross-country correlations between the unemployment rate and two measurements of spending on active measures in 31 OECD countries. There is no correlation at all between total spending as a percentage of GDP and unemployment (right panel of the figure). However there is a negative correlation between spending per unemployed person and the unemployment rate (left panel): countries that spend more tend to feature lower unemployment rates on average in the 2000s. Now, we cannot infer any causal relationship—by which more resources dedicated to activating each unemployed person would tend to lower unemployment—from a mere correlation of this kind. Indeed, there are many other factors we do not control for that could explain this correlation. For one thing, unemployment might cause spending as much as spending might cause unemployment. Note that if some reverse causality of this sort were in play, with unemployment causing more spending per unemployed for instance, then we should observe a correlation of the opposite sign, which is not the case. Second, and more important in this case, spending and unemployment might both be the result of other factors, such as the institutions in the labor market and the quality of other

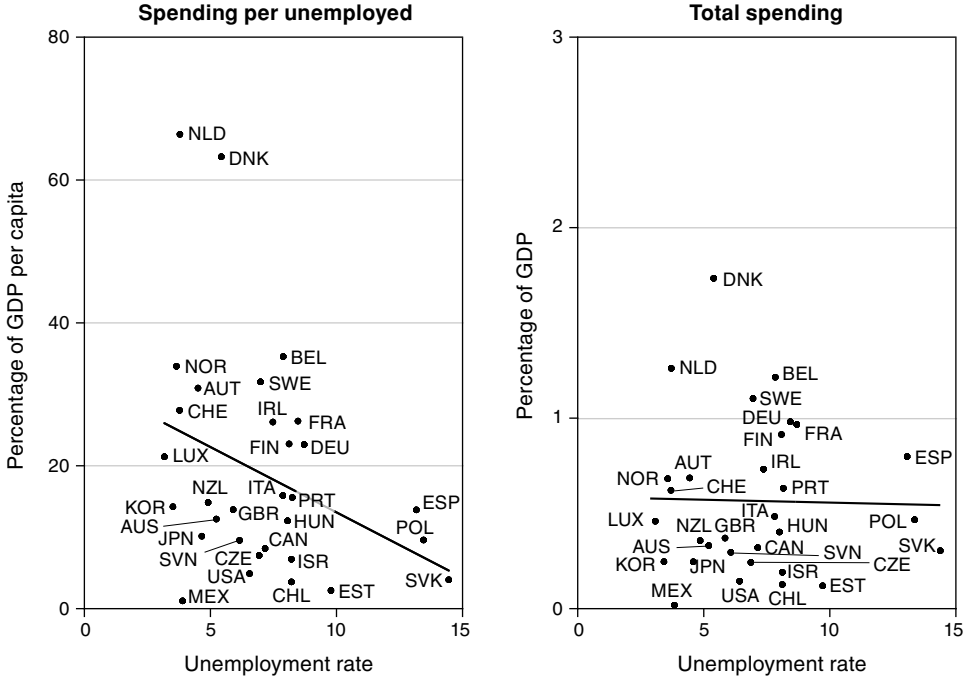


Figure 14.4
Spending on active labor market programs and unemployment, averages over the 2002–2011 period.

Note: In the left panel total spending in $PPP is divided by the number of unemployed, and the resulting spending per unemployed is then divided by the GDP per capita in $PPP.

Source: OECD Labor Market Programs and Labor Force Statistics databases.

public policies. For instance, if countries that spend more on active measures are also countries where educational policies or low-wage policies are the best suited to ensure a low rate of youth unemployment, this type of correlation may arise. Last, spending per unemployed person is going to be mechanically lower when unemployment is high, by construction. As we will see, the impact of active programs on unemployment is best identified with high-quality microeconomic data and well-tailored strategies.

Actually, some institutional features of the labor market are also positively correlated with spending on active measures. First, as shown on the left panel of figure 14.5, countries where the net replacement rate of unemployment benefits[1] is high also feature high spending per unemployed person. This is notably the case in the Northern European countries. Contrastingly, in the United States and Japan, where replacement

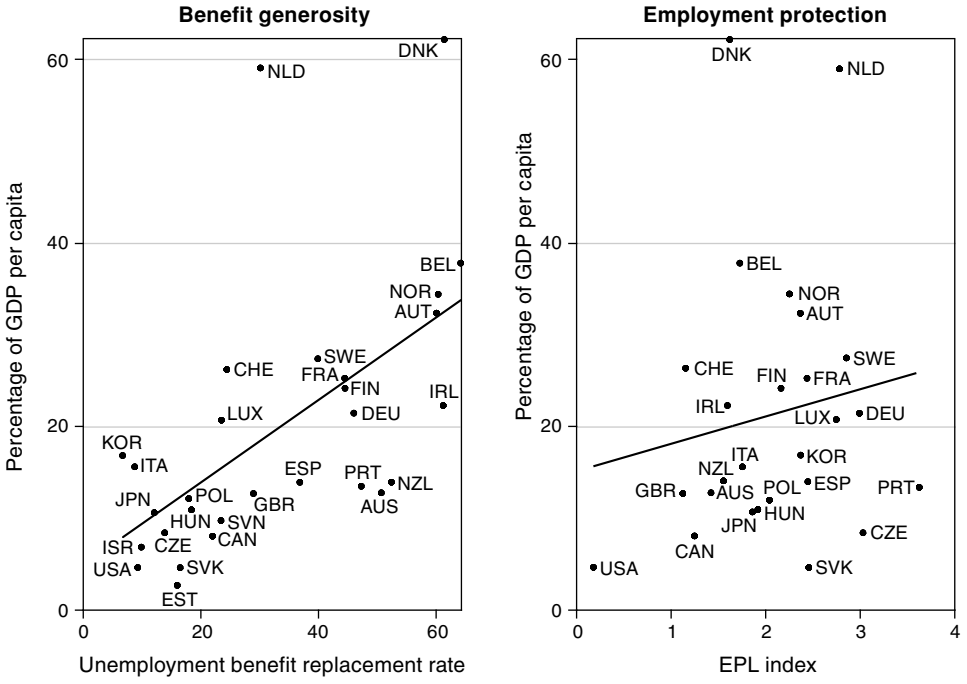


FIGURE 14.5

Spending on active programs per unemployed person and insurance systems in the labor market, averages over the period 2005–2010.

Notes: The replacement rate of unemployment benefits is the average ratio of net benefits to net income over a period of 5 years of unemployment based on the rules of benefits and tax systems. EPL index is the OECD index of the stringency of employment protection legislation (see chapter 13, section 2). Spending per unemployed person as a percentage of GDP per capita is defined as in figure 14.4.

Source: OECD Labor Market Programs and Employment Protection databases.

[1] The net replacement rate of unemployment benefits is the ratio of net unemployment benefits to net past earnings.

rates are low, spending is also low. There is evidence that in the Northern European countries, activation is seen as a way to ensure generous unemployment benefits while holding unemployment at low levels. Activation strategies, notably job search assistance and monitoring accompanied by sanctions, and also mandatory participation in active programs at a given stage in the unemployment spell, may indeed counteract the potential disincentives created by generous benefits. Second, as shown on the right panel of figure 14.5, spending per unemployed person is also positively correlated across countries with the strictness of employment protection, although the correlation is weaker than the previous one. We have seen already in chapter 13 that in countries where employment is more protected by legislation regulating layoffs, the exit rate from unemployment tends to be lower. Countries appear to counteract this effect by active measures aimed at stimulating outflows from unemployment. Actually, the correlation is even stronger when only job subsidies in the private sector are considered.

1.2.5 Examples of Active Policy in Several Countries

By way of illustration, we compare the American case with that of two European countries, Sweden and the United Kingdom. The United States and the United Kingdom display a degree of convergence, while the rise in unemployment during the 1990s brought a palpable change of direction to Swedish policy.

The United States

In the United States active employment policy targets economically disadvantaged groups, and the beneficiaries are often defined with reference to a poverty threshold.

The public job creation programs born in the 1970s, especially under the umbrella of the Comprehensive Employment and Training Act (CETA) of 1973, were gradually restricted to persons in difficulty before being abolished in 1983 by the government of Ronald Reagan. The New Jobs Tax Credit, set up in 1977, was a very large-scale program of nontargeted subsidies for employment in the private sector. It was replaced at the beginning of the 1980s by the more limited Targeted Jobs Tax Credit, which, as its title indicates, was intended for economically disadvantaged groups. More recently, in the aftermath of the 2008–2009 crisis, the HIRE Act (2010) temporarily renewed nontargeted hiring subsidies for employment in the private sector.

Programs of this kind, which aim to increase labor demand, are the exception in the United States. Most of the active policy measures which have followed one another since the beginning of the 1960s in this country are "supply-side" measures that aim to increase the human capital of the recipients. This approach is shared by the Manpower Development and Training Act (MDTA, 1962), the Comprehensive Employment and Training Act (CETA, 1973), and the Job Training Partnership Act (JTPA, 1983). Thus, the JTPA seeks to promote on-the-job training, classroom training, and work experience. This emphasis on education was maintained throughout the Clinton presidency. Another major item of active policy expenditure in the United States is job search assistance: figure 14.2 indicates that about 25% of active policy expenditure goes to public employment services and over 30% to labor market training. The Worker Profiling and Reemployment Services System set up in 1993 obliges all recipients of unemployment

insurance to draw up an individual list of their skills. In exchange, they gain access to many services to help them improve their job search strategy.

Sweden

The "Swedish model" created after World War II long combined a macroeconomic policy privileging competitiveness in international trade with a wage policy indexed to productivity growth in the sector exposed to international competition, and an active employment policy favoring mobility of labor from declining industries towards growing ones. But after the first oil shock, combating unemployment became a new objective of employment policy. The creation of temporary jobs in the public and private sectors, and subsidies for hires, then became prominent. The crisis of the 1990s, which saw the unemployment rate exceed 8% in 1996 (it had been less than 3% before 1990), caused doubts about, and even accusations against, this type of active employment policy (Calmfors, 1994; Calmfors and Lang, 1995). Since then, active policy has privileged labor market training and subsidized employment, especially for young people and the long-term unemployed. After a period of unsuccessful job search, participation in labor market programs becomes compulsory, with the aim of reducing the risk of long-term unemployment. Only a few other countries have obligatory programs for all benefit recipients who remain unemployed beyond a clearly specified period (Australia, Denmark, and the United Kingdom).

A number of major reforms have come about in Sweden since 2006. Notably, the long-standing operator of the system was replaced in 2007 by an integrated national agency charged with supervising the 300 local agencies that ensure nationwide coverage. The amount of unemployment benefit was cut back and the conditions of eligibility made more rigorous. Benefit payments were made degressive and were capped so as to favor the option of returning to work. At the same time, guarantees of activity aimed at the job seekers with the weakest ties to the labor market were reinforced. A general subsidy for employers, targeting migrants and the long-term unemployed, replaced many of the selective labor market schemes hitherto in place. For example, the measure labeled *Nystartsjobb* (literally, "New Start Jobs") offers complete relief from payroll taxes for the duration of a year to firms that hire young people under 25 who have been unemployed for six months or persons over 25 who have been unemployed for more than a year.

The United Kingdom

The Thatcher government progressively abandoned all the measures put in place by Labour governments to support demand, in favor of "supply-side" policies. So, the Job Start Allowance set up in 1986 offers a lump-sum bonus to long-term unemployed persons who agree to take low-wage jobs. But in general, active employment policy in the United Kingdom focuses on unskilled youth. The Youth Training Scheme set up in 1983 and continued in the 1990s as Youth Training provides periods of training, financed by the public authorities, for this social category. Training policies addressed to broader categories of workers are in place as well, like the Training Enterprise Councils set up in 1991, which are decentralized organizations charged with creating professional training programs under the auspices of large local firms. With the creation of Job Centers in 1987, emphasis was also placed on measures to enhance job searching. This policy direction was continued under the Labour government headed by Tony Blair, with the New Deal for Young People set up in 1998, which targets all unemployed

benefit recipients between 18 and 24 years old who have been unemployed for at least six months. It is compulsory, and begins with a period, lasting no longer than four months, of intensive job search assistance and small, basic skills courses. If the unemployed person does not find a job during this phase, the program provides several options, including the possibility of offering a subsidy to potential employers, and enrollment in a full-time training course (see the presentation of Blundell et al. [2004] in section 3.3.1 below for a detailed description of this program). The New Deal program was also extended to older workers and the disabled. In 2011 all New Deal programs were replaced by the Work Programme targeted at the most disadvantaged groups and making more intensive use of private providers for placement services.

Countries have assayed a great many policies over time and have adopted a variety of approaches to activating the unemployed. These various policies—placement, training, subsidies to the private sector, and public job creation, to name only the main strategies—influence the labor market in very different ways and might even have effects on the unemployed and indeed on employed workers not (in principle) meant to feel their effects, which are often neglected or ignored by policy makers. We now need to analyze in detail these policies to throw into relief their overall impact on employment and wages.

2 ACTIVE POLICIES: THEORETICAL ANALYSIS

If we are to (efficiently) assay the efficiency of active labor market policies, it is important to work with an equilibrium model that takes into account the combined reactions of labor demand and wages, as well as possible inefficiencies arising from the functioning of the labor market. In this regard, the search and matching model used to this point proves particularly useful, allowing us to represent a labor market which functions inefficiently for reasons that have to do with the process of job destruction and creation and the mode of wage formation. Within this framework, a positive study of employment policy is possible. It is important to note that we will be studying the consequences of active employment policies without reference to how they are financed, so throughout this section there is an implicit assumption that active policies are paid for by a lump-sum tax, one independent of income. This hypothesis is evidently unrealistic. Its only purpose is to highlight the consequences of public expenditure on employment and earnings, independently of any distortions that may arise from how the outlay is financed.

2.1 Manpower Placement Services

Manpower placement agencies, whether public or private, have a double mission. On one hand, they are charged with registering the unemployed and verifying that their clients are indeed looking for work, so that if necessary they may receive unemployment insurance. On the other, these agencies assemble offers of, and demands for, employment, and help the unemployed search for a job more effectively. The existence of such agencies is justified if, in their absence, individual decisions result in an insufficient

allocation of the resources devoted to job search. By reducing individual search costs, placement agencies can improve labor market efficiency, collecting all available information and putting it at the disposal of workers. From another perspective, the justification of the *public* character of some of these agencies must lie in imperfections inherent in the functioning of the "market" for job placements, which, for example, requires very large networks to be set up. Fixed costs for such infrastructure are very high, and congestion effects may occur. That being so, the decentralized functioning of the placement market leads to an inefficient allocation of resources. Table 14.1 shows that public agencies predominate when it comes to managing job offers; they share this role with private firms in some countries, like the United States and the United Kingdom, but tend to monopolize it in others, like France, Germany, and Sweden.

If we are to analyze placement agencies, private or public, we need to adapt our basic model laid out in chapter 9 so as to include placement activity. It will then be possible to characterize efficient outcomes and compare them with market equilibria.

2.1.1 A Matching Model with Placement Agencies

Yavas (1994) set out a formal framework for analyzing the efficiency of a labor market with placement agencies. The essential hypothesis is that an agency can ensure a *better* matchup between unemployed persons and vacant jobs than individual job searches can. This improvement in the contacting process comes at the cost of an extra drain on the resources of society (figures 14.1 and 14.2 give an order of magnitude for the amount of this cost). Fundamentally then, to set up a placement agency is to create a different kind of matching technology as an alternative to the one spontaneously available to all workers and employers. We assume that this alternative technology has increasing returns, since placement agencies generally make large outlays in order to set up a network of connections that will enable them to fill jobs at low marginal cost.

TABLE 14.1
The activity of public placement agencies at the end of the 1990s.

Country	Regulation	Vacancy registration ratio as a % of labor force	Vacancy registration ratio as a % of labor force
Australia	C	8	37
Belgium	M	—	39
Denmark	M	8	43
France	M	12	39
Japan	M	9	76
Netherlands	C	4	28
Sweden	M	11	—
United States	C	9	44
United Kingdom	C	5	—

Notes: The vacancy registration ratio equals the ratio of the job vacancies handled annually by the public agencies either to the total hirings in the economy or to the labor force. M signifies a public monopoly, and C signifies the coexistence of public and private agencies.

Source: Lippoldt and Brodsky (2004, table 7.2, p. 216).

Let us assume, for simplicity, that the labor force is of constant size, normalized to one, and let $x \in [0,1]$ be the number of unemployed persons resorting to the services of placement agencies. There is also a continuum of these agencies, indexed by $i \in [0,a]$. The agencies are assumed to be uniformly distributed, such that the mass[2] of agencies is equal to a. Let us also assume, again for simplicity, that these agencies are *instantaneously* capable of locating an entrepreneur ready to hire anyone looking for a job (which indubitably represents an improvement in the matching process). Under these conditions, we can simply denote by $c(x_i)$ the cost attached to the placement of x_i individuals by agency i. It is composed of a fixed cost $c_0(a)$ and a variable cost $c_v(x_i)$, that is, $c(x_i) = c_0(a) + c_v(x_i)$. The fixed cost $c_0(a)$ is assumed to rise with the number of agencies and satisfies $c_0(0) \geq 0$, $c_0''(a) > 0$ as well. The hypothesis that the fixed cost rises with the number of agencies gives us a simple way to take into account the congestion effects that occur in job placement. Job placement consists of creating networks so as to bring employers and workers into contact with one another, and this occasions fixed costs that probably increase when more agencies are involved. The variable cost is increasing, convex, and satisfies $c_v(0) = 0$.

Since an individual who resorts to the services of an agency finds a job immediately, only persons who undertake to look for a job on their own are described as the unemployed. We designate the number of unemployed persons by $u \in [0,1]$, and assume that the number of matches per unit of time is defined by a matching function $M(u,v)$ with the usual properties. In this expression, v again designates the number of vacant jobs, so the exit rate from unemployment is equal to $\theta m(\theta)$ with $\theta = v/u$. Let q be the exogenous job destruction rate. At stationary equilibrium the number of persons who have lost their jobs, $q(1-u)$, must be equal to the number of persons who have found jobs, $x + \theta m(\theta)u$. Hence, the mass, $x = \int_0^a x_i \mathrm{d}i$, of individuals resorting to the services of placement agencies is defined as a function of u and θ by the equality:

$$x = q(1-u) - \theta m(\theta)u \tag{14.1}$$

We should point out that this last equation also characterizes the Beveridge curve adapted to the matching model with placement agencies.

2.1.2 THE SOCIAL OPTIMUM IN THE PRESENCE OF PLACEMENT AGENCIES

In chapter 9, section 4, we saw that the social optimum is characterized very simply when the interest rate r goes to 0. Let us again place ourselves in this situation; the planner's problem then amounts to the maximization of *instantaneous* aggregate production subject to the constraint of the Beveridge curve. If, at every date, an employed individual is capable of producing an exogenous quantity y of goods, whereas an unemployed person can only make a quantity $z < y$ of these same goods "at home," instantaneous aggregate production is equal to total production $(1-u)y + uz$ from which we must deduct the total costs $hu\theta + \int_0^a c(x_i)\mathrm{d}i$ corresponding to the "natural" process of

[2] In what follows, we refer indifferently to the mass or the number of agencies.

matching and to the placements made by agencies. We thus have:

$$\omega = (1-u)y + uz - hu\theta - \int_0^a c(x_i)\mathrm{d}i \tag{14.2}$$

Equation (14.1) of the Beveridge curve allows us to eliminate the unemployment rate u from the definition (14.2) of instantaneous production, which then takes the form:

$$\omega = -\int_0^a [c_0(a) + c_v(x_i)]\,\mathrm{d}i + y - \frac{(q - \int_0^a x_i \mathrm{d}i)(y - z + h\theta)}{q + \theta m(\theta)} \tag{14.3}$$

The planner's problem consists simply of maximizing ω with respect to x_i, a, and θ. Scrutiny of the expression (14.3) of aggregate production ω shows that this problem is *dichotomic*. For all values of a and x_i, the optimal value of the labor market tightness is the solution of the problem:

$$\max_{\theta} \frac{y - z + h\theta}{q + \theta m(\theta)}$$

We thus come back to the planner's problem described in chapter 9, section 4.2. In other words, the presence of placement agencies has no influence on the optimal value of labor market tightness. This value is thus always given by equation (9.24) from chapter 9:

$$\frac{(y-z)\,[1-\eta(\theta)]}{q + \eta(\theta)\theta m(\theta)} = \frac{h}{m(\theta)} \quad \text{with} \quad \eta(\theta) = -\frac{\theta m'(\theta)}{m(\theta)} \tag{14.4}$$

For this optimal value of θ, assuming that there exists a unique interior solution[3] such that $a > 0$ and $x_i \in (0,1)$, maximization with respect to x_i and a of criterion (14.3) immediately yields:

$$c_v'(x_i) = \frac{y - z + h\theta}{q + \theta m(\theta)} = \frac{h}{[1-\eta(\theta)]\,m(\theta)}, \quad \forall\, i \in [0, a] \tag{14.5}$$

$$ac_0'(a) + c_0(a) + c_v(x_a) = x_a \frac{y - z + h\theta}{q + \theta m(\theta)} \tag{14.6}$$

Equation (14.5) indicates that it is optimal to use the services of placement agencies up to the point where the marginal cost of a placement is equal to its marginal gain. This equation thus determines the volume of placements by each agency. Equation (14.6) defines the number of agencies a. The left-hand side of (14.6) corresponds to the marginal cost of a supplementary agency, while the right-hand side represents its marginal gain. At the optimum, the two sides must be equal. Since the fixed cost c_0 rises with a, agencies are created up to the point where the cost of adding one more agency exceeds its gain.

[3] Let us recall that if $g(x) = \int_{a(x)}^{b(x)} f(x,i)\mathrm{d}i$, where f, a, and b are continuously derivable functions, then $g'(x) = b'(x)f[x,b(x)] - a'(x)f[x,a(x)] + \int_{a(x)}^{b(x)} \frac{\partial f(x,i)}{\partial x}\mathrm{d}i$.

2.1.3 Decentralized Equilibrium with Private Placement Agencies

From now on we assume that there are private placement agencies, charging for their services at price p_v for firms and price p_u for unemployed workers. So a firm can instantly fill one of its vacant jobs by paying price p_v, and an unemployed worker can instantly find a job by paying price p_u. That being the case, if a firm decides to turn to a placement agency for one of its vacant positions, it receives an expected gain equal to $\Pi_e - p_v$, where Π_e designates the expected profit from a filled job. At equilibrium, the free entry condition entails that the value Π_v of a vacant job is null, and equality $\Pi_e = p_v$ will thus always be satisfied. Symmetrically, at equilibrium, the tariff of the placement agencies will be such that the expected utility V_u of an unemployed person who does not make use of an agency's services will equal the expected utility $V_e - p_u$ of a person who has found a job immediately thanks to these services (V_e designates the expected gain from a filled job). We will thus have $p_u = V_e - V_u$. Let us assume that wage bargaining takes place in decentralized fashion, in such a way that an employee obtains fraction $\gamma \in [0, 1]$ of the global surplus $S = \Pi_e - \Pi_v + V_e - V_u$. Bearing in mind that the condition of free entry likewise dictates that the profit expected Π_e from a filled job is equal to the average cost $h/m(\theta)$ of a vacant job, and that the sharing of the surplus entails $(1 - \gamma)(V_e - V_u) = \gamma\Pi_e$, we have:

$$p_v = \frac{1-\gamma}{\gamma} p_u = \frac{h}{m(\theta)} \tag{14.7}$$

When placement agencies are in a perfectly competitive market, they do not take into account the linkage (14.7) between labor market tightness—which depends on the mass x of individuals who have resorted to placement agencies, through the medium of the Beveridge curve (14.1)—and the prices p_u and p_v. In other words, each agency considers these prices as given and determines the volume x_i of its placements in such a way as to maximize its profit $(p_u + p_v)x_i - c(x_i)$. Since relation (14.7) defining prices p_u and p_v entails $p_u + p_v = h/(1 - \gamma)m(\theta)$, this maximization arrives at a relation between x_i and θ, which takes the form:

$$c_v'(x_i) = \frac{h}{(1-\gamma)m(\theta)}, \quad \forall\, i \in [0, a] \tag{14.8}$$

Moreover, free entry into the market for placement services entails that firms are created as long as profit opportunities exist. Since the fixed cost rises with the number of agencies, at equilibrium the zero-profit condition in this market determines the number of firms a:

$$(p_u + p_v)x_i - [c_0(a) + c_v(x_i)] = 0 \Leftrightarrow c_0(a) + c_v(x_i) = x_i c_v'(x_i) \tag{14.9}$$

Since, for given θ, the presence of placement agencies does not change the wage setting on each job, the model yields a wage curve identical to the one obtained in the basic model of chapter 9. In particular, the equilibrium value of labor market tightness is given by equation (9.22) in chapter 9:

$$\frac{(1-\gamma)(y-z)}{r + q + \gamma\theta m(\theta)} = \frac{h}{m(\theta)} \tag{14.10}$$

By comparing relations (14.4) to (14.6), characterizing the social optimum, to equations (14.7) to (14.10) with $r = 0$, we see that decentralized equilibrium is not efficient, even if the Hosios condition $\gamma = \eta(\theta)$ is satisfied. This result arises from the existence of congestion effects among the placement agencies. In this economy, there is no mechanism giving placement agencies entering the market an incentive to take account of the losses they inflict on agencies already present. The upshot is that decentralized equilibrium leads to an excessive number of agencies and an overproduction of placements when the Hosios condition is satisfied. This result is easily verified by comparing equations (14.6) and (14.9). The notion that free competition in the placement agencies market leads to a situation of overproduction should nevertheless be put into perspective. Inasmuch as the size of the fixed costs attached to this type of business limits the number of firms present in this market, it is likely that monopolistic behavior in the form of restricted supply will appear.

The existence of congestion effects and the size of the fixed costs attached to the job placement business suggest that decentralized equilibrium probably leads to an inefficient allocation characterized by states of under- or overproduction. This inefficiency, and the need to check on the search effort being made by those receiving unemployment benefits, generally justify state intervention in the job placement market. But this intervention must itself be efficient. The empirical research on this problem is presented in section 3.2 below.

Placement agencies are not the only type of search method used by the unemployed. There are other approaches, such as contacting relatives, family, family friends, and ex-colleagues (see also chapter 5, section 1.3). These social networks may be less effective but are also less costly than recourse to placement agencies. Cahuc and Fontaine (2009) show that decentralized decisions to use social networks in the job search process can also be inefficient, inasmuch as social networks may be overutilized with respect to an efficient allocation, with the consequence that congestion effects are intensified. This happens for instance when formal and more costly job search methods are underutilized. Moreover, the existence of different job search methods can give rise to a higher job search intensity than the efficient one. Hence, recourse to placement agencies may help steer the labor market away from inefficient equilibria.

2.2 Why Promote Training?

A large portion of the money spent on labor market policy goes to promote training. Leaving aside the question of how they are financed, these measures have the capacity to increase employment by raising labor productivity. Nonetheless, public intervention is only justified if individual decisions lead to levels of training inadequate with respect to what would be socially desirable. We saw in chapter 4 that in a perfectly competitive economy, where it is possible to sign complete contracts, individual training decisions are socially efficient. It would be difficult to justify the need for public intervention in such a setting.

Individual decisions about training are no longer necessarily efficient, though, when competition is imperfect. Imperfection in competition may arise from many sources, which create distortions and give private agents an incentive to take inefficient decisions. We pointed out, in chapter 4, that the unobservability of the characteristics of employees drives them, in certain circumstannces, to overeducate themselves in

order to signal their quality to employers. In many cases, imperfect competition is also revealed by too low a level of investment in education. For example, the imperfection of the credit market may block access to training that would pay off, both individually and socially, and so impede individuals with few resources from acquiring some kinds of training (see Becker, 1964).

In this section, we concentrate on the consequences of imperfections in the labor market as regards education. In particular, we demonstrate, on the basis of the work of Acemoglu (1997), Acemoglu and Pischke (1998, 1999a, 1999b), and Stevens (1994), that the existence of *transaction costs* in the labor market generally leads to *underinvestment* in training when state intervention plays no part. Such underinvestment reduces productivity and proves harmful to employment.

To examine decisions about training, it is best to adopt the distinction introduced by Becker (1964) between *general training*, which enhances the productivity of the individual concerned for all types of jobs, and *specific training*, which only enhances her productivity for one type of job. This distinction is clearly theoretical, to the extent that all training has a certain degree of specificity, but it is analytically useful. General training is fundamentally associated with the worker who can apply it in different types of jobs and so bring employers to compete for her services. The structure of competition between employers is thus capable of affecting decisions about training which potentially concern a multitude of individuals. Specific training, on the other hand, is associated with a match between a particular worker and a particular employer, and the payoff it brings depends only on the relations between these two agents.

We begin by studying the problems linked to general training, showing that the length of time matching takes, and the costs it incurs, are sources of underinvestment. We then study specific training, emphasizing that the difficulty of signing complete contracts is the source of underinvestment for this type of training.

2.2.1 Acquiring General Training

Decisions about general training in a perfectly competitive economy were presented in chapter 4. According to the standard analysis of Becker (1964), in that context investment in general training is entirely financed by workers. Moreover, the level of investment chosen corresponds to a social optimum. The costs of achieving matches and the monopsony power of employers, however, entail an underinvestment in general training with respect to the socially desirable situation (Stevens, 1994; Acemoglu, 1997; Acemoglu and Pischke, 1998, 1999a, 1999b). This we will demonstrate, initially by integrating investment in general training into the matching model of chapter 9 and then by proceeding to characterize the social optimum of this economy and compare it with decentralized equilibrium.

The Labor Market with Matching Costs and Investment in General Training

In order to represent decisions to invest in general training in the presence of matching costs without too much difficulty, we assume that a person entering the labor market possesses no training of this kind at the outset. At the time she finds her *first* employer, she decides to invest an amount i in general training. For simplicity, the duration of training is assumed to be null. Once trained, each worker is capable of producing

quantity $y(i)$ of goods at every future instant. In other words, workers never need to be retrained. As workers are always assumed to have infinite lifetimes, this property obliges us to consider that the labor force is always growing, for if it were not, everyone would have acquired the necessary general training at the end of some greater or lesser period of time, and at the stationary state, the optimal level of investment would be zero. Thus, we assume that the labor force increases at the constant exogenous rate $n > 0$ and that all the new entrants into the labor market are unemployed persons, who by hypothesis have no general training. They find themselves in competition with older unemployed persons, who have the general training they got when they were first hired.

As in the preceding sections and in chapter 9, the imperfection of the process by which firms and workers match up is summarized by a matching function possessing the usual properties. The exit rate from unemployment is then equal to $\theta m(\theta)$, where labor market tightness θ represents the ratio V/U between the stock of vacant jobs and the stock of unemployed persons. In what follows, we omit, with no risk of confusion, the time index, and we denote by U_f, U_n, and N the number of trained unemployed persons, the number of unemployed persons with no training, and the size of the labor force at any date. We then have $U = U_f + U_n$. The unemployed, trained or not, have the same probability of exiting from unemployment, for employers are incapable of telling them apart, before meeting them. We use $u_f \equiv U_f/N$ and $u_n \equiv U_n/N$ to designate the number of unemployed in each of these categories with respect to the labor force, and $u \equiv U/N$ to designate the unemployment rate. At every instant, the stock of unemployed persons without training increases by nN units but loses $\theta m(\theta) U_n$ individuals who find jobs. The instantaneous variation $\dot{U}_n$ in the number of untrained unemployed is thus defined by the equality $\dot{U}_n = nN - \theta m(\theta) U_n$. Since $\dot{U}_n \equiv nNu_n + N\dot{u}_n$, the law of motion of u_n is:

$$\dot{u}_n = n - [n + \theta m(\theta)]\, u_n \tag{14.11}$$

From that we deduce the stationary level of unemployed persons for this category:

$$u_n = \frac{n}{n + \theta m(\theta)} \tag{14.12}$$

Let us further assume that the job destruction rate q is an exogenous constant; the instantaneous variation $\dot{U}$ in the total stock of unemployed persons is equal to the difference between the number of persons who at every instant become unemployed, $qN(1-u) + nN$, and the number $\theta m(\theta) U$ of persons who find a job. Since $\dot{U} \equiv nNu + N\dot{u}$, the time path of the unemployment rate is given by:

$$\dot{u} = q + n - [q + n + \theta m(\theta)]\, u \tag{14.13}$$

The stationary unemployment rate is then written:

$$u = \frac{q + n}{q + n + \theta m(\theta)} \tag{14.14}$$

We are back to the equation of the Beveridge curve, which defines a decreasing relation between the unemployment rate and the rate of vacant jobs.

The Social Optimum

In chapter 9, section 4, we saw that if we assume that all agents are risk neutral, the social optimum is found by maximizing the present discounted value of net aggregate output, taking into account the dynamics of the variables that enter into this discounted value. With the notations employed to this point, net instantaneous aggregate output Ω is defined as follows:

$$\Omega = N(1-u)y + zU - hV - \theta m(\theta)U_n i \tag{14.15}$$

In this formulation, the variable y represents the average production per employed worker, which must formally be distinguished from the production $y(i)$ realized by a person who has benefited from an investment i at the current date, precisely because the production of employed workers depends exclusively on investments in general training made in the past. It should also be noted that the training costs $\theta m(\theta)U_n i$ of the untrained unemployed who find a job form part of Ω. Let $Y = N(1-u)y$ be the instantaneous gross production of employees. This variable increases at each instant by the production $\theta m(\theta)U_f y$ of trained unemployed persons who find a job and the production $\theta m(\theta)U_n y(i)$ of unemployed persons trained at the current date because they have just found their first job. Taking into account the losses due to the destruction of jobs, the instantaneous variation in gross aggregate output is defined by $\dot{Y} = \theta m(\theta)[U_f y + U_n y(i)] - qY$. Since by definition $\dot{Y} \equiv (1-u)(n y + N\dot{y}) - N\dot{u}y$ and $u \equiv u_n + u_f$, relation (14.13) allows us, after several easy calculations, to arrive at an equation describing the law of motion of average production per employed person. It comes to:

$$\dot{y} = \frac{\theta m(\theta) u_n}{1-u}\,[y(i) - y] \tag{14.16}$$

At any instant t, the size N of the labor force is equal to $N_0 e^{nt}$ where N_0 designates the exogenous size of this population at date $t = 0$. With the help of expression (14.15) of instantaneous net aggregate output, the planner's problem takes the following form:

$$\max_{\theta,\, i} \int_0^{+\infty} [(1-u)y + (z - \theta h)u - \theta m(\theta) u_n i]\, e^{-(r-n)t} \mathrm{d}t \quad \text{s.c. (14.11), (14.13), and (14.16)}$$

Socially Efficient Investment

Let λ, μ, and ν be the multipliers respectively linked to constraints (14.11), (14.13), and (14.16); the Hamiltonian of the planner's problem is written:[4]

$$H = [(1-u)y + (z - \theta h)u - \theta m(\theta) u_n i]\, e^{-(r-n)t}\mathrm{d}t + \lambda \dot{y} + \mu \dot{u} + \nu \dot{u}_n$$

The first-order conditions are given by the equations:

$$\frac{\partial H}{\partial i} = 0, \quad \frac{\partial H}{\partial \theta} = 0 \quad \text{and} \quad \frac{\partial H}{\partial y} = -\dot{\lambda}, \quad \frac{\partial H}{\partial u} = -\dot{\mu}, \quad \frac{\partial H}{\partial u_n} = -\dot{\nu} \tag{14.17}$$

[4] The principles of dynamic optimization are set out in mathematical appendix B at the end of the book.

Differentiating the Hamiltonian with respect to i, the first of the conditions (14.17) immediately entails:

$$\lambda = \frac{(1-u)e^{-(r-n)t}}{y'(i)} \tag{14.18}$$

Differentiating the Hamiltonian now with respect to y, condition $\partial H/\partial y = -\dot{\lambda}$ brings us to:

$$(1-u)e^{-(r-n)t} - \lambda \frac{\theta m(\theta) u_n}{1-u} = -\dot{\lambda} \tag{14.19}$$

Henceforth we are at stationary equilibrium where $\dot{\theta} = \dot{u} = 0$; differentiating relation (14.18) with respect to t gives $\dot{\lambda} = -(r-n)\lambda$. Bringing this value of $\dot{\lambda}$ into (14.19), we deduce the value of the multiplier λ. Equation (14.18) then yields $y'(i)$ as a function of u, u_n, and θ. Using definitions (14.12) and (14.14) of the unemployment rates at stationary equilibrium, we can express $y'(i)$ as a function of the variable θ alone. It comes to:

$$y'(i^*) = r + \frac{nq}{n + \theta m(\theta)} \tag{14.20}$$

This equation completely characterizes the level of efficient investment i^* for any value of labor market tightness θ. For given θ, integrating differential equations (14.11) and (14.13) does indeed allow us to express the unemployment rates u_f and u_n as a function of the variable θ alone. There is then no more need to take constraints (14.11) and (14.13) into account in the planner's problem. Since relation (14.20) was only obtained on the basis of conditions $\partial H/\partial i = 0$ and $\partial H/\partial y = -\dot{\lambda}$, it is thus satisfied for any given value of θ. Note that we find the level corresponding to perfect competition, that is, $y'(i) = r$, when $\theta m(\theta)$ goes to $+\infty$, when it is possible for a person who has lost her job to be rehired immediately.

Decentralized Equilibrium

We will now establish that decentralized equilibrium is characterized by underinvestment in general training even if firms and workers are capable of entering into complete contracts (this result was obtained by Acemoglu, 1997). It is assumed that a complete contract is negotiated when a match occurs and is not renegotiable later. In chapter 9, section 5, we showed that investment decisions in the presence of complete contracts lead to the maximization of the surplus net of investment costs. The level of the wage negotiated depends on the share of the surplus obtained by each party and the amounts they each invest.

By definition, the surplus from a match that takes place with a worker who has not yet acquired any general training is equal to the sum of the expected profit $\Pi_e(i)$ and the expected utility $V_e(i)$, reduced by the value Π_v of a vacant job, and of the expected gains V_u of an untrained unemployed person, where i designates the level of investment made in the job in question. When an untrained worker is hired, the optimal investment maximizes the net surplus. When the free entry condition $\Pi_v = 0$ is satisfied, the net surplus reads:

$$S_n(i) = V_e(i) - V_u + \Pi_e(i) - i \tag{14.21}$$

Let us denote respectively by i_e and i_f, with $i_e + i_f = i$, the amount of investment made by the employee and the firm, and let us assume that a part γ of the net surplus goes to the worker; the negotiated wage is implicitly determined by the surplus sharing rules:

$$V_e(i) - i_e - V_u = \gamma S_n(i) \quad \text{and} \quad \Pi_e(i) - i_f = (1-\gamma) S_n(i)$$

These equations indicate that the wage of workers without initial training depends not just on the amount of total investment i but also on their personal contribution to this investment. For a given amount of investment i, the wage negotiated is evidently lower, the smaller the worker's contribution is. We simply denote this wage by w.

It is important to point out that the expected utility of a trained worker, should she lose her current job, depends on her training, since in negotiating with potential employers, she can make her productive abilities, equal to $y(i)$, pay off. Consequently, we denote by $V_u(i)$ the gains expected by an unemployed person who has had the benefit of an investment in general training amounting to i. The expected gains are then defined by the usual equations:

$$rV_e(i) = w + q\left[V_u(i) - V_e(i)\right] \tag{14.22}$$

$$r\Pi_e(i) = y(i) - w + q\left[\Pi_v - \Pi_e(i)\right] \tag{14.23}$$

Let $\bar{V}_e(i)$ and $\bar{w}(i)$ be respectively the expected utility and the wage of an employee hired when she was already trained (for whom the investment i in general training was thus made on a previous job); we then have:

$$rV_u(i) = z + \theta m(\theta)\left[\bar{V}_e(i) - V_u(i)\right] \quad \text{and} \quad r\bar{V}_e(i) = \bar{w}(i) + q\left[V_u(i) - \bar{V}_e(i)\right] \tag{14.24}$$

For trained workers, bargaining covers only the wage level $\bar{w}(i)$, since it is no longer necessary to invest in their general training. At this stage, the model becomes identical to the basic model of chapter 9 and the outcome of the negotiation is described by equation (9.21):

$$\bar{w}(i) = z + [y(i) - z]\,\Gamma(\theta) \quad \text{with} \quad \Gamma(\theta) = \frac{\gamma\,[r + q + \theta m(\theta)]}{r + q + \gamma\theta m(\theta)}$$

Relations (14.24) then allow us to express $V_u(i)$ as a function of i and θ; it comes to:

$$rV_u(i) = z + [y(i) - z]\,\frac{\gamma\theta m(\theta)}{r + q + \gamma\theta m(\theta)} \tag{14.25}$$

This formula indicates how the investment i in general training made today increases the expectation of future gain of a worker in search of a job. It should be taken into account at the time of choosing the amount of optimal investment. Taking relations

(14.22) and (14.23) into account, when the free entry condition $\Pi_v = 0$ is satisfied, the surplus net of investment costs (14.21) is written:

$$S_n(i) = \frac{y(i) + qV_u(i)}{r+q} - i - V_u \tag{14.26}$$

With the help of definition (14.25) of $V_u(i)$, the maximization of the net surplus gives an investment i_m defined by:

$$y'(i_m) = r + \frac{rq}{r + \gamma\theta m(\theta)} \tag{14.27}$$

Setting aside the case of perfect competition (which is obtained by making $\theta m(\theta)$ go to $+\infty$), comparison of this relation with equation (14.20) characterizing the socially efficient level of investment i^* shows that if $r > n$ then $y'(i_m) > y'(i^*)$ for all values of θ. The concavity of function $y(.)$ then entails $i^* > i_m$. In an imperfectly competitive labor market, there is thus a tendency to underinvest in general training even if agents can sign complete contracts.[5] That derives from the fact that a part of the investment decided upon by a worker and an employer will necessarily benefit future employers, who are not parties to the investment decision.

Underinvestment and Incomplete Markets

We have just seen that agents underinvest in general training because it is not possible for them to negotiate with *future* employers. The latter will benefit from the investment made today, for in an imperfectly competitive market they will capture a part of the surplus produced by workers. This positive externality is not taken into account by the market, and this in turn justifies state intervention in the area of general training (on these questions, see Acemoglu, 1997, and Acemoglu and Pischke, 1998, 1999a, 1999b). We note that if decentralized equilibrium with complete contracts is inefficient, it is so a fortiori with incomplete contracts.

There are many other sources of externality associated with training decisions. Most often the acquisition of human capital by an agent represents a positive externality for her immediate circle without these benefits being acknowledged through any remuneration. The transmission of know-how through simple discussions, or by observation, are classic examples of such externalities. Individual training has social consequences on which the market does not necessarily place a value. Many studies have shown that the performance of students is influenced by the average level of performance of the students in their vicinity (Coleman et al., 1966; Ioannides and Topa, 2010; Pattachini and Zenou, 2011). These externalities play a very important role in models of endogenous growth (Lucas, 1988; Benabou, 1996; Aghion and Howitt, 1998).

Formally, these direct externalities can be brought into account in the model developed above by taking the view that a worker's productivity is an increasing function of her own investment i and of the average level of investment $\bar{i}$ of all workers.

[5] It is easy to verify that employers have no interest in reinvesting in workers who are already trained. If they did, they would maximize a net surplus defined by $\frac{y(i)+qV_u(i)}{r+q} - i - V_u(i)$, which necessarily gives a level of investment inferior to i_m, since $V_u'(i) > 0$.

Individual production is then represented by the concave function $y(i, \bar{i})$. If we go back to the model with this formulation, the possibility arises of a multiplicity of market equilibria when a rise in average investment improves the marginal return to individual investment (that is, if the second derivative y_{12} is positive). In the terminology of Cooper and John (1988), the decisions of agents are then characterized by "strategic complementarities" capable of causing coordination failures and holding the market at a low level of investment.

Complex contracts obliging possible future employers to pay a transfer to the initial employer or to pay a wage supplement to previously trained workers would in theory allow the social optimum to be reached (Acemoglu, 1997). But this contractual structure is not realistic because for it to be put into practice there would have to be commitments binding *all* employers, something very hard to envisage. Snower (1995), Ulph (1995), and Acemoglu (1997) have also shown that firms might be given an incentive to choose technologies using mainly low-skilled manpower, if workers have little training. Such behavior by firms would accentuate underinvestment in general training, since the incentive for workers to invest in this type of training increases with the demand for skilled labor.

The imperfection of the financial markets is another barrier to investment in general training. When wage earners are obliged to borrow in order to get training, the difficulties of access to credit do indeed lead to an insufficient level of training. The imperfection of financial markets most often arises from an asymmetry of information between the organizations granting credit and the investors. Uncertainty about the capacities of individuals applying for credit, and the chance that they might use the money for purposes other than training, constitute sources of inefficiency in the credit market which must lead to rationing. Becker (1964) emphasizes that this type of problem ought to be solved by public intervention in the credit market instead of by regulating the general training of workers. Thus underinvestment in training does not always necessitate subsidies or action by the state in this area.

The imperfect information employers have about the characteristics of workers is another potential source of underinvestment in general training. If employers observe the amount invested in human capital, and the return on it, imperfectly, then workers are at risk of not being able to make their training pay off fully, which leads them to invest less. So employers have an interest in completing general training after hiring (Katz and Ziderman, 1990; Chang and Wang, 1996). In that case, investment by firms will be optimal if it is possible to sign complete, non-renegotiable contracts.

2.2.2 ACQUIRING SPECIFIC TRAINING

Unlike general training, specific training demands a new investment every time a worker changes firms. In that context, the incompleteness of the labor contract becomes the principal source of inefficiency in decentralized decisions. We prove this point, beginning with a definition of the social optimum in the presence of transaction costs in the labor market and costs of specific training. We then show that decentralized equilibrium coincides with the social optimum when there are complete contracts. This result is thus different from that obtained within the framework of general training, where the costs of matching constitute a source of inefficiency in decentralized decisions. Conversely,

when labor contracts are incomplete, decentralized decisions entail underinvestment with respect to the socially desirable level.

The Social Optimum with Specific Training

With no risk of confusion, we again denote by i the investment in specific training from which a worker benefits at each new hire. Once this investment is made, the employee is capable of producing a quantity $y(i)$ of goods *solely* in the firm she has just joined. The function $y(i)$ possesses the same properties as before: it is increasing, concave, and such that $y(0) > z$. Formally, the analysis of the social optimum with specific training is deduced from that with general training, with these addenda: an unemployed person never possesses specific training, and an investment i must be made in every unemployed person when she finds a job. In other words, from now on we have $u_f \equiv 0$ and $u_n \equiv u$. Relations (14.13) and (14.14) describing the law of motion of the unemployment rate u and the stationary value of this variable apply here as well. On the other hand, we must replace u_n by u in equation (14.16) characterizing the evolution of average production y per employed person. Thus we will now have:

$$\dot{y} = \frac{\theta m(\theta) u}{1-u}\,[y(i) - y] \tag{14.28}$$

The planner's problem is then written as follows:

$$\max_{\theta, i} \int_0^{+\infty} [(1-u)y + (z - \theta h)u - \theta m(\theta) u i]\, e^{-(r-n)t} \mathrm{d}t \quad \text{s.c. (14.13) and (14.28)}$$

Let λ and μ again denote the multipliers respectively associated with constraints (14.28) and (14.13); the Hamiltonian of the planner's problem takes the form:[6]

$$H = [(1-u)y + (z - \theta h)u - \theta m(\theta) u i]\, e^{-(r-n)t} + \lambda \dot{y} + \mu \dot{u}$$

The first-order conditions are given by equations:

$$\frac{\partial H}{\partial i} = 0, \quad \frac{\partial H}{\partial \theta} = 0 \quad \text{and} \quad \frac{\partial H}{\partial y} = -\dot{\lambda}, \quad \frac{\partial H}{\partial u} = -\dot{\mu} \tag{14.29}$$

Differentiating the Hamiltonian with respect to i, the first of conditions (14.29) again brings us to the equality (14.18) giving the value of the multiplier λ as a function of u and of i. If we now derive the Hamiltonian with respect to y, condition $\partial H/\partial y = -\dot{\lambda}$ entails:

$$(1-u)e^{-(r-n)t} - \lambda \frac{\theta m(\theta) u}{1-u} = -\dot{\lambda} \tag{14.30}$$

At stationary equilibrium where $\dot{\theta} = \dot{u} = 0$, the derivation of relation (14.18) with respect to t gives $\dot{\lambda} = -(r-n)\lambda$. Bringing this value of $\dot{\lambda}$ into (14.30), we deduce from

[6] The technique of dynamic optimization is set out in mathematical appendix B at the end of the book.

that the value of the multiplier λ. Equation (14.18) then yields $y'(i)$ as a function of u and θ. Using definition (14.14) of the unemployment rate at stationary equilibrium, we can then express $y'(i)$ as a function of θ alone. The socially optimal level of investment in specific training, again denoted i^*, thus satisfies:

$$y'(i^*) = r + q \tag{14.31}$$

It should be pointed out that efficient investment in specific training depends neither on the matching process nor on labor market tightness θ. These properties are highly intuitive, for the investment in specific training is only made after the matchup between a worker and a firm, and this investment has to be made again at each new matchup. The time spent searching for a job thus plays no part in the decision to invest in specific training.

Equilibrium with Complete Contracts and Specific Training

Contrary to the result we reached in the case of general training, here we will show that decentralized equilibrium selects a socially optimal amount of investment in specific training when firms and workers are capable of committing themselves to complete contracts. Formally, the only difference from the case of general training lies in the independence of the expected utility of any unemployed person when an investment in specific training is made, which means that it is enough to set $V_u(i) = V_u$ in the decentralized market model with general training in order to find equilibrium with specific training. Therefore, setting $V'_u(i) = 0$ in the expression (14.26) of the surplus from a filled job, we see that the equilibrium value, again denoted i_m, of the global investment in specific training satisfies the equality $y'(i_m) = r + q$. In a decentralized equilibrium, the investment in specific training is thus socially optimal. The absence of externalities arising from specific training ensures that the privately chosen investment is socially efficient. Note that to arrive at this result, it is not necessary to specify the exact form of V_u, nor to refer to the matching process that takes place in the labor market. The efficiency of decentralized equilibrium when it comes to investment in specific training is thus a property that is satisfied with or without labor market frictions. The reason for this is the same as the one adduced for the determination of efficient investment i^*: the time spent searching for a job plays no part in the decision to invest in specific training.

The hypothesis that there is commitment to complete contracts renders the participation of agents in financing the investment inconsequential. As in the case of general training, to the extent that there are binding commitments, the parties agree to compensate changes in workers' share of investment in training by changes in the wage. In what follows, we show that this compensation does not operate if contracts are incomplete.

Equilibrium with Incomplete Contracts and Specific Training

A necessary condition (but not always a sufficient one; see the case of general training) of the efficiency of investment decisions is that it must be possible to sign long-term, nonrenegotiable contracts in such a way as to avoid the holdup problem. But it is impossible under many circumstances to have the clauses of a contract verified by a third party (see chapter 6), and this leads to the adoption of incomplete contracts—ones that are vulnerable to renegotiation. That being so, there is a risk of underinvestment. This situation is

illustrated for physical capital in chapter 7, section 3.3, and investment in training is no different.

This will emerge clearly if we go back to the previous model: but now we assume that each party decides, at the time of hiring, how much to contribute to the investment in specific training, knowing that the wage might be renegotiated at any time. It is easiest to represent this situation by a two-stage game. In the first stage, the employer and the worker choose, simultaneously and without cooperation, their respective specific investments i_f and i_e. Total investment $(i_f + i_e)$ is always denoted by i. In the second stage, the wage is negotiated in such a way as to share the surplus in accordance with the bargaining power of each of the agents. The outcome of this game is found by backward induction.

The expected utility of an employee and the expected profit from a filled job are again given by relations (14.22) and (14.23) on condition that we replace $V_u(i)$ by V_u in (14.22). In the second stage of the game, the gains of the employer and the worker are respectively equal to $\left[\Pi_e(i) - i_f\right]$ and $[V_e(i) - i_e]$ if the bargaining is successful. But if the bargaining fails, the respective gains amount to $(\Pi_v - i_f)$ and $(V_u - i_e)$ since at this stage the investment has already been made. So the surplus released by a match is equal to:

$$S(i) = V_e(i) - V_u + \Pi_e(i) - \Pi_v = \frac{y(i) - rV_u}{r + q} \tag{14.32}$$

The wage bargaining that takes place at this stage shares out the surplus in accordance with the bargaining power of each of the agents. Since V_u does not depend on i, relations (14.22), (14.23), and (14.32) defining the gains of agents and the surplus show that this stage of the game is formally identical to wage bargaining in the basic model from chapter 9. We thus have:

$$w = \gamma y(i) + (1 - \gamma) rV_u \tag{14.33}$$

In the first stage of the game, the employer determines the amount i_f of his investment by maximizing his net profit $\Pi_e(i) - i_f$. He then knows the reaction of the negotiated wage described by equality (14.33) and considers the investment i_e of the employee as given. So with the help of the definition of $\Pi_e(i)$ given by (14.23), we arrive at:

$$(1 - \gamma) y'(i) = r + q \tag{14.34}$$

Symmetrically, the worker knows the reaction of the wage and decides her investment i_e by maximizing her net gain $V_e(i) - i_e$ with given i_f. The definition of $V_e(i)$ given by (14.22) then entails:

$$\gamma y'(i) = r + q \tag{14.35}$$

Relation (14.34) describing the best response from the employer indicates that he announces a *global* amount of desired investment, denoted $\widetilde{i}$ and defined by the equality $(1 - \gamma) y'(\widetilde{i}) = r + q$. Relation (14.35) likewise shows that the employee desires a global amount of investment, denoted $\widehat{i}$, such that $\gamma y'(\widehat{i}) = r + q$. In a noncooperative equilibrium, the agent with the highest level of desired investment will assume the entire

cost of the investment. Consequently, if $\gamma > 1/2$, $\widehat{i}$ is superior to $\widetilde{i}$ and only the worker invests in her own specific training. At market equilibrium, this investment amounts to $\widehat{i}$. Relation (14.31) giving the value i^* of the socially efficient investment then shows that $\widehat{i} \leq i^*$, with $\widehat{i} = i^*$ if $\gamma = 1$. On the other hand, if $\gamma < 1/2$, the employer assumes the entire burden of the investment, which then comes to $\widetilde{i}$. Relation (14.31) again shows that we have $\widetilde{i} \leq i^*$, with $\widetilde{i} = i^*$ if $\gamma = 0$. If $\gamma = 1/2$, there is a range of equilibria, all of them inefficient. Hence market equilibrium leads to underinvestment in specific training, except when one of the agents has all the bargaining power. In that situation, the fact that no commitment can be made no longer matters, for the agent with all the power is also the only one to benefit from the payback on the investment; this explains why she invests in an efficient fashion.

We have just shown that transaction costs in the labor market constitute sources of underinvestment in training, both specific and general. This justifies state intervention in this area in order to upgrade all levels of training. The intervention itself has to be adequately efficient as well. Many empirical studies have been dedicated to this problem, and the results are brought together in sections 3 and 4 below.

2.3 Employment Subsidies and the Creation of Public-Sector Jobs

When the matching process is imperfect, social efficiency requires strictly positive unemployment so that vacant jobs can be filled. To try to get rid of unemployment by creating a great many vacant jobs would be a waste of resources. Nevertheless, there are a number of reasons why an excessively high unemployment rate may occur at market equilibrium. When that happens, employment subsidies and the creation of public jobs may constitute means to reduce the unemployment rate while improving overall welfare.

Nevertheless, policies of this kind have nontrivial effects on labor market equilibrium. To grasp their impact, we must bring into account the interaction between the behavior of firms and the behavior of workers. As we shall see, such interaction may generate obstacles to the achievement of the intended policy goals.

2.3.1 Employment Subsidies

The main limitation on the efficiency of employment subsidies lies in the upward pressure they exert on wages, which has a tendency to bid up the cost of labor and reduce labor demand. This phenomenon emerges clearly in the case of a perfectly competitive labor market as seen in detail in chapter 3, section 1.2, and illustrated in figure 14.6. An increase in labor demand on account of a fall in the cost of labor increases wages. These increases are greater, the less the wage elasticity of the labor supply. At the limit, if the wage elasticity of the labor supply is null, the shift in labor demand simply leads to a wage rise, with no impact on employment. In practice, as shown in chapter 3, employment subsidies generally induce significant wage increases inasmuch as the elasticity of the labor supply is relatively small, which in turn reduces the employment effect of wage subsidies. This mechanism implies that employment subsidies are more efficient when there is a binding minimum wage. In that case, the employment impact of employment subsidies is entirely determined by their effect on labor demand.

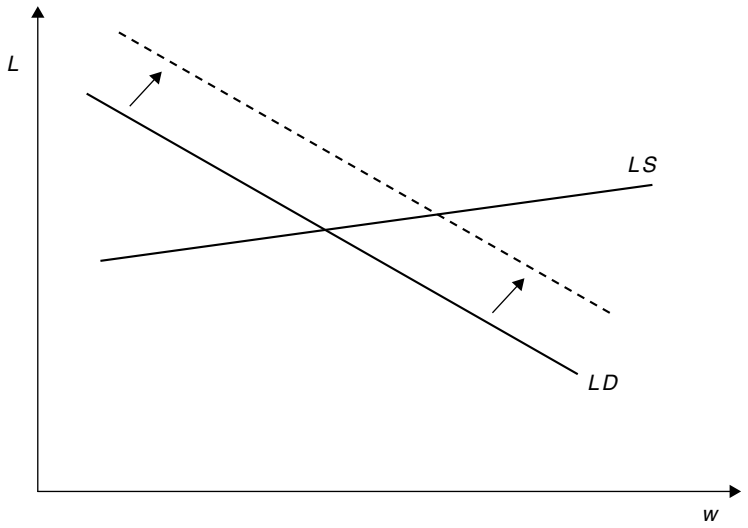


FIGURE 14.6
The effect of employment subsidies in the competitive model.

The search and matching model arrives at similar conclusions. It shows that the employment impact of employment subsidies depends on the response of wages. In the search and matching model with taxes (chapter 12), we concluded that increases in average tax rates (keeping the marginal rate constant) decrease employment, increase unemployment, and decrease wages. We also concluded that changes in average tax rates (keeping marginal rates constant) have no effect on the unemployment rate if the income of unemployed workers varies by a same amount as that of employees (see chapter 12, section 1.2.2, equation [12.13]). These results obviously apply to employment subsidies, which are equivalent to decreases in tax rates.

2.3.2 THE CROWDING-OUT EFFECTS OF PUBLIC-SECTOR JOBS

In comparison with employment subsidies, the creation of public-sector jobs presents the advantage of making it possible actually to create jobs within a short timeframe. For this reason it is often adopted either as a remedy for unemployment or as a springboard to regular jobs for persons who have difficulty entering the labor force. The creation of public-sector jobs poses problems of two kinds. First, it is not certain that jobs of this kind do significantly boost one's chances of obtaining a job in the unsubsidized private sector after the public-sector program terminates. This will emerge clearly when we review the empirical research. Second, the creation of public jobs is liable to crowd out private-sector ones through the same mechanism as employment subsidies: the increase in labor demand provokes a wage rise that may, over time, completely cancel out the impact of the public-sector jobs created, if the labor supply is insensitive to wages (Calmfors, 1994; Calmfors and Lang, 1995; Algan et al., 2002). We will begin by looking at the crowding-out effect induced by the creation of public-sector employment in the matching model, before proceeding to a quantitative assessment of the extent of this effect.

It is possible to represent the impact of the creation of public-sector jobs schematically, using the matching model and assuming that these jobs have the same characteristics as those in the private sector (less rudimentary models may be studied in Holmlund and Linden, 1993; Calmfors and Lang, 1995; and Algan et al., 2002).

The Beveridge Curve with Public-Sector Jobs

By hypothesis, workers in the private and public sectors receive the same wage w and face the same probability q of losing their jobs. The assumption is that the state aligns civil service wages with those negotiated in the private sector. For the sake of simplicity, the size of the labor force is assumed to be constant, equal to 1; we denote public-sector employment by L_g. If L designates employment in the private sector, the unemployment rate u is defined by the equality:

$$u = 1 - L_g - L$$

We assume that the matching process in the public sector is perfectly efficient. The state recruits its employees by a random draw from among all the unemployed. Let g be the rate at which an unemployed person is hired in the public sector. At stationary equilibrium, the volume of jobs destroyed per unit of time in this sector, qL_g, must equal the volume gu of jobs created. Hence rate g depends on the unemployment rate, the job destruction rate, and the volume of public-sector jobs, according to the formula:

$$g = \frac{qL_g}{u} \tag{14.36}$$

Assuming that the usual matching process goes on in the private sector, at every instant there are $[g + \theta m(\theta)]u$ jobs created and $q(1 - u)$ jobs destroyed in the economy as a whole. At stationary equilibrium, these two quantities are equal, and using definition (14.36) of g, the unemployment rate is expressed as follows:

$$u = \frac{q(1 - L_g)}{q + \theta m(\theta)} \tag{14.37}$$

This equation defines the Beveridge curve in the presence of a public sector of size L_g. It turns out that the creation of public-sector jobs reduces the unemployment rate when the vacancy rate in the private sector is given. But the number of vacancies is an endogenous variable determined by the profit outlook of firms, so we must focus on the determinants of labor demand and negotiated wages to understand the impact of public employment on unemployment.

Labor Market Equilibrium

Wages and the job destruction rate being identical in both sectors, an employee has the same expected utility V_e everywhere. Since an unemployed person finds a job in the public and private sectors at respective rates g and $\theta m(\theta)$, his expected utility V_u satisfies the relation:

$$rV_u = z + [g + \theta m(\theta)]\,(V_e - V_u) \tag{14.38}$$

Comparing this relation with the definition of V_u in the basic matching model of chapter 9, it turns out that this matching model with public-sector employment is formally equivalent to the basic model, on condition that we replace the probability $\theta m(\theta)$ of returning to employment by the sum $g + \theta m(\theta)$. Consequently, the negotiated wage is written as follows:

$$w = z + \Gamma(\theta, g)(y - z) \quad \text{with} \quad \Gamma(\theta, g) = \frac{\gamma\,[r + q + g + \theta m(\theta)]}{r + q + \gamma[g + \theta m(\theta)]} \tag{14.39}$$

It is, moreover, possible to eliminate the unemployment rate u between relations (14.36) and (14.37), which allows us to write g as a function of L_g and θ. We thus get $g = L_g[q + \theta m(\theta)]/(1 - L_g)$. Bringing this value of g into the wage equation (14.39), we find the remuneration of an employee as a function of labor market tightness θ and the level L_g of public-sector employment, that is:

$$\begin{gathered} w = z + \widehat{\Gamma}(\theta, L_g)(y - z) \\ \text{with} \quad \widehat{\Gamma}(\theta, L_g) = \frac{\gamma\,[r(1 - L_g) + q + \theta m(\theta)]}{r + q + \gamma\theta m(\theta) - L_g\,[r + q(1 - \gamma)]} \end{gathered} \tag{14.40}$$

In the (θ, w) plane, labor market equilibrium lies at the intersection of the wage curve (WC), represented by equation (14.40), with labor demand. The latter arises from the equality between the average cost $h/m(\theta)$ of a vacant job and the expected profit $(y - w)/(r + q)$ from a filled job, so it does not depend on the size L_g of the public sector. On the other hand, it is easy to verify that for given θ, the negotiated wage rises with L_g. In the (θ, w) plane, the wage curve shifts to the right. Labor market equilibrium is represented in figure 14.7. It turns out that by increasing the exit rate from

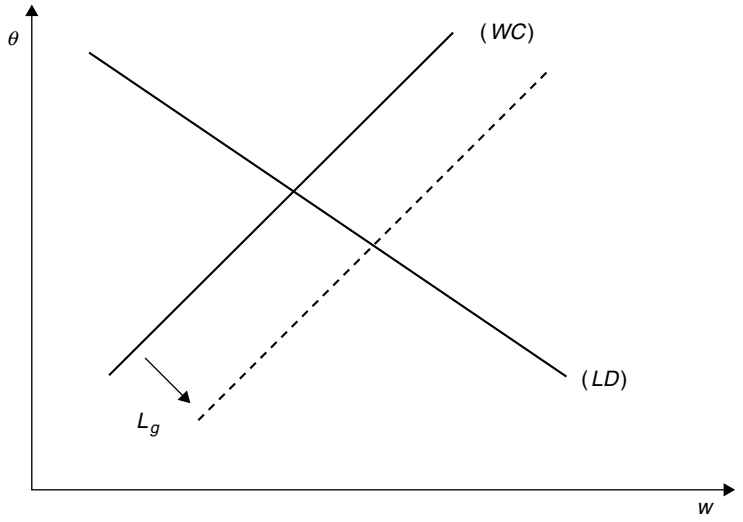


FIGURE 14.7
The effects of public-sector jobs on wages.

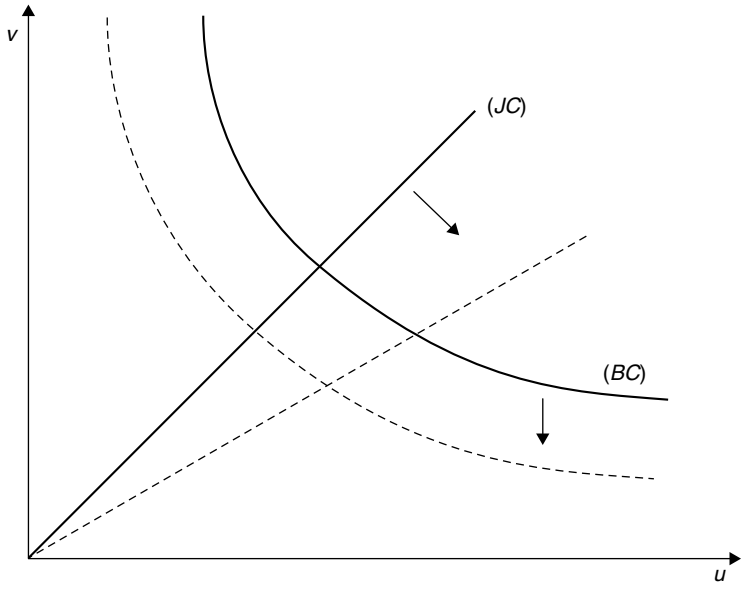


Figure 14.8
Labor market equilibrium with public-sector job creation.

unemployment, public-sector employment exerts upward pressure on the negotiated wage and thus proves liable to crowd out private employment.

The equilibrium unemployment rate is obtained by focusing on the intersection of the Beveridge curve (BC) defined by equation (14.37) with the line issuing from the origin with slope θ. Figure 14.8 sums up this situation. An increase in public-sector employment also leads to a downward shift of the Beveridge curve, so it is equivalent to greater efficiency in the matching process. This improved efficiency runs counter to the crowding-out effect on private-sector jobs and, in short, the variations in the unemployment rate are ambiguous. Therefore, the creation of public job does not necessarily reduce the unemployment rate.

All in all, these analyses suggest that the widespread subsidization of private-sector jobs and the creation of public-sector jobs are measures that do not systematically increase overall employment. Given that such measures can also be very costly, we may provisionally conclude that they should not be adopted in systematic fashion to combat unemployment.

2.4 The Equilibrium Effects of Targeted Measures

Having established that the creation of public-sector jobs may exert a negative effect on the creation of private-sector jobs, let us now proceed to demonstrate that targeted measures to aid certain populations or certain firms may likewise have negative effects on populations or firms not directly affected by the measures in question. To this point we have analyzed the effects of job search help and subsidies for employment on the assumption that all unemployed persons and all firms receive the same aid. In actuality, employment policies are often differentiated by type of firm or type of worker in such a

way as to target those who, in the first instance, are most in need of help. Such targeting aims in principle to cut back on windfall effects and thus improve the cost–benefit ratio of these policies by not aiding, or by furnishing less aid to, firms that would have hired or to job seekers who would have found a job in any case. We start by presenting the impact of job search assistance that, while aimed at job seekers of a certain type, affects overall labor supply. Next we examine the impact of a hiring subsidy that, while aimed at certain firms, affects overall labor demand.

2.4.1 Helping Some Workers Find a Job

The simple matching model presented in chapter 9 can be extended to the case where we have two groups of unemployed persons, those receiving special job search support and those not receiving it. In this setting, we will see that an improvement in the odds of getting a job for those receiving intensive support is not without influence on the same odds for other people. This was shown notably by Cahuc and Le Barbanchon (2010) and Crépon et al. (2013). In what follows we present the model of the former.

A Matching Model with Counseled and Uncounseled Unemployed

Workers are in principle identical and can be in one of three states: (1) employed, (2) unemployed and counseled, or (3) unemployed and not counseled. Upon entering unemployment, workers are not counseled. They may enter into counseling at a rate $\mu > 0$ and if so they keep on receiving counseling until they find a job. We denote by u and $\widetilde{u}$ the number of noncounseled and counseled unemployed workers respectively (the total number of workers is normalized to one so that these are also rates). By assumption the only potential effect of counseling is to increase the arrival rate of job offers to the counseled unemployed workers (not always the case in practice, though it is the stated goal of this type of measure; see section 4 below). Let us normalize to one the number of efficiency units of job search per unit of time of each noncounseled unemployed worker. Counseled unemployed workers are assumed to produce a different number of efficiency units of search, denoted $\delta > 1$. Parameter δ is estimated by econometricians who evaluate the impact of counseling by comparing the exit rates out of unemployment of counseled and noncounseled workers, on the assumption that the arrival rate of job offers to the noncounseled workers is not influenced by counseling.

In this setting, the number of efficiency units of job search per unit of time amounts to $s = u + \delta\widetilde{u}$. Hence, the number of employer–worker contacts per unit of time is given by $M(s, v) \geq 0$, where $v \geq 0$ denotes the number of job vacancies and M is the matching function, as in chapter 9. A vacant job meets on average $M(s, v)/v = m(\theta)$ unemployed workers per unit of time, where $\theta = v/s$ is the tightness of the labor market. Similarly, the rate at which counseled and noncounseled unemployed job seekers can meet jobs is $\delta\theta m(\theta)$ and $\theta m(\theta)$ respectively. In this simple model, a firm that creates a vacant job has a chance to fill it with either a counseled or a noncounseled worker. Hence, at the steady state, the value of a vacant job is:

$$r\Pi_v = -h + m(\theta)\left[\alpha\widetilde{\Pi}_e + (1-\alpha)\Pi_e - \Pi_v\right] \tag{14.41}$$

where h is the search cost, $\alpha = \delta\widetilde{u}/s$ is the probability of meeting a counseled worker, $\widetilde{\Pi}_e$ is the value of a job filled with a counseled worker, and Π_e is the value of a job filled

with a noncounseled worker. If we assume that jobs are destroyed at the exogenous rate q, the asset value of a job satisfies:

$$r\widetilde{\Pi}_e = y - \widetilde{w} + q(\Pi_v - \widetilde{\Pi}_e) \quad \text{and} \quad r\Pi_e = y - w + q(\Pi_v - \Pi_e) \tag{14.42}$$

where y is the productivity of jobs and $\widetilde{w}$ (resp. w) the wage of counseled (resp. uncounseled) workers. Using (14.41) and (14.42), the free entry condition $V = 0$ implies that:

$$\frac{h}{m(\theta)} = \alpha\widetilde{\Pi}_e + (1-\alpha)\,\Pi_e = y - [\alpha\widetilde{w} + (1-\alpha)\,w] \tag{14.43}$$

At this stage we may note that increasing the share of counseled workers can either increase or decrease labor market tightness. If the wage of counseled workers is higher than that of noncounseled workers, an increase in the share of counseled workers raises the proportion of highly paid workers and cuts back the expected profit for a vacant job. This reduces labor market tightness and the job finding rate of noncounseled workers $\theta m(\theta)$. The contrary holds if the wage of noncounseled workers is higher. We now turn to wage formation.

As in chapter 9, we assume that wages are negotiated. Let us denote by z the exogenous gain of an unemployed person: since a noncounseled worker has a probability μ to enter counseling per unit of time, the value of job search and of a job for a noncounseled worker are:

$$rV_u = z + \mu(\widetilde{V}_u - V_u) + \theta m(\theta)(V_e - V_u) \quad \text{and} \quad rV_e = w + q(V_u - V_e) \tag{14.44}$$

while the value of job search and of a job for a counseled worked are simply:

$$r\widetilde{V}_u = z + \delta\theta m(\theta)(\widetilde{V}_e - \widetilde{V}_u) \quad \text{and} \quad r\widetilde{V}_e = w + q(V_u - \widetilde{V}_e) \tag{14.45}$$

Note that if a previously counseled worker loses her job, she reverts to the pool of noncounseled workers. Again, let $\gamma \in [0,1]$ be the relative power of the worker in the bargaining process. The surplus of a job filled by a previously counseled worker and that of a job filled by a noncounseled worker are:

$$\widetilde{S} = (\widetilde{V}_e - \widetilde{V}_u) + (\widetilde{\Pi}_e - \Pi_v) \quad \text{and} \quad S = (V_e - V_u) + (\Pi_e - \Pi_v) \tag{14.46}$$

We assume that the result of the wage bargaining gives the following outcome for workers (see chapter 9 for details):

$$\widetilde{V}_e - \widetilde{V}_u = \gamma\widetilde{S} \quad \text{and} \quad V_e - V_u = \gamma S \tag{14.47}$$

With this sharing rule, using the asset values equations (14.42), (14.44), and (14.45) and the definition of surpluses (14.46), we arrive at the value of surpluses $\widetilde{S}$ and S as a function of the model parameters and θ:

$$(r+q)\,\widetilde{S} = y - z - \delta\theta m(\theta)\gamma\widetilde{S} + q\Delta \tag{14.48}$$

$$(r+q)\,S = y - z - \theta m(\theta)\gamma S + \mu\Delta \tag{14.49}$$

where $\Delta = \widetilde{V}_u - V_u = \theta m(\theta)\gamma(\delta\widetilde{S} - S)$ stands for the difference in value of job search for counseled and noncounseled workers.[7] Using the free entry condition and the sharing rules (14.47), one gets a last relation between $\widetilde{S}$, S, and θ:

$$\frac{h}{m(\theta)} = (1-\gamma)\left[\alpha\widetilde{S} + (1-\alpha)S\right] \tag{14.50}$$

This equation defines an increasing relationship between the surplus and labor market tightness. It can be interpreted in terms of labor demand: a larger expected surplus (which is a convex combination of the surplus of jobs filled with counseled workers and the surplus of jobs filled with noncounseled workers) increases the number of vacant jobs created by firms. It is represented by curve (*JC*) in figure 14.9. Now, α is an endogenous parameter which depends on the values of $\widetilde{u}$ and u. Hence, we need two more equations in order to find the equilibrium solution. They are given by the laws of motion of unemployment for the two categories of unemployed workers:

$$\frac{\mathrm{d}\widetilde{u}}{\mathrm{d}t} = \mu u - \delta\theta m(\theta)\widetilde{u} \quad \text{and} \quad \frac{\mathrm{d}u}{\mathrm{d}t} = q(1-u-\widetilde{u}) - \mu u - \theta m(\theta)u$$

Positing $\frac{\mathrm{d}\widetilde{u}}{\mathrm{d}t} = \frac{\mathrm{d}u}{\mathrm{d}t} = 0$ at the steady state, one gets the equilibrium values of total unemployment $u^* = \widetilde{u} + u$:

$$u^* = \frac{q\left[\mu + \delta\theta m(\theta)\right]}{q\left[\mu + \delta\theta m(\theta)\right] + \delta\theta m(\theta)\left[\mu + \theta m(\theta)\right]} \tag{14.51}$$

and the Beveridge curve, represented by (*BC*) in figure 14.9, is simply given by this latter equation. The system of equations (14.48) to (14.51) allows us to determine the equilibrium values of the four unknown variables $\widetilde{S}$, S, u^*, and θ. Analysis of these sheds light on the impact of counseling on labor market equilibrium. Equations (14.48) and (14.49) show that the surplus of jobs filled with counseled workers $\widetilde{S}$ is smaller than the surplus of jobs filled with noncounseled workers S since $\delta > 1$. The surplus of jobs filled with counseled workers is smaller because counseled workers, who get more contacts with firms offering vacant jobs than noncounseled workers, have a reservation wage, equal to $r\widetilde{V}_u$, which is higher than the reservation wage of noncounseled workers, equal to rV_u.

The Properties of the Labor Market Equilibrium

With this property in mind, and recalling that $m'(\theta) < 0$, it can easily be understood how counseling may reduce labor market tightness by the differentiation of equation (14.43) that results from the free entry condition:

$$\frac{h}{m(\theta)^2}m'(\theta)\frac{\partial\theta}{\partial\mu} = \frac{\partial\alpha}{\partial\mu}(\widetilde{w} - w) + \left[\alpha\frac{\partial\widetilde{w}}{\partial\mu} + (1-\alpha)\frac{\partial w}{\partial\mu}\right] \tag{14.52}$$

[7] In that case, the negotiated wages are $\widetilde{w} = \Gamma(\theta)[y - z - \lambda\Delta] + z - q\Delta$ and $w = \Gamma(\theta)[y - z - \mu\Delta] + z - \mu\Delta$ with $\Delta = \widetilde{V}_u - V_u = \frac{\gamma}{1-\gamma}\frac{\theta m(\theta)}{(r+q)(r+\mu)}[(\delta-1)y - \delta\widetilde{w} - w]$. So $\Delta > 0$, and $\widetilde{w} - w = (\mu - \lambda)(1 - \Gamma(\theta))\Delta > 0$ if $\mu > \lambda$.

- First, an increase in the proportion of counseled workers (induced by a higher μ) reduces the probability that noncounseled workers get a job offer and raises the probability that vacant jobs are matched with counseled workers who get higher wages. The upshot is fewer surplus jobs (first term of the right-hand side of equation (14.52), with $\frac{\partial \alpha}{\partial \mu} > 0$). This *composition effect* reduces the expected profits of filled jobs and then induces firms to create fewer job vacancies altogether. This effect pushes the job creation curve (*JC*) to the right on figure 14.9, which corresponds to fewer vacancies for a given level of unemployment.
- Second, everything else being equal, an increase in the proportion of counseled workers decreases the value of the surplus of jobs filled with noncounseled workers because it improves their outside option (and their wage): in the future, should they lose their jobs, previously noncounseled workers would then have higher chances to get counseling (second term of the right-hand side of equation (14.52), with $\frac{\partial w}{\partial \mu} > 0$, whereas in contrast $\frac{\partial \tilde{w}}{\partial \mu} < 0$). When the number of counseled workers is relatively small, this *wage effect* contributes to reduce expected profits and hence reduces labor market tightness. This induces the job creation curve (*JC*) to move to the right in figure 14.9.
- Third, everything else being equal, the value of the surplus of jobs filled with counseled workers increases when there is more counseling because these jobs are filled more rapidly thanks to higher search intensity. This is the *direct effect* of counseling, which causes counseled job seekers to leave unemployment faster, thus helping firms save on vacancy costs. This effect induces the job creation curve (*JC*) to move to the left in figure 14.9.

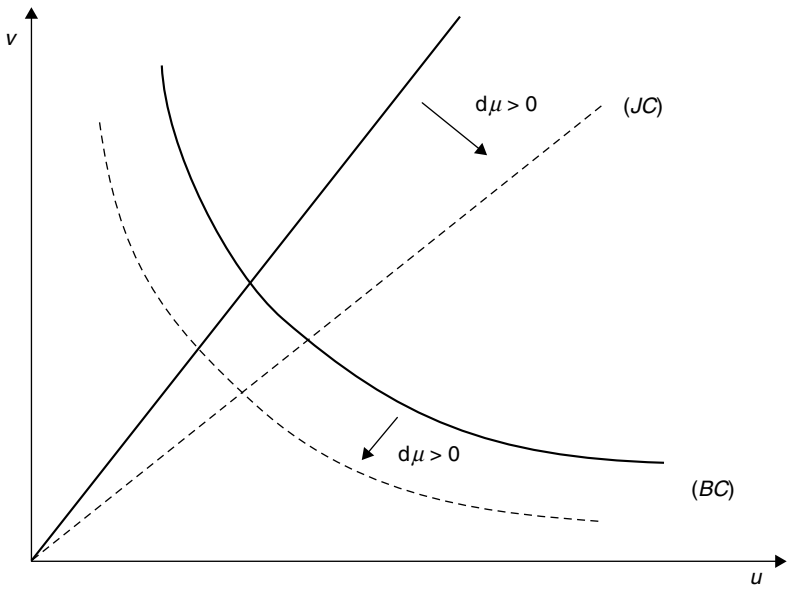


FIGURE 14.9
The equilibrium effect of a targeted counseling program.

Note: Parameter μ is the probability of entering the program for the unemployed.

If the first two effects dominate, which is the case for simulations done with a wide range of relevant values of the parameters, and when the share of counseled workers is relatively modest, as shown by Cahuc and Le Barbanchon (2010), then the increase in the share of counseled unemployed persons pushes the job creation curve (*JC*) to the right. The direct effect may dominate only when the number of counseled workers is sufficiently high.

Once the effect of counseling on labor market tightness is known, it is possible to look at its impact on the steady-state unemployment rate u^*. Differentiating u^*, given by equation (14.51), with respect to the entry rate into counseling μ gives:

$$\frac{\mathrm{d}u^*(\theta,\mu)}{\mathrm{d}\mu} = \frac{\partial u^*(\theta,\mu)}{\partial \mu} + \frac{\partial u^*(\theta,\mu)}{\partial \theta}\frac{\mathrm{d}\theta}{\mathrm{d}\mu}$$

Since $\partial u^*(\theta,\mu)/\partial\mu < 0$ and $\partial u^*(\theta,\mu)/\partial\theta < 0$ (a property typical of the Beveridge curve), the interpretation of the sign of these partial derivatives is straightforward. First, an increase in the entry rate into counseling raises the share of the unemployed who exit unemployment at a higher rate. The effect on unemployment, everything else being equal, is negative: $\partial u^*(\theta,\mu)/\partial\mu < 0$. This effect induces the Beveridge curve (*BC*) to move inward in figure 14.9. Second, when labor market tightness is increased, the exit rate out of unemployment is higher and unemployment drops: $\partial u^*(\theta,\mu)/\partial\theta < 0$. If more counseling always increases labor market tightness, $\mathrm{d}\theta/\mathrm{d}\mu > 0$, then counseling unambiguously reduces unemployment; but if the contrary is the case, $\mathrm{d}\theta/\mathrm{d}\mu < 0$, then the total impact of counseling on steady-state unemployment is ambiguous. Now in the previous paragraph we saw that the sign of $\mathrm{d}\theta/\mathrm{d}\mu$ depends on which of the three identified effects dominates, something which also depends of the value of the parameters.

Cahuc and Le Barbanchon (2010) simulate this model for the French labor market using a matching function of the type $m(\theta) = m_0\theta^{-0.5}$ and do indeed find that the relationship between unemployment and the entry rate of the unemployed into counseling is hump-shaped (see figure 14.10). This model shows that the equilibrium effects of labor market policies can outweigh their direct effects, notably when programs are small in scale, affecting fewer than 10% of workers. Thus a naive evaluation, relying on a simple comparison of outcomes for the treated and the nontreated, could lead to erroneous results if the policy induces equilibrium effects which change the baseline arrival rate of job offers $\theta m(\theta)$. Such an evaluation would neglect equilibrium effects by taking for granted that nontreated job seekers would not be affected by the programs, that is, job offers to the nontreated in the absence of the policy would be the same as those observed by the econometrician in the presence of the policy. Thus the counterfactual arrival rates for this category of worker would be wrong, and that could lead to the conclusion that counseling always decreases unemployment—which is not necessarily true, especially for small programs.

2.4.2 Helping Some Firms Hire Workers

Equilibrium effects can also hinder the impact of targeted subsidies that help some but not all firms. Again, this can be seen easily in the framework of the search and matching model.

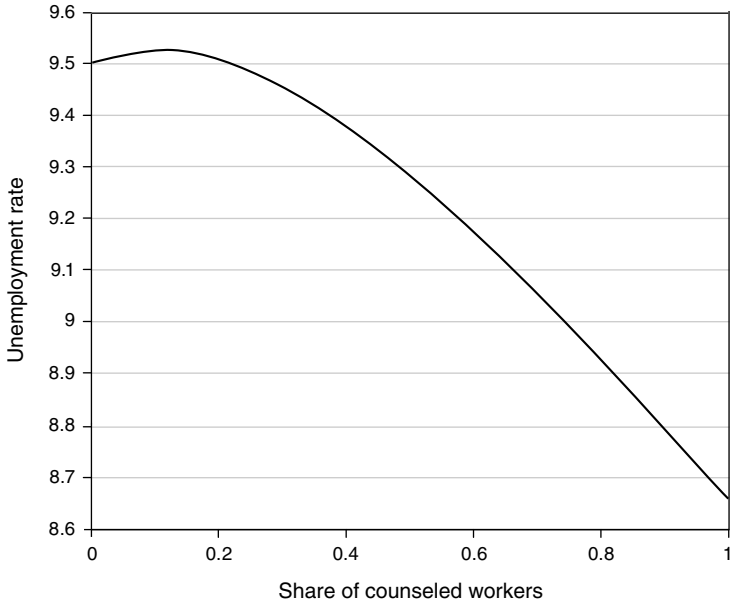


Figure 14.10
The relationship between the unemployment rate and the share of counseled workers.

Source: Cahuc and Le Barbanchon (2010, figure 1).

Consider a set of ex ante identical firms. Only a fraction α of firms can benefit from a subsidy s granted by the government. For simplicity, let us assume that firms do not know, upon creating a vacant job, whether they will benefit from the subsidy or not. They only find out if they have been selected once the match is formed. Let $\widetilde{\Pi}_e$ be the value of jobs for a firm receiving subsidies and Π_e the corresponding value for firms not receiving them. We then have:

$$r\widetilde{\Pi}_e = y - \widetilde{w} + s + q(\Pi_v - \widetilde{\Pi}_e) \quad \text{and} \quad r\Pi_e = y - w + q(\Pi_v - \Pi_e) \tag{14.53}$$

where $\widetilde{w}$ is the wage of subsidized jobs. The value of a vacancy then depends on the probability of getting the subsidy:

$$r\Pi_v = -h + m(\theta)\left[\alpha\widetilde{\Pi}_e + (1-\alpha)\,\Pi_e - \Pi_v\right]$$

The expected utilities of an employee who occupies either a subsidized job or a nonsubsidized job are respectively defined by:

$$r\widetilde{V}_e = \widetilde{w} + q(V_u - \widetilde{V}_e) \quad \text{and} \quad rV_e = w + q(V_u - V_e) \tag{14.54}$$

Unemployed workers may be matched up with either a subsidized job or a nonsubsidized job and may accept the job offer as long as it has a value above their reservation wage rV_u, which must be the case for either type of job in equilibrium (otherwise

one type of vacancy would never be filled). Hence:

$$rV_u = z + \theta m(\theta)\left[\alpha \widetilde{V}_e + (1-\alpha)V_e - V_u\right] \tag{14.55}$$

Using equations (14.53) and (14.54), we find that the surplus of subsidized jobs, $\tilde{S} = \widetilde{V}_e - V_u + \widetilde{\Pi}_e - \Pi_v$, and of nonsubsidized jobs, $S = V_e - V_u + \Pi_e - \Pi_v$, can be written:

$$\tilde{S} = \frac{y + s - rV_u - \Pi_v}{r+q} \quad \text{and} \quad S = \frac{y - rV_u - \Pi_v}{r+q}$$

It is evident that the subsidy increases the surplus in exactly the same way as if the productivity of the job were increased by an amount equal to the subsidy. Now, as seen in section 3.3 of chapter 9, the negotiated wages are determined by wage bargaining, which yields a share γ of the job surplus to the worker. Bargaining induces wages similar to those obtained in chapter 9, equation (9.20):

$$\tilde{w} = \gamma(y+s) + (1-\gamma)rV_u \quad \text{and} \quad w = \gamma y + (1-\gamma)rV_u = \tilde{w} - \gamma s \tag{14.56}$$

The expression for the wage $\tilde{w}$ of subsidized jobs shows that the subsidy raises the wage because bargaining implies that workers and employers share the increased surplus provided by the subsidy. Moreover, equation (14.56) shows that the two wages are linked by the value of the reservation wage rV_u, which is common to all workers. Hence, even though the wage of nonsubsidized jobs is lower than that of subsidized jobs, both wages depend on the subsidy. It is evident that $\tilde{w}$ increases both directly, thanks to the subsidy, which is shared between the employer and the employee, and indirectly, thanks to the increase in the reservation wage; it is also evident that the wage of nonsubsidized jobs only increases to the extent that rV_u increases with the subsidy.

The value of job search rV_u can be obtained from equations (14.54), the entry condition, $\Pi_v = 0$, and the surplus sharing rule:

$$rV_u = z + \frac{\gamma}{1-\gamma}\theta h \tag{14.57}$$

It can easily be verified that the equilibrium value of labor market tightness is determined by an equation similar to that obtained in equation (9.22) in chapter 9, which is now written:

$$\frac{(1-\gamma)(y + \alpha s - z)}{r + q + \gamma\theta m(\theta)} = \frac{h}{m(\theta)} \tag{14.58}$$

This equation shows that labor market tightness always increases with the amount of the employment subsidies s and with the share α of firms that benefit from the subsidy. However, the overall impact of the subsidy can be decomposed into a direct effect, beneficial to employment, which increases the surplus, the profits, and the wage of subsidized jobs, and an indirect effect, detrimental to employment, which increases rV_u, as shown by equation (14.57). The increase in rV_u raises the wage of nonsubsidized jobs and decreases the surplus and the profits of these jobs. The indirect effect, which is

detrimental to profits, diminishes the positive impact of employment subsidy programs on employment—an aspect of policy that evaluation strategies must not neglect to take into account.

3 EVALUATING LABOR MARKET POLICIES

Most empirical research tries to judge the value of labor market policies by comparing the observed impact of a policy measure on the agent who benefits (for example, the number of hires by a firm receiving subsidies) with what would have been the outcome if the measure in question had not been applied to that agent. The difficulty of this exercise lies in the fact that the latter result is not observed. The solution to the problem of missing data is to assume that available data on the behavior of *other* agents can, under certain conditions, take its place. The impact of a policy measure on a particular agent should only be the first step in the assessment. In line with the theoretical structures presented in this chapter, we must pursue the analysis with the help of an equilibrium model of the whole labor market. As we will see, empirical research conforming to this prescription is still rare.

The evaluation of labor market policies is grounded in the notion of potential gain, which represents the difference in the levels taken by a given indicator (wages, for example) in the presence and in the absence of the policy measure being examined. In practice, potential gain is pinpointed with the help of several standard estimators, the calculation and the validity of which depend on the available data. Data of this kind most often come from surveys (or, more rarely, from pre-existing administrative data sets) so we speak of observational or nonexperimental data. Selection bias is the main weakness of assessments made on this type of data and, in response, the "social experiment" approach has undergone considerable development in recent years. Such experiments aim to reproduce in the field of economics the experimental techniques employed in sciences like agronomy, biology, and medicine. The other limitation of impact evaluation methods is the identification of potential equilibrium effects, by which the program may exert externalities on nonparticipants and affect their labor market outcomes.

3.1 The Challenges Ahead: Selection Bias and Externalities

Every labor market policy has a precise goal: for example, a training placement is intended to increase the human capital of an individual. The success of such policies will be judged on the basis of a tangible result, which, in this example, might be a higher wage or a higher probability of gainful employment. In the literature on labor market policy, this result is often referred to as the individual's *response*. The observer generally knows the gross impact of a policy on the beneficiary, for example the wage received after a training placement. But in order to assess the efficiency of this policy, the observer must also know what wage the *same* person would be receiving if he had not had the benefit of the placement. This is the nub of the problem, since the latter wage is not observed. Hence the essential question facing any evaluation of a policy measure is this: how would a person or a firm who has benefited from a measure—a "treated" person or firm—have responded if they had not benefited from that measure?

This approach to the evaluation problem is therefore based on the notion of "*potential* outcome," attributed to, among others, Fisher (1935), Roy (1951), Quandt (1972), and Rubin (1974). The literature on the subject often refers to the Roy-Rubin model, which is the basis of various evaluation methods based on either experimental or observational data (difference-in-differences, regression discontinuity, etc.). For a detailed presentation see for instance Wooldridge (2010, chapter 21) and DiNardo and Lee (2011).

3.1.1 Potential Outcome

Let t_P be the period, assumed to be unique, over which the "treatment" is applied. Let us assume further that we can attribute two potential responses to each individual, which we designate by y_{it}^1 and y_{it}^0. The variable y_{it}^1 represents the response of agent i that would be observed at date t if he were treated, while the variable y_{it}^0 represents the response of agent i that would be observed at date t if he were not treated. Readers should note that date t can be posterior or anterior to the period t_P of the treatment and should pay close attention to the terminology used. Before the treatment, a person referred to as treated has not yet undergone the treatment but will definitely do so during period t_P. Conversely, after the treatment a person referred to as treated has in fact undergone the treatment. Results y_{it}^1 and y_{it}^0 are described as *potential*, for "to be treated" and "not to be treated" are two mutually exclusive states: it is not possible to observe the responses of the *same* individual i at the *same* date t in these two states.

To distinguish potential outcomes from actual ones, it is best to work with a dummy variable δ_i, which takes a value of 1 if agent i has actually benefited from the measure, and 0 if not. The challenge of the evaluation problem comes from the fact that the econometrician does observe the realizations of the variable $y_{it} = \delta_i y_{it}^1 + (1 - \delta_i) y_{it}^0$ (i.e., for a given individual observes either y_{it}^1 or y_{it}^0) but never observes *simultaneously* the realizations of variables y_{it}^1 and y_{it}^0 for the same individual. In particular, he never observes the realizations of the gain of the treatment defined by $\Delta_{it} = y_{it}^1 - y_{it}^0$. This unobserved result is called the "counterfactual outcome." For a treated person i, the counterfactual outcomes correspond to realizations of y_{it}^0 whereas for an untreated agent j, the counterfactual outcomes correspond to realizations of y_{jt}^1. Formally, our evaluation problem is thus a missing data problem.

3.1.2 Contrast Variables, Selection Bias, and Identifying Assumptions

If we limit ourselves to direct effects, the efficiency of a measure is generally assessed with the help of a *contrast variable*; the one most commonly adopted is the ATT, the *average effect of the treatment on the treated*. The ATT is defined by (omitting indices i and t for simplicity):

$$\mathbb{E}(\Delta \mid \delta = 1) = \mathbb{E}(y^1 \mid \delta = 1) - \mathbb{E}(y^0 \mid \delta = 1) \tag{14.59}$$

In principle, the data allow us to know $\mathbb{E}(y^1 \mid \delta = 1)$ and $\mathbb{E}(y^0 \mid \delta = 0)$, which represent respectively the average responses of a treated person and an untreated one, but they do not allow us to determine $\mathbb{E}(y^0 \mid \delta = 1)$, which represents what would, on average, have been the response of that person if he had not undergone the treatment that in

fact he did undergo.[8] To assess the average gain from treatment defined by (14.59), the econometrician is thus obliged to make an *identifying assumption*, which gives him the means to estimate the expected value of the counterfactual outcome $\mathbb{E}(y^0 \mid \delta = 1)$ using the available data. The identifying assumption lets the econometrician link the *unobserved* responses of the treated group to the *observed* responses of the members of the control group. As we will see below, the identifying assumption depends on the data available and influences the estimation procedure. In making it, the econometrician is facing a difficult problem: finding a group of individuals, the control group, who have not undergone the treatment and are as nearly identical to the treated group as possible. In practice, people self-select to participate or not in a program based on personal characteristics or situations which are not always observable. The existence of such selection biases makes it a challenge to arrive at a good counterfactual outcome.

Policy measures can also be judged with the help of other contrast variables, like the average treatment effect (ATE), which is the unconditional mean $\mathbb{E}\left(y^1 - y^0\right)$, that is, the effect of this type of program for the whole population. Measuring the ATE makes less sense in the case of targeted measures such as labor market programs since most of the population would not be eligible. Another example is $\Pr(\Delta > 0 \mid \delta = 1)$, which represents the proportion of participants for whom the program was beneficial. For simplicity, we take the view that the only contrast variable is the average gain from the treatment on the treated (ATT), but what follows can easily be applied to any contrast variable. In general the assessment of the "success" of the treatment is achieved by comparing this average gain to an indicator of the cost of the treatment.

3.1.3 Indirect Effects: From Partial Equilibrium to General Equilibrium

Most of the studies that aim to evaluate labor market policies try to assess the behavior of an agent reacting to a precise measure without taking into account the effect this measure might have on the decisions of other agents—which might, in turn, change the environment within which the first agent responds to the measure under consideration. Indeed, if the control group is, say, negatively affected by the measure, then the outcome among workers in this group is lower than it would have been absent the program. If we want to assess these "indirect" effects, to use the terminology of Lewis (1963), then we have to work with an equilibrium model of the entire labor market. Layard and Nickell (1986) and Calmfors (1994) established the following typology of the principal indirect effects which the Roy-Rubin model leaves out.

1. Displacement or crowding-out effects: the jobs created by a measure destroy existing jobs to which the measure does not apply. This happens when, for example, firms benefiting from subsidies increase their production and their market share at the expense of other firms. Another channel is the effect on wages: subsidized firms might not only recruit more but also offer higher wages, which also forces the unsubsidized firms to increase their wages and reduce their employment.

[8] Obviously, it is also not possible to determine $\mathbb{E}(y^1 \mid \delta = 0)$, which represents the response with treatment of a person not treated.

2. Substitution effects: the jobs created flow to the beneficiaries of a particular measure at the expense of those who are not targeted. For instance, tailoring measures to specific groups typically aims at enhancing their efficiency by limiting windfall effects (also called deadweight effects). Windfalls occur when, for some units, the impact of a measure differs hardly at all from what would have happened if it had not been implemented. Subsidies can represent a "windfall" for the firm and a "deadweight" for society if firms would have hired workers anyway. Tailoring subsidies to fit the profile of hard-to-place workers can limit such deadweight effects. But doing so might in turn generate substitution effects if firms then make it a priority to hire workers for whom the subsidy was tailored instead of ones who are not eligible for it.

3. Tax effects: the taxes needed to finance a measure affect the decisions of all agents.

Now, most labor market policies are targeted at specific groups, types of firms, or jobs and do entail such indirect equilibrium effects. The great majority of empirical studies have looked only at the direct effects of labor market policies, neglecting their effects on the general equilibrium of the economy. Aside from the fact that it is clearly harder to make a global assessment anyway, the emphasis on direct effects arises from the predominance of U.S. research in this area. In the United States the amounts budgeted for employment policy are relatively small, so it seems reasonable to assume that their macroeconomic effects are negligible. Heckman et al. (1999) do however argue that global effects ought to be given more prominence in the assessment, since, apart from its costs, a policy measure affects the behavior of both the beneficiaries and the nonbeneficiaries. We will see below how several empirical studies deal with this issue.

3.2 Evaluation Based on Controlled Experiments

For existing programs, the identification of a counterfactual outcome is often a difficult task, even an impossible one in the case of very large-scale interventions. For that precise reason, controlled experiments are often regarded as the gold standard of policy evaluation. Data from such experiments do indeed escape the problem of selection bias, in principle. Let us suppose that we want to assess the benefits of a training program. A social experiment will consist of dividing the individuals eligible for the program, and who agree to take part in the experiment, into two randomly chosen groups: a treatment group which does in fact benefit from the program and a control group which does not. This random division of the participants is called "randomization." If the two groups are large enough, randomization entails that, on average, observed and unobserved characteristics will be identical within each group. That being so, the differences in the average results observed between these two groups will depend only on the program or "treatment" administered, and selection bias will be eliminated. We illustrate this approach with a job placement experiment conducted in France in the 2000s and analyzed by Crépon et al. (2013). The related data and programs are available at www.labor-economics.org.

It is worth noting that "social experiments" must not be confused with "natural experiments." The latter term applies to studies that use an exogenous change in a policy measure, like a rise in the minimum wage or a tax reduction, to estimate the effects of this measure on a given population. The treated group is then the set of persons

belonging to the population who benefit from this change, and the control group is the set, or a subset, of the persons in the same population to whom it does not apply. The data produced by natural experiments do not, therefore, automatically respect the conditions imposed by randomization and must be considered, strictly speaking, non-experimental.

3.2.1 An Example of Randomization: The Impact of a Targeted Job Placement Program for Skilled Youth

As we saw in section 2.4, labor market programs can entail equilibrium effects that may lead the investigator to overestimate their impact on wages or employment if the evaluation does not make it possible to control for them. Crépon et al. (2013) provide a good example of the method used to identify this type of effect. They base their results on a large-scale, randomized social experiment that took place in France in 2007, which focused on intensive job seeker assistance. This type of program is very common in the OECD countries, whether targeted at youth, displaced workers, the long-term unemployed, or various groups at risk of long-term unemployment.

The Experiment

Under the experimental program in question, private agencies were contracted to provide intensive placement services to young graduates (with at least a two-year college degree) aged below 30 and having spent at least 6 of the last 18 months continuously unemployed or underemployed. The private providers were paid partially on delivery, that is, conditional on the individual finding a job with a contract of at least six months and staying employed for at least six months. The intervention unfolded in two phases. In the first one, lasting for a maximum of six months, the agency counseled the job seekers. In the second phase, if a job was found, the newly employed worker was given further support and counseling for a maximum of her first six months at work to help her keep that job or get another one if she lost it. The experiment took place in 235 cities across the country, in 10 administrative regions covering about half of France.

All eligible young unemployed workers in a given area were first identified by the public employment services and then a fraction p was randomly assigned to the program. This approach avoids the selection bias that can undermine the correct comparison of outcomes between treated and nontreated groups. Indeed, randomization ensures that the two groups are comparable as long as the size of the sample is sufficiently large: here there is no reason to think that the two groups differed with respect to some unobserved characteristics that might influence employment outcomes. Since participants were assigned to the experiment over 14 months, the authors benefit from 14 monthly cohorts, starting in September 2007. They focus on cohorts 3 through 11 to ensure comparability. In these cohorts, 29,636 individuals were randomly selected to be surveyed. In this way, the authors come up with a large volume of cross-sectional data specifying the responses of treated and untreated persons after the end of the program.

The Identifying Assumptions

The first step of the evaluation procedure consists in measuring the gain from the program, without accounting for potential equilibrium effects. Let δ be a dummy which equals 1 if the individual is assigned to participate in the program. If y_A^0 and y_A^1 denote

the outcomes for nontreated and treated individuals respectively, at a period *after* the program has been implemented, then the average treatment effect on the treated (ATT) we want to identify, conditional on the vector $\mathbf{X}$ of observable characteristics of individuals, is:

$$\mathbb{E}\left(y_A^1 - y_A^0 \,|\, \mathbf{X}, \delta = 1\right) \tag{14.60}$$

that is, the difference in the outcomes among participants ($\delta = 1$) when treated (y_A^1) and when not treated (y_A^0). Conditioning on the set $\mathbf{X}$ of observable characteristics helps to control for remaining differences in characteristics across groups that may affect the outcome variable. The problem is that the econometrician does not observe y_A^0 when $\delta = 1$ because the same participants cannot have been both treated and untreated. So the strategy is to use the outcome observed for nonparticipants after the program was implemented. This creates a problem of *causal inference*: under what assumptions can we infer the impact of the program by observing only y_A^0 when $\delta = 0$? The answer is that two identifying assumptions are required:

1. Nonparticipants are "comparable" to participants.

2. Nonparticipants are not affected in any manner by the existence of the program.

Let us discuss briefly these two important assumptions.

The identifying assumption related to (1) is called the *conditional independence assumption*:[9]

$$\mathbb{E}\left(y_A^0 \,|\, \mathbf{X}, \delta = 1\right) = \mathbb{E}\left(y_A^0 \,|\, \mathbf{X}, \delta = 0\right) \tag{14.61}$$

This equality signifies that conditional on the vector $\mathbf{X}$ of observable characteristics, the average effect of nontreatment (y_A^0) is the same for a participant ($\delta = 1$) in the experiment and a nonparticipant ($\delta = 0$). Condition (14.61) implies that participation is unrelated to what individuals would earn in the absence of the program. Hence, observing the outcomes of the nontreated individuals who are not participating in the program—which the econometrician can do—determines exactly what would have happened to the treated individuals if they had not participated in the program. This assumption

[9] The expression of ATT (14.60) can be rewritten:

$$\begin{aligned}\mathbb{E}\left(y_A^1 - y_A^0 \,|\, \mathbf{X}, \delta = 1\right) &= \mathbb{E}\left(y_A^1 \,|\, \mathbf{X}, \delta = 1\right) - \mathbb{E}\left(y_A^0 \,|\, \mathbf{X}, \delta = 1\right) \\ &= \mathbb{E}\left(y_A^1 \,|\, \mathbf{X}, \delta = 1\right) - \mathbb{E}\left(y_A^0 \,|\, \mathbf{X}, \delta = 0\right) \\ &\quad + \underbrace{\mathbb{E}\left(y_A^0 \,|\, \mathbf{X}, \delta = 0\right) - \mathbb{E}\left(y_A^0 \,|\, \mathbf{X}, \delta = 1\right)}_{\text{selection bias}}\end{aligned}$$

Hence if the following condition were satisfied:

$$\mathbb{E}\left(y_A^0 \,|\, \mathbf{X}, \delta = 1\right) = \mathbb{E}\left(y_A^0 \,|\, \mathbf{X}, \delta = 0\right)$$

the ATT could be estimated based on observable outcomes only.

is reasonable for randomized experiments on condition that the sample is sufficiently large. Randomization aims precisely at getting rid of selection bias, and any remaining differences across individuals based on observables can be controlled by conditioning on $\mathbf{X}$. Note that this is less likely for natural or quasi-experiments, which typically require other methods such as difference-in-differences and matching (see section 3.3 below).

The identifying assumption related to (2) is called the *stable unit treatment value assumption* (SUTVA) in the literature (see Rubin, 1980, 1990). This assumption rules out cases where the treatment of one individual affects another's outcome. Most of the studies aiming to evaluate labor market policies implicitly make this assumption. However, as we saw in section 2.4, there are many reasons to suspect the existence of such effects. If these effects on nonparticipants do in fact exist, then the ATT estimator will be biased unless we adopt a strategy to account for them. This is what Crépon et al. (2013) do, and we will get back to their procedure after running a "naive" estimation that ignores these effects.

The Cross-Section Estimator

The *cross-section* estimator of the average gain from the program (also called difference-in-means), denoted $\tilde{\Delta}_{CS}$, is then given by:

$$\tilde{\Delta}_{CS} = \overline{y}_A^T - \overline{y}_A^C \tag{14.62}$$

where $\overline{y}_A^T$ and $\overline{y}_A^C$ are the average observed outcomes among individuals respectively assigned and not assigned to the program *after* the program has taken place. Thus we simply need to compare the average result of individuals actually treated and those untreated at dates following the treatment. When the conditional independence assumption (14.61) and the stable unit treatment value assumption hold good, the estimator $\tilde{\Delta}_{CS}$ is an unbiased estimator of the average gain $\mathbb{E}\left(y_A^1 - y_A^0 \mid \mathbf{X}, \delta = 1\right)$. Note that, as opposed to the difference-in-differences method often used in the quasi-experimental approaches, there is no need for a time dimension to identify the average gain in the case of randomized experiments. This is because it is not necessary to net out potential differences in outcome across groups that would stem from differences in characteristics, since both groups are made fully comparable through the randomization. For the same reason, this approach does not require us to make the "common trend assumption," unlike the difference-in-differences methods studied in previous chapters (e.g., chapters 1, 5, and 12, and also below).

A "Naive" Estimation

To implement the cross-section estimator, one can compute the corresponding averages of outcomes among participants and nonparticipants. One can also simply estimate the following equation using the OLS, which has the advantage of allowing us to control for observed characteristics:

$$y_i = \alpha_1 + \beta_1 \delta_i + \mathbf{X}_i \boldsymbol{\gamma}_1 + \varepsilon_i \tag{14.63}$$

In this equation y_i is a dummy taking a value of 1 if the person is under a fixed-term contract of six months or more (LTFC) or, alternatively, under any long-term job arrangement (fixed-term contract of more than six months or permanent contract, LT), eight months after the beginning of the experiment; y_i takes a value of 0 if the person is not under LTFC or LT; δ_i is a dummy equal to 1 if the individual is assigned to participate in the program (whether actually treated or not) and 0 otherwise; X_i is a vector of control variables including a dummy for each cohort of entry into the program, and individual-level control variables (age, gender, education, past duration of unemployment, and its square). ε_i is a random term with zero mean. With these settings, the estimated coefficient $\widehat{\beta}_1$ yields the impact of the program on the probability of long-term employment of participants compared to nonparticipants.

Note that in practice, there is a difference between being assigned to participate in the program and actually being treated, that is, benefiting from the offered services. Indeed, some individuals assigned to the experiment actually declined the services. This is very often the case in social experiments, for individuals who have been randomly selected and assigned to a treatment group may balk at participating in the program. In other words, there is always some *non-take-up*, which in the present case amounts to about two thirds of the workers assigned to treatment. But non-takers cannot be excluded from the sample of participants, because take-up is itself subject to selection bias: among the assigned individuals, those who declined the services probably had specific characteristics that might also influence their odds of employment (actually non-take-up occurs mainly among those who were already in employment at the time of the experiment and who may have been more employable than others). Excluding non-takers from the pool of participants would make the two groups of participants and nonparticipants no longer comparable and would violate the conditional independence assumption (14.61), to the extent that self-selection can be based on unobservable characteristics that cannot be controlled for. Non-take-up limits the direct impact of program assignment on employment outcomes, and it also reduces the power of the identification strategy (for due to standard error it is more difficult to identify a small significant effect than a large one with a given number of observations) but does not annihilate it if there are enough takers among assigned participants.

Results from estimating equation (14.63) are presented in table 14.2. Overall, participants assigned to treatment are only 0.7 percentage point more likely to have obtained a fixed-term contract of six months or longer (LTFC) and 0.2 percentage point more likely to have any long-term job (LT). These estimates are not significantly different from zero at the 10% level of confidence. This is in comparison with a mean of 20% for the control group (column 3). However, let us consider those who were not employed at the beginning of the study and for whom take-up was higher (about 43%): they were 1.7 percentage points (11% if we compare this with a mean of 16% for the control group) more likely to have an LTFC and 1.5 percentage points more likely to have an LT (4%) if they were assigned to treatment than if they were not (column 4). The authors also show that the effects are stronger for young men than for young women. For the latter category, the effects are actually insignificant for all type of workers (working or not).

Table 14.2
The impact of intensive job placement counseling on the employment outcomes of young educated workers, leaving out equilibrium effects.

Outcome	Variable	All participants	Not employed
LTFC	Assigned to treatment (β_1)	.007 (.005)	.017*** (.006)
	Control mean (α_1)	.20	.16
LT	Assigned to treatment (β_1)	.002 (.007)	.015 (.010)
	Control mean (α_1)	.47	.37
Observations		21,431	11,806

Note: The table reports ordinary least squares (OLS) regressions controlling for gender, education, past duration of unemployment and its square, cohort dummies, and 47 dummies for local area quintuplets. The dependent variables are employment outcomes when surveyed eight months after the random assignment: long-term fixed contracts (LTFC) are fixed-term contracts with a length of at least six months; long-term employment (LT) is either a long-term fixed contract or an indefinite-term contract. Column 4 restricts the sample to job seekers who did not report that they were employed at the time of randomization. Standard errors in parentheses are robust to heteroskedasticity and clustered at the local area level. *** significant at the 1% level.

Source: Crépon et al. (2013, table III).

3.2.2 Accounting for Equilibrium Effects

The estimated effect obtained from equation (14.63) is a gross effect, for it ignores the potential impact of the treatment on nonparticipants. These externalities potentially bias the estimates of the true effects of the program on participants. Indeed, if nonparticipants are indirectly and negatively affected by the treatment of participants, as we may suspect based on models reviewed previously, then the observed difference between participants and nonparticipants, coefficient β_1, reflects only in part the positive effect of the program on the probability of employment of participants but also only in part its negative effect on that of nonparticipants. At the limit, if the program had no effect on participants but still exerted negative externalities on nonparticipants, β_1 could still be positive and significant, and a researcher might draw the wrong conclusion about the positive impact of the measure.

Unfortunately, randomization of the samples of treated and nontreated individuals is not in itself a solution. It ensures, in the best case, that the conditional independence assumption is met. But it cannot obviate economic and social interaction, which leads to a violation of the stable unit treatment value assumption. One strategy for dealing with these equilibrium effects is to make use of any heterogeneity in the size of treatment groups across areas. For if significant equilibrium effects are present, these should be smaller in areas where the program is of smaller scale (the treatment of just a few participants cannot influence the labor market outcomes of several thousand nonparticipants) and larger in areas where the program is of large scale. Implementing this strategy requires the size of the treated group to be unrelated to the labor market outcome under study. If, for instance, the pool of eligible participants is larger in economically depressed areas or larger in areas with many young people entering the labor

market, the researcher cannot pinpoint the equilibrium effect with precision, since it might be blended inextricably with these economic or demographic situations. Hence, even if individuals are randomly selected, the *number* of people who are treated within a local market is most often not itself a random assignment, unless the design of the experiment incorporates that aim. Comparison across markets may thus lead to biased estimates of the equilibrium effects.

To address this issue, Crépon et al. (2013) implemented a two-step randomization: in the first step, the 235 local employment areas were assembled into 47 groups of five agencies that covered areas similar in size and with comparable local populations. Each of these 47 strata was randomly assigned a proportion p of job seekers for eventual treatment in the second step: 0%, 25%, 50%, 75%, or 100%. In the second step, the fraction p of all the eligible job seekers in the area was randomly selected to be assigned to treatment. In the simplest specification, the authors pooled all areas having a positive share of treated job seekers to compare them with areas where no one was treated. The local markets which were randomly selected to feature no treatment can be considered a "super control": they only comprise control group individuals who cannot by assumption be influenced by the program. We can compare individuals in these areas with individuals in other areas by fitting the following equation:

$$y_i = \alpha_2 + \beta_2 \delta_i \pi_i + \lambda_2 \pi_i + \mathbf{X}_i \boldsymbol{\gamma}_2 + \varepsilon_i \tag{14.64}$$

where π_i is a dummy equal to 1 for a person living in a local market with a strictly positive fraction of individuals assigned to treatment and zero otherwise; $\mathbf{X}_i$ now also includes 47 dummies for agencies quintuplets (the randomization strata) on top of individual characteristics and cohort dummies. In this specification, β_2 is the difference between those assigned to treatment ($\delta_i = 1$) and those who are in treatment zones but are not themselves assigned to treatment ($\delta_i = 0$). Parameter λ_2 is the measure of externalities: it measures the effect of being untreated in a treated zone compared to being untreated in an untreated zone. The sum $\beta_2 + \lambda_2$ is the effect of being assigned to treatment compared to being in an entirely unaffected labor market.

The results from this specification are presented in table 14.3 for the sample of those not employed at the time of the assignment. The first line shows that those assigned to treatment are 2.3 percentage points more likely to have an LTFC than those assigned to control status in the treatment labor markets. This gross effect of the program is of roughly the same order of magnitude as in table 14.2 for those not employed. But the net effect of program assignment is not significant anymore. Columns 4 and 5 of table 14.3 show that both the gross effect of treatment and the externalities are stronger and more significant for men than for women. For men equilibrium effects almost fully offset the gross effect of treatment. Focusing on any type of long-term contract yields similar results. These results indicate that a part of the program effects measured with the naive approach, in which equilibrium effects were left out of account, were due to an improvement in the search ability of some workers, which reduced the relative job search success of others. All in all, this study shows that it is indeed important to account for equilibrium effects when evaluating labor market policies.

In principle, social experiments constitute the most convincing approach to evaluating the impact of labor market policies because selection bias is obviated. In practice, this conclusion depends on several hypotheses, and the impact of these on each

TABLE 14.3
The impact of intensive job placement counseling on the employment outcomes of young educated workers, with equilibrium effects incorporated.

Outcome	Variable	Not employed	Not employed Men	Not employed Women
LTFC	Assigned to treatment (β_2)	.023*** (.008)	.043*** (.013)	.013 (.010)
	In a program area (λ_2)	−.013 (.009)	−.036*** (.013)	−.001 (.012)
	Net effect of program assignment ($\beta_2 + \lambda_2$)	.010 (.008)	.007 (.011)	.012 (.011)
	Control mean (α_2)	.16	.131	.177
LT	Assigned to treatment (β_2)	.025** (.012)	.037** (.018)	.019 (.014)
	In a program area (λ_2)	−.021* (.013)	−.043** (.020)	−.010 (.018)
	Net effect of program assignment ($\beta_2 + \lambda_2$)	.003 (.011)	−.006 (.018)	.009 (.016)
	Control mean (α_2)	.365	.372	.36
Observations		11,806	4,387	7,419

Note: The table reports ordinary least squares (OLS) regressions controlling for gender, education, past duration of unemployment and its square, cohort dummies, and 47 dummies for local area quintuplets. The dependent variables are employment outcomes when surveyed eight months after the random assignment: long-term fixed contracts (LTFC) are fixed-term contracts with a length of at least six months; long-term employment (LT) is either a long-term fixed contract or an indefinite-term contract. Column 2 restricts the sample to job seekers who did not report that they were employed at the time of randomization; column 3 restricts the sample to those who did. Standard errors in parentheses are robust to heteroskedasticity and clustered at the local area level. *** significant at the 1% level, ** significant at the 5% level, * significant at the 10% level.

Source: Crépon et al. (2013, table V).

particular experiment must be assessed. In the first place, it must be remembered that a social experiment aims to gain knowledge about a specific measure, so that it may eventually be applied in a "normal," that is, a nonexperimental, context. In other words, it is assumed that the average gain from a measure, as evaluated through a social experiment, is equal or nearly equal to the average gain that will flow from the same measure in a "normal" setting. For that to be true, it is necessary in particular that the mere existence of a random draw does not change the composition of the population agreeing to participate in the experiment.[10] In the second place, we often observe that a significant proportion of the treatment group drops out of the experimental protocol along the way and that an equally significant proportion of the control group is benefiting from services more or less similar to those offered in the program being tested but originating elsewhere. These biases of *attrition* and *substitution* do not disqualify the experimental data because they also exist in nonexperimental data. The assessment of the effects of a measure must simply take them into account appropriately (see Heckman et al., 1999, pp. 1907–1914).

[10]In fact, this hypothesis is sufficient but not really necessary. Heckman et al. (1999, p. 1901) supply two hypotheses, measurably less stringent, for which the experimental data make it possible to obtain unbiased estimators of the average treatment effect.

3.3 Evaluation Based on Observational Data

Setting up social experiments is not an easy task. Apart from the cost of organizing the procedure, which can be substantial, policies are often implemented under pressure and governments often cannot, or do not want to, wait for the results of experiments. Besides, the fact that some individuals are randomly selected to participate in a pilot social program and others randomly excluded can sometimes be difficult to accept at the local level. For these reasons, controlled experiments are still relatively rare in labor economics, and researchers often have to rely on preexisting data.

The econometrician wishing to assess a policy measure generally disposes of data flowing from surveys that give the responses of individuals who have had the benefit of the measure—the treated group—and those of untreated individuals—the control group. These data do not in themselves allow her to distinguish the specific impact of the measure (which is what she wants to know) from the impact of differences that may exist between the *characteristics* of the two groups (which is what she wants to eliminate). In the real world, an individual decides to take part in a program, or benefit from a policy measure, according to his characteristics and personal desires. It is possible in addition that only a portion of the individuals who wish to benefit from a measure are chosen by the agency in charge. Therefore, estimates based on data from surveys are subject to *selection biases*, which the econometrician strives to minimize through appropriate methods.

3.3.1 An Example of Difference-in-Differences: The Impact of a Job Placement Program for Youth

In practice, a number of labor market programs adopt not just one type of intervention but an array of interventions to help workers find a job, which often makes the evaluation of individual components of the program a real challenge. For instance, in dealing with the low-skilled or the long-term unemployed, programs often combine placement services with hiring subsidies, as was the case for the New Deal for Young People (NDYP) introduced in the United Kingdom in 1998 and in effect until 2011. The program was targeted at the 18- to 24-year-old, longer-term unemployed, and participation was compulsory after six months of unemployment. The program comprised a first stage of four months called "Gateway," which included intensive counseling with interviews at least once a week with a personal advisor and small, basic-skill courses. At the end of this period clients could get a wage subsidy for a maximum of six months. In the second stage, four options were offered for those still unemployed: a voucher for subsidized employment in the private sector for six months; a period of paid training or full-time education for up to 12 months; work in the voluntary sector for six months; or a job on the "Environmental Task Force," which was most often a public-service job. From January to March 1998, the program was implemented as a pilot in 12 areas before it was rolled out nationally in April 1998. Since the policy was only introduced in selected areas at the beginning of 1998, it is possible to evaluate the impact of the "Gateway"—enhanced job search assistance—on the chances of finding a job. However, the selected areas, as well as the participants, were not chosen at random, as opposed to the previous example studied by Crépon et al. (2013). The most common strategy in the case of such nonexperimental settings is to make use of areas comparable to those

where the pilot was implemented in order to build a control group, using the difference-in-differences method. Blundell et al. (2004) provide such a framework for evaluating the impact of the NDYP.

Why Not Do an After-Before Difference Assessment?

If we have longitudinal data or repeated cross-section data on the same population, the first idea that springs to mind is to compare the average response of the persons treated *before* and *after* their participation in the program. Let us denote y_B^0 and y_B^1 the employment status of nontreated and treated individuals respectively, at a period *before* the program was implemented. As in the previous section, let δ be a dummy which equals 1 if the individual is assigned to participate in the program. Even with longitudinal data, the econometrician still does not observe the realization of the potential response y_A^0 of this person if he had not undergone the treatment which he did in fact undergo. So, without a supplementary hypothesis, the econometrician cannot infer the average treatment effect, which is $\mathbb{E}\left(y_A^1 - y_A^0 \,|\, \mathbf{X}, \delta = 1\right)$. With longitudinal data, however, the realizations of the response y_B^1 of a representative participant before the application of the program are known. Then a possible identifying hypothesis would be:

$$\mathbb{E}\left(y_A^0 - y_B^0 \,|\, \mathbf{X}, \delta = 1\right) = 0 \tag{14.65}$$

This hypothesis signifies that *for a person having taken part in the program* ($\delta = 1$), the responses if he had not benefited from the treatment would have been the same, on average, before and after the period when the program was applied. *For the participants in the program*, let $\overline{y}_A^T$ and $\overline{y}_B^T$ be the empirical average responses of the treated after and before the period when the program was applied; the "before-after" (BA) estimator of the average gain from the treatment, denoted $\tilde{\Delta}_{BA}$, would then be:

$$\tilde{\Delta}_{BA} = \overline{y}_A^T - \overline{y}_B^T \tag{14.66}$$

This estimator offers the advantage of making it possible to dispense with data on nonparticipants, which is clearly helpful when these data are not available from comparable individuals. If hypothesis (14.65) is satisfied, the estimator $\tilde{\Delta}_{BA}$ is unbiased. But unfortunately, in most cases this hypothesis must be rejected. In the first place, hypothesis (14.65) excludes any influence from unobserved heterogeneity. Suppose for example that there are two classes of workers, the "good" ones and the "bad" ones, such that the productivity of the "good" ones rises between dates A and B independently of their participation in the program (because labor demand shifts in their favor, for example), whereas the productivity of the "bad" ones only rises if they take part in the program. If the fact of being "good" or "bad" is not observed, and if there is at least one "good" worker who takes part in the program, hypothesis (14.65) is not satisfied.

Another reason to reject hypothesis (14.65) is that the global state of the economy and/or the situation of an individual taking part in the program are liable to undergo change between dates B and A. In that case, the estimator will credit the program for successes or failures which are in fact due to macroeconomic and/or life-cycle factors.

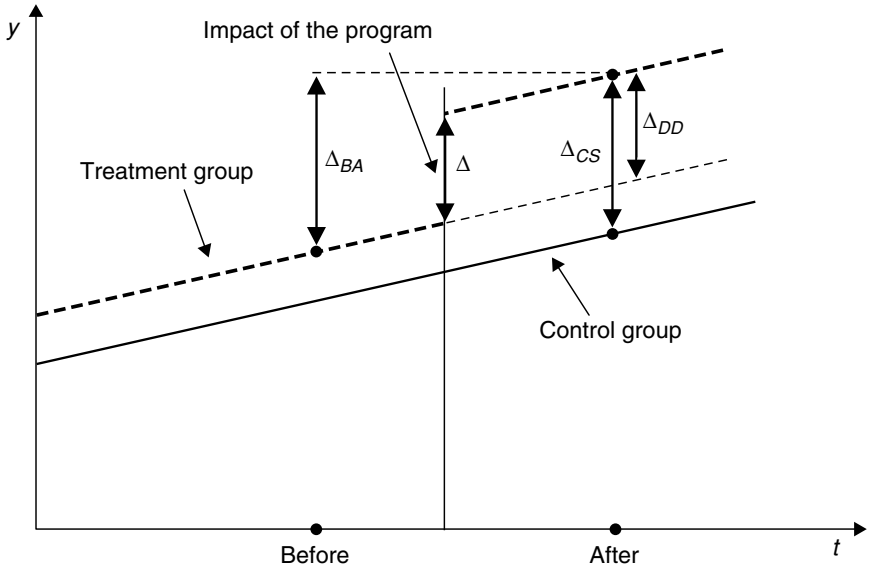


FIGURE 14.11
The difference-in-differences estimator compared with other differences between the treatment and the control groups, when the common trend assumption is satisfied.

Note: Δ is the true impact of the program; Δ_{DD} is the difference-in-differences estimator. The common trend assumption posits that the distance before and after the treatment is the same for the two groups. For the treatment group, the light dotted line represents what would have happened without the treatment, which is never observed. In that case, both the cross section Δ_{CS} and the before-after Δ_{BA} estimators are biased: the cross section is biased because the characteristics of the control and the treatment groups differ, leading to differences in the level of the outcome variable y; the before-after is biased because various factors influence the outcome variable y over time independently of the policy.

This point is illustrated by figure 14.11, which displays the evolution of the outcome for a control group and for a treatment group in the situation where this outcome increases over time, following a common trend and in the absence of the program. The program increases the outcome of the treatment group by an amount Δ. Figure 14.11 shows that the before-after estimator, Δ_{BA}, does not provide a correct estimate of the true effect Δ because it does not account for the common trend in the outcome. The cross-section estimator, Δ_{CS}, defined by equation (14.62) and utilized in the case of controlled experiments, is not appropriate either because the treated and nontreated groups might have characteristics leading to differences in outcome (even if they share a common trend). If present, these differences need to be "differenced-out."

The Difference-in-Differences Estimator

If various factors may influence the odds of employment over the period of the treatment, then we need to control for them. The identifying hypothesis (14.65) signifies that the gain from nontreatment is null for the participants. It says nothing about the value of this gain for nonparticipants. But if we have data for the latter, it is possible to find the average gain from nontreatment for the group of nonparticipants. We can then

postulate that this average gain is the same as that for the group of participants. This identifying hypothesis is written thus:

$$\mathbb{E}\left(y_A^0 - y_B^0 \,|\, \mathbf{X}, \delta = 1\right) = \mathbb{E}\left(y_A^0 - y_B^0 \,|\, \mathbf{X}, \delta = 0\right) \tag{14.67}$$

This equality clearly postulates that the (observed) average gain $\mathbb{E}\left(y_A^0 - y_B^0 \,|\, \delta = 0\right)$ from nontreatment for the nonparticipants is equal to the (unobserved) average gain $\mathbb{E}\left(y_A^0 - y_B^0 \,|\, \delta = 1\right)$ of nontreatment for the participants: had they not been treated, participants ($\delta = 1$) would have had the same gains as nonparticipants ($\delta = 0$). In practice, for this assumption to be credible, it is necessary that participants and nonparticipants be comparable with regard to the outcome under consideration. This conditional independence assumption in the case of the difference-in-differences is often called the "common trend assumption" (adopting the terminology of Blundell and MaCurdy, 1999). It means that the trends that may affect the results of participants *and* nonparticipants are identical, and it can be rewritten:

$$\mathbb{E}\left(y_A^0 \,|\, \mathbf{X}, \delta = 1\right) = \mathbb{E}\left(y_B^0 \,|\, \mathbf{X}, \delta = 1\right) + m_t \tag{14.68}$$

where $m_t = \mathbb{E}\left(y_A^0 - y_B^0 \,|\, \mathbf{X}, \delta = 0\right)$ is the aggregate growth of the outcome variable, here employment, in the nontreated areas. The common trend assumption (14.68) says that participants ($\delta = 1$), had they not benefited from the program, would have had the same increase or decrease in employment as the one observed for the nonparticipants. Note that if we make this assumption, we also assume that there are no externalities specifically influencing individuals in the control group, since we assume that the average difference in employment after-before is not group specific and is common to the treatment group as well. In practice, this common trend assumption requires us to check whether or not, before the program unfolds, the average outcome under consideration evolved in the same manner for participants and nonparticipants. Also, it requires us to check that from the date of the program's commencement, no other factor could have generated any divergence within the average outcomes among participants and nonparticipants apart from the program itself. *For the nonparticipants in the program*, let $\overline{y}_A^C$ and $\overline{y}_B^C$ be respectively the observed average responses after and before the period over which the program is applied. If we remark that $\overline{y}_B^T$ is the empirical estimate of $\mathbb{E}\left(y_B^0 \,|\, \mathbf{X}, \delta = 1\right)$, the difference-in-differences (*DD*) estimator, denoted $\tilde{\Delta}_{DD}$, is defined by:

$$\tilde{\Delta}_{DD} = \left(\overline{y}_A^T - \overline{y}_B^T\right) - \left(\overline{y}_A^C - \overline{y}_B^C\right)$$

Thus the difference-in-differences estimator is equal to the difference between the before-after estimator of the treated group and the before-after estimator of the control group. It can easily be verified that this is an unbiased estimator of the average gain from the program, $\mathbb{E}\left(y_A^1 - y_A^0 \,|\, \mathbf{X}, \delta = 1\right)$, if the identifying hypothesis (14.67) is satisfied. Note that with this hypothesis, the difference-in-differences estimator eliminates the biases due to time-invariant heterogeneity.

In principle, the difference-in-differences estimator has the advantage of being insensitive to changes in the global state of the economy that affect the control group

and the treatment group uniformly. However, this is not always the case in practice. For this reason, the common trend assumption may not be fulfilled if other policies or institutional changes influencing employment, which occur at the same time as the program, could impact the treatment and the control groups differently.

Ashenfelter's "dip" is another example which may imply that the common trend assumption is not satisfied. Ashenfelter (1978) observed that the wages of (future) participants in a training program had a tendency to fall off in the period before they entered the program. Many subsequent studies have confirmed this observation, both in the United States and in certain European countries (see, for example, Regnér, 1997; Heckman and Smith, 1998). If the Ashenfelter dip exists only in the wage profile of treated individuals, the common trend assumption is not satisfied. Heckman and Smith (1998) and Regnér (1997) find that this is the case, so the difference-in-differences estimator overestimates the impact of the program when the Ashenfelter dip is at play.

We may also note that this assumption is not satisfied in the example of the "good" and "bad" workers imagined above if the composition of the group of participants and the group of nonparticipants is not stable over time. Longitudinal data are preferable because they allow us to study the same individuals before and after the measure is implemented and control for time-invariant unobserved heterogeneity. But this estimator can also be implemented based on two cross sections, on the condition that the characteristics of individuals within groups remain stable over time.

Treatment Effects and the Search for Externalities

To implement the difference-in-differences estimator, one simple way is to run the following linear probability model:

$$y_{it} = \alpha + \beta(\eta_t \times z_i) + \lambda z_i + \mathbf{T}\boldsymbol{\delta} + \mathbf{X}_{it}\boldsymbol{\gamma} + \varepsilon_{it} \tag{14.69}$$

where y_{it} is a dummy that equals 1 if the individual is in employment and 0 otherwise, η_t is also a dummy that equals 1 after the treatment has taken place (i.e., after the introduction of the pilot zone—February 1998) and 0 otherwise, z_i is another dummy that equals 1 if the individual belongs to the treatment group (i.e., long-run unemployed persons aged between 19 and 24 and living in a pilot zone) and 0 otherwise, $\mathbf{T}$ is a vector of year dummies that reflects common aggregate time effects, and $\mathbf{X}_{it}$ is a vector of individual characteristics that need to be controlled for, to correct for differences in observable characteristics that could impact the employment probability between individuals and areas registered at the time when eligibility is checked (completion of six months of unemployment). Since y_{it} is a dummy variable, one can also run a logit or probit regression based on this specification, which Blundell et al. (2004) do. But the results do not change qualitatively. The authors make use of the fact that the program has two eligibility criteria—areas and age—to define various control groups.

- In a first specification, they compare the treated (long-term unemployed 19- to 24-year-olds living in pilot zones) to a control group composed of 19- to 24-year-olds with the same unemployment duration and *not* living in the pilot areas. Comparing these two groups allows the researchers to identify the effect of the treatment on the treated. This is a *net* effect, since individual members of the control group all live in other areas and cannot be affected by the intervention (they constitute a "super control"). As shown in the first row of table 14.4,

TABLE 14.4
The impact on employment of intensive job placement assistance for young and long-term unemployed men at the tenth month after starting an unemployment spell.

Treatment group	Control group	Number of observations	Effect on employment probability (β)
19- to 24-year-olds living in treatment areas	19- to 24-year-olds living in all control areas	3,716	.110** (.039)
19- to 24-year-olds living in treatment areas	25- to 30-year-olds living in treatment areas	1,096	.104* (.055)
25- to 30-year-olds living in treatment areas	25- to 30-year-olds living in matched control areas	983	.055 (.058)
19- to 30-year-olds living in treatment areas	19- to 30-year-olds living in all control areas	6,896	.066** (.029)

Notes: The table reports ordinary least squares (OLS) regressions controlling for marital status, sought occupation, region, age, and labor market history (number of unemployment spells). The dependent variable is whether an individual has left unemployment between the sixth and the eighth months of an unemployment spell, among individuals having completed a six-month spell of unemployment which began over a predefined time interval. Standard errors in parentheses. ** significant at the 5% level, * at the 10% level.

Source: Blundell et al. (2004, table 1).

which reports results of the regression for men, the program has improved participants' exits into employment very significantly—11 percentage points—compared with outflow rates in the preprogram period of only 24% of individuals in the treatment group over the similar four-month period. Hence the probability of getting a job rose by about 45% (=11/24). Excluding those who exited the program thanks to a wage subsidy, Blundell et al. (2004) find a lower bound for the "pure" effect of job placement of 4 to 5 percentage points.

- In a second specification, the researchers compare the treated to a control group made of 25- to 30-year-olds with the same unemployment duration but this time living *in* the pilot areas. This setup allows them to measure a *gross* effect of the program, for those young unemployed living in the pilot areas but not eligible for the program might suffer from externalities which could reduce their employment outlooks (or improve them in some cases, as noted in section 2.4) and thus bias the estimated average treatment effect on the treated. If there are significant externalities at play, a contrast ought to emerge between the estimated impact and the one obtained in the previous setting. As shown in the second row of table 14.4, these externalities are probably not very significant in the case of this pilot program, since the result does not change.
- In a third specification, the researchers compare the nontreated, long-term unemployed young (25- to 30-year-olds) in a pilot area to young individuals of the same age and condition not living in any of the pilot areas. In this way, the authors can measure directly whether ineligible youth living in the treated areas are affected by program externalities compared with youth of the same age and same unemployment spell living in other areas. If the common trend assumption between

these two groups is satisfied, any difference should stem only from program externalities. As seen in the third row of table 14.4, no such effect can be found.

- In a fourth specification, all young long-term unemployed (19- to 30-year-olds) in a pilot area are compared to all young people of the same age and condition not living in the pilot areas. This setup estimates the average treatment effect of the program in the "whole" youth labor market. The point estimate is still positive and significant but smaller due to the larger size of the treatment group.

Blundell et al. (2004) run the same regression on women but, like Crépon et al. (2013), they surprisingly do not find any significant impact of this type of program. This is not a general rule though, as other papers studying placement or monitoring programs for young women did find significant effects in other contexts (see Dolton and O'Neill, 2002; Bergemann and van den Berg, 2008).

3.3.2 MATCHING

The crucial identification assumption in the difference-in-differences method is the common trend between the treated and the control groups. This assumption may not be satisfied if the two groups differ significantly in their characteristics and if these characteristics influence the outcome under study. In other words, if the labor market policy is heterogeneous with respect to some observable characteristics—which many probably are—then its estimated effect is an average impact across different effects. That being the case, it is important that the treatment and the control groups be comparable with respect to these characteristics. It is also important that the characteristics of the two groups remain constant over time, before and after the policy is implemented. In the previous examples, this problem was dealt with by controlling the impact of observable characteristics on the outcome variable in a linear and constant manner across groups and over time.

The Technique of Matching

Another approach, labeled *matching,* can help solve this problem in a more flexible way. It consists of extracting from the sample a control and a treated group of individuals *similar* on the basis of *observable* characteristics. Several techniques exist in the literature to make the control group comparable to the treatment group. The more straightforward, which does not require reliance on estimations, is to perform what is called *matching on the covariates*. In this case researchers would try to set a single match for each person in the treated group. The key questions are of course the choice of the vector $\mathbf{X}$ of relevant covariates and what metric to use to measure the distance between two individuals. A common metric is called the Mahalahonis distance, which for two individuals h and i is the square root of $(\mathbf{X}_h - \mathbf{X}_i)\,\widehat{\Sigma}_{\mathbf{X}}^{-1}\,(\mathbf{X}'_h - \mathbf{X}'_i)$.[11] This means that for one individual i in the treated group, the researcher picks out the corresponding individual $h(i)$ who is in closest proximity to i and assigns $h(i)$ to the control group. Another possibility is to match one treated individual i to a weighted average of nontreated individuals most closely proximate to i.

[11] $\mathbf{X}$ is a vector with one row and $\mathbf{X}'$ is the transpose of $\mathbf{X}$. $\widehat{\Sigma}_{\mathbf{X}}^{-1}$ is the inverse of the estimated variance-covariance matrix.

When $\mathbf{X}$ is a vector with discrete values, a simple matching technique is to build cells of treated and nontreated individuals, calculate the average difference in outcome between the two groups in each cell, and calculate a total weighted average of these differences, using the size of cells as weights (see for an example Angrist, 1998). If the conditional independence assumption (14.61) holds, then:

$$\mathbb{E}\left(y_A^1 - y_A^0 \,|\, \mathbf{X}, \delta = 1\right) = \sum\nolimits_x \Delta_x P(\mathbf{X} = \mathbf{x} | \, \delta = 1) \tag{14.70}$$

where $\Delta_x = \mathbb{E}\left(y_A^1 - y_A^0 \,|\, \mathbf{X} = \mathbf{x}, \delta = 1\right)$ is the average treatment effect on the treated (ATT) and $P(\mathbf{X} = \mathbf{x} | \, \delta = 1)$ is the probability to be a participant in a given cell where $\mathbf{X} = \mathbf{x}$.

The Technique of the Propensity Score

Another technique, called *propensity score matching,* uses the propensity score $p(\mathbf{X}) = P(\delta = 1|\mathbf{X})$, which is the probability for an individual to be a participant conditional on his or her observed characteristics. The propensity score matching techniques use the propensity score to match individuals in both groups (Rosenbaum and Rubin, 1983).

The propensity score can be estimated in a first stage with a logit or a probit model. To do this, we regress the dummy of reception of treatment δ_i on the vector of covariates $\mathbf{X}_i$ that determines selection into the treatment: $\delta_i = \alpha + \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i$, where ε_i is an error term. Then the propensity score for an individual i is simply the predicted value $\widehat{\delta}_i$ obtained from this regression, given the observed characteristics $\mathbf{X}_i$ and the estimated coefficients $\widehat{\alpha}$ and $\widehat{\boldsymbol{\beta}}$.

To understand this approach better, suppose we pick a propensity score $p(\mathbf{X})$ at random from the population. Suppose further that we can select two agents from the population sharing this propensity score and that we randomly allocate one of them to the treatment group and the other to the control group. If we make the weak conditional independence assumption $\mathbb{E}\left(y_A^0 \,|\, p(\mathbf{X}), \delta = 1\right) = \mathbb{E}\left(y_A^0 \,|\, p(\mathbf{X}), \delta = 0\right)$, then the expected difference in the observed outcomes for these agents, which defines the average treatment effect on the treated, is:

$$\mathbb{E}\left(y_A^1 \,|\, p(\mathbf{X}), \delta = 1\right) - \mathbb{E}\left(y_A^0 \,|\, p(\mathbf{X}), \delta = 0\right)$$

In practice, we can use propensity scores to match individuals across groups in the same way we used covariates for matching, except that we match on the value of the score instead of the value of covariates. The more straightforward approach to this exercise is to adopt the "nearest neighbor." That means that for a given individual i in the treated group, we pick out a corresponding individual $h(i)$ to be assigned to the control group, such that the differences in the estimated propensity scores between the two individuals is the smallest possible:

$$h(i) = \arg\min_h \left[\widehat{p}(\mathbf{X}_h) - \widehat{p}(\mathbf{X}_i)\right]$$

where $\widehat{p}$ denotes the estimated propensity score (see Wooldridge, 2010, chapter 21). Again, instead of using the nearest neighbor in terms of propensity score, we can use a weighted average of these scores. This is referred to as "kernel-based matching" in the literature: each observation in the treatment group is matched to a weighted sum

of observations in the control group that have similar propensity scores, while giving greater weight to observations with closer scores.

The Limits of the Matching Techniques

Under the conditional independence assumption, whatever the technique used, matching the two groups allows us to employ a simple cross-section estimator (14.62) to assess the impact of the policy measure in question, using the matched groups to compute the sample average $\overline{y}_A^T$ and $\overline{y}_A^C$. Note that this is only possible if we assume that no *unobservable* characteristics are influencing the effect of the policy and if all relevant observable characteristics can be used to match the two groups. Indeed, the aim of the matching method is to eliminate, or reduce as much as possible, selection biases that depend only on the *observable* characteristics of individuals; hence it assumes that agents' decisions to take part in a program, and their responses, depend mainly on their observable characteristics. If the econometrician has reason to believe that this is not the case, then he should try whenever possible to apply the difference-in-differences method to the matched groups based on panel data.

The matching method does have one major limitation though: since it is necessary to have overlap (also called "common support") between the two groups in terms of characteristics, a number of individuals in either group might be eliminated from the analysis at the end of the matching process because researchers must limit themselves to covariate values for which both treated and control values exist. So, in practice, matching can only measure the effect of the treatment within the common support region. As a result, the estimated treatment effects may vary substantially based on the matching method chosen and the set of characteristics selected. Apart from demographic variables, such as age, gender, and education, the inclusion of a variable describing the labor market experience of individuals (e.g., the number and duration of unemployment spells, the characteristics of former employers, etc.) can improve the quality of matching (see Heckman, Ichimura, and Todd, 1997, 1998; Lechner et al. 2011).

Regressions or Matching?

Basically, matching amounts to defining cells of observations in which individuals have homogeneous characteristics: a specific treatment–control comparison is then carried out for each cell and together they are weighted to produce an overall average treatment effect. The method is in principle more flexible than the regression approach, in that it allows the effect of the policy to vary across cells. But regressions that include covariates can also be viewed as a species of matching with specific weights. So overall and in principle, the regression, matching, and propensity score matching approaches should not produce major empirical divergences, especially when the sample is large. This is the case notably for the study of Blundell et al. (2004) presented above on the impact of NDYP in the United Kingdom on youth employment. Table 14.5 shows that the estimates with the linear regression including covariates are indeed very close to those obtained with propensity matching. (See other examples in Caliendo and Kopeinig, 2008; Angrist and Pischke, 2009, chapter 3, pp. 69–91; and Wooldridge, 2010, section 21.3. These contributions also show the results of the various techniques on the assessment of the impact of a training program.) Most statistical software packages offer modules to perform matching.

Table 14.5
The impact of intensive job placement assistance for young and long-term unemployed men on employment at the tenth month after starting an unemployment spell using linear matching (regression with covariates) or propensity score matching.

Treatment group	Control group	Obs.	Effect on employment probability (β)	
			OLS regression with covariates	Propensity score matching
19- to 24-year-olds living in treatment areas	19- to 24-year-olds living in all control areas	3,716	.110** (.039)	.104** (.046)

Note: See table 14.4. Standard errors in parentheses. ** significant at the 5% level.

Source: Blundell et al. (2004, table 1).

3.3.3 Duration Models: The "Timing of Events" Approach

Sometimes very detailed information on the duration of unemployment spells and on the characteristics of the unemployed persons in question is available to researchers. This type of data allows us to use the "timing of events," meaning in our case the timing of entry by a client into an active program, in order to identify the causal impact of interventions. This approach is also well suited to estimating their effect not only on employment or unemployment status at a given point in time but also on the duration of time spent in registered unemployment. The idea is to use the random component that determines the timing of entry into a given program in order to build a control group without having to rely on any external factor influencing the treatment, such as an age threshold or any other aspect of the program regulations that might exclude some groups of workers from participation. This for instance is the strategy that Sianesi (2004) adopts to estimate the impact of active labor market programs in Sweden in the 1990s. We detail her analysis below to clarify the method.

The Conditional Independence Assumption

Sianesi (2004) analyzes the impact of any entry into any type of program (training, public-sector employment, subsidized jobs in the private sector, programs for specific groups, etc.) on a number of outcomes, such as the employment and unemployment status of individuals, from 1994 to 1999. In 1994 Sweden was in the trough of a deep recession, which caused unemployment to reach 13.5% the same year. At that time, programs had a maximum duration of six months and participants remained enrolled for four months on average. Before 2001 participation in a program would renew job seekers' eligibility for generous unemployment compensation and was therefore likely to reinforce the work disincentives associated with the benefit system.

Sianesi uses very rich, longitudinal administrative data on both program participation and benefit receipt, which provide each person's labor market status information over time (unemployed, on a given program, employed, etc.), together with a set of individual characteristics including previous working conditions. She was able to follow individuals over a maximum of five years. Restricting herself only to individuals who became unemployed for their first time in 1994, aged 18 to 55, and not disabled, she came up with a sample of 116,130 individuals, followed from the moment they

registered in 1994 to the end of 1999. The data cover 60 months, denoted from $t = 0$ to $t = 60$.

Let us consider a given outcome variable, say employment status, and let us denote $y_t^{1(u)}$ the potential outcome at time t for an individual with a duration of unemployment u, if he has joined a program in his uth month of unemployment, and $y_t^{0(u)}$ if he has not enrolled in any program at least up until the u month point. The maximum unemployment duration is $u_{\max} = 18$ months, which covers 94% of the sample. At any time, individuals can be either in benefit-compensated or noncompensated unemployment, enrolled in a program, in employment, in education, or inactive. Then, the average impact at time t, for those joining a program in their uth month of unemployment versus persisting in the unemployed state (irrespective of whatever happens in the future, including program participation), is:

$$\Delta_t^u = \mathbb{E}\left(y_t^{1(u)} - y_t^{0(u)} | \mathbf{X}, \delta^u = 1\right)$$

where δ^u is a dummy that equals 1 if the individual is joining a program at the uth month of unemployment, and $\mathbf{X}$ is a vector of observable characteristics. The problem, again, is that the counterfactual situation $E\left(y_t^{0(u)} | \mathbf{X}, \delta^u = 1\right)$ is not observed by the econometrician. But it is possible to use instead the observed outcome of those who have the same duration of unemployment (u months) and did not enter any program that month, if we assume the following condition:

$$\mathbb{E}\left(y_t^{0(u)} | \mathbf{X}, \delta^u = 1\right) = \mathbb{E}\left(y_t^{0(u)} | \mathbf{X}, \delta^u = 0\right) \tag{14.71}$$

This equality is analogous to the conditional independence assumption (14.61). It means that conditional on the vector $\mathbf{X}$ of observable characteristics, the average effect of nontreatment ($y_t^{0(u)}$) is the same for those who joined the program at the uth month of unemployment ($\delta^u = 1$) and for those who did not ($\delta^u = 0$). In other words, condition (14.71) means that entering a program or not entering a program at month u of unemployment is *random*, not driven by any factor that could have an influence on the employment outcome under consideration, once we have controlled for all the relevant observables. Sianesi argues that an individual's decision to participate in a program largely depends on that individual's subjective likelihood of finding employment, which can be predicted quite well for those who register as unemployed for the first time with the information she has in hand. Condition (14.71) means that nonparticipants do not turn down a program offer at the uth month because they anticipate a better offer in the future; otherwise, program participation would not be independent of unemployment status. Such dependency would also create an Ashenfelter dip problem due to reduced job search prior to participation. The fact that there was a continuous flow of different programs on offer made it less likely that unemployed job seekers would, on any individual basis, be able to predict future program opportunities.

The Average Treatment Effect

If condition (14.71) is met, we can identify what would have happened to new program participants who entered a program at the uth month if they had not participated, by

simply observing the outcome of those with the same duration of unemployment at the date of entry into the program, but who did not enter any program—conditional on their having the same characteristics as new participants. Of course, for this assumption to be credible, (i) we need to have a good set of characteristics to control for, (ii) we need large cohorts of participants and nonparticipants, and (iii) we need several "events," that is, several dates of entry into a program. If these needs are met, as they are in Sianesi (2004), then the overall effect of programs can be estimated. Since this strategy assumes that the control group was selected at random, in order to calculate the average effect of treatment we need only compare the average outcomes of the two groups at a given point in time. Thus we revert to the cross-section estimator (14.62). This means that we assume that there are no unobservable characteristics which might drive the decision to enter any program at the uth month and which might also influence the outcome. Because the treated group has in fact been divided into $u_{\max}$ subgroups, where $u_{\max}$ is the maximum preprogram unemployment duration, then the average treatment effect of program participation is a weighted average of the treatment effects Δ_t^u of those groups, weighted according to the observed month of placement distribution of the treated:

$$\mathbb{E}\left(\Delta_t^u | \mathbf{X}, \delta = 1\right) = \sum\nolimits_{u=0}^{u_{\max}} \Delta_t^u P\left(\delta^u = 1 \mid \mathbf{X}, \delta = 1\right)$$

where δ is a dummy that equals 1 if an individual has entered a program at any duration of unemployment, and $P\left(\delta^u = 1 \mid \mathbf{X}, \delta = 1\right)$ is the relative size of the cohort that entered into a program at the uth month compared with the total size of the group of program participants. To ensure maximum comparability between the treated and the control groups, Sianesi uses the "nearest neighbors" propensity score matching presented above over the common support of a large set of characteristics, including those that may describe the individual's past employment history and current employment prospects.

An Illustration: The Impact of Active Labor Market Programs in Sweden in the 1990s

Figure 14.12 shows one of the main results of Sianesi's study. Although on average joining a program initially reduces the chance of finding employment by up to 4 percentage points (which is evidence of a lock-in effect), from the fifth month forward participants perform significantly better than their nontreated counterparts, displaying significantly higher and increasing employment rates over time. Over the first five years from the start of any program, the treated are seen to enjoy an employment probability 6% higher on average. Sianesi further shows that while taking nonresponses (individuals "lost" and ones whose employment status is unknown) into account would probably halve this impact, it would still remain significantly positive. She also finds that the effect on employment is greater for those entering sooner rather than later into active programs. Moreover, Sianesi analyzes the impact of program participation on the duration of benefit payments. Because, at that time, program participation allowed clients to renew their benefit entitlements, she finds that the probability of later benefit receipt outside of any program is also increased by program participation, especially for individuals close to or at the deadline of benefit exhaustion. This is evidence of the work disincentive embedded in the institutional setup of the Swedish active programs at that time.

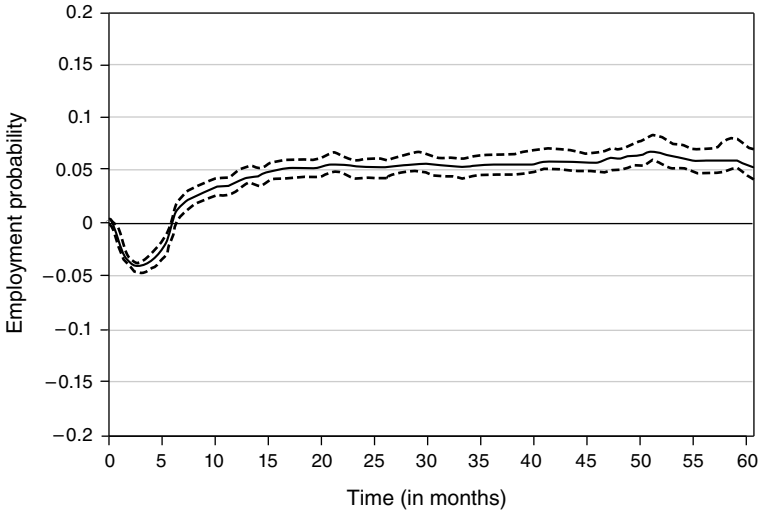


Figure 14.12
The impact of active program participation on employment probability over time in Sweden in the 1990s.

Note: The horizontal axis corresponds to the time in months elapsed since entry into program. The confidence interval at 95% is represented above and below the estimate.

Source: Sianesi (2004).

4 THE MAIN EMPIRICAL RESULTS

The evaluation of postprogram effects has been a thriving field of research over the last 20 years, notably in the developed economies, and especially in Europe, where spending on active labor market policies represents a significant drain on the public purse. Hence it will be worth presenting meta-analyses which provide an overview of the results of these impact evaluations. We then come back to detailed evaluations of different kinds of active labor market policy in order to get a more fine-grained picture of the efficacity of each of them. Thus we review the main results of empirical research on job search assistance, training, hiring subsidies, and temporary public job creation. In conclusion we focus on what can be learned about the equilibrium effects of some of these policy measures.

4.1 An Overview: Results from Meta-analysis

Hundreds of evaluations have been published over the last 20 years that shed light on the impact of labor market policies in varying contexts and for various types of worker. But evaluation methods vary widely across publications (experimental, quasi-experimental, etc.), and so do the outcomes measured (unemployment, employment, gains) and the time horizons (short-run, long-run). All this makes comparisons difficult to achieve, which is why meta-analysis can be useful to disentangle the results.

4.1.1 Two Recent Meta-analyses on Labor Market Policies

Meta-analysis methods assemble results from many studies, as well as methodological and contextual information on these studies, in a database in order to identify the impact of one type of program. Meta-analysis was originally developed in the field of health care studies, where it is usually employed to generate robust evidence on the effectiveness of a given drug by combining the data from a large set of experiments while controlling for the method used and laboratory conditions. In health studies the sample size of experimental trials is usually small, and meta-analyses can produce more precise estimates. In its simplest form, the method comes down to identifying a common measure of effect size and then building a weighted average across studies. In the social sciences, empirical evidence is often based on large samples, but controlled experiments are scarcer than in health care studies, and their results are sensitive to the environment in which they are implemented. Moreover the majority of studies are based on natural experiments where the reliability of the results may be weaker due to the conditions of identification. Hence it is important to gather information from a wide variety of studies while controlling for covariates that may influence their results.

Two recent meta-analyses of labor market policies are those of Card et al. (2010) and Kluve (2010). Card et al. (2010) based their analysis on 199 program impact evaluations from 97 studies conducted between 1995 and 2007 in 26 OECD countries. Kluve (2010) restricted the analysis to 137 program evaluations from 19 European countries but over a longer period (1983 to 2006). Table 14.6 taken from Card et al. (2010) shows that evaluations of labor market policies based on experimental design are quite rare. Most studies use observational data with a longitudinal dimension. Matching is also frequently used to correct the selection bias in observational data. As noted above, most of these studies identify program effects in partial equilibrium, meaning that the potential equilibrium effects are rarely taken into account.

The most difficult task for meta-analyses, especially in labor market studies, is to define a standardized measure of program impact because the measurement of

Table 14.6
Evaluation methods used in program effect evaluations, 1995–2007.

	Overall sample	Austria, Germany, and Switzerland	Nordic countries	Anglophone countries
Basic methodology (%)				
Cross sectional with comparison group	3.0	0.0	5.7	0.0
Longitudinal with comparison group	51.3	80.6	30.2	75.0
Duration model with comparison group	36.2	19.4	43.4	0.0
Experimental design	9.1	0.0	18.9	25.0
Covariate adjustment method (%)				
Matching	50.8	73.1	30.2	45.0
Regression	42.7	26.9	52.8	40.0

Note: Percentage of estimates on a total of 199 estimates of treatment effect drawn from 99 studies.

Source: Card et al. (2010, table 4, p. F461).

outcomes varies across studies. Some studies report treatment effects on the exit rate from registered unemployment or on the probability of unemployment registration in the future, some measure the gains (wages and other sources of income) generated in the labor market, while others (more rarely) report the effects of the program on the probability of employment at some date after the completion of the program. This makes it almost impossible to build a standardized effect size estimate for each study in the sample. For this reason Card et al. (2010) and Kluve (2010) classified the estimates based on sign and significance into three categories: significantly positive, insignificantly different from zero, and significantly negative.[12]

Nevertheless, Card et al. (2010) and Kluve (2010) are able to identify the sign and significance of the program impact at three points in a large number of cases: a short-term impact at approximately one year after completion of the program, a medium-term impact roughly two years after program completion, and a long-term impact roughly three years after program completion. Then their empirical strategy is to run an ordered probit model to fit the sign/significance of estimated program impact on a set of study characteristics. The ordered probit is a generalization of the probit model to the case of more than two outcomes of the dependent variable (for example, -1 for a significant negative impact estimate, 0 for an insignificant impact estimate, and +1 for a significant positive impact estimate; see Wooldridge, 2010, chapter 16, for more detail on this method).

Basically it comes down to regressing the probability that a given type of program yields a positive, negative, or insignificant effect, given its characteristics and the characteristics of the evaluation method used to assess it. Card et al. (2010) include in the set of regressors the type of program (training, placement, job subsidies in the private sector, public employment, and other types of interventions), the square root of the size of the sample used, and other covariates to control for the evaluation design and the type of participants. Their results are presented in table 14.7.

Kluve (2010) controls as well for the institutional and macroeconomic contexts, and the timing of the evaluation, but does not distinguish studies by the horizon of impact evaluation. His results are presented in table 14.8. In the regressions that make it possible to build the tables 14.7 and 14.8, program types are identified by dummies, and one type of program must be omitted for the model to be identifiable. Thus the coefficients are relative to the omitted program (i.e., more or less effective than "mixed interventions and other programs" in table 14.7 and more or less effective than "training" in table 14.8).

4.1.2 What These Meta-analyses Teach Us

Although their samples differ, these two meta-analysis studies yield quite consistent conclusions:

- Job creations in the public sector are more often ineffective than other interventions and even appear detrimental, with negative treatment effects. This is consistent with earlier literature reviews, including ones such as Heckman

[12] Other studies, such as Greenberg et al. (2003) in the case of training programs, and based on a much more homogeneous and larger set of experimental-based evaluation studies in the United States, are able to use both effect size (and not just its sign) and standard error in the outcome variable to build a standardized measure of program effect.

TABLE 14.7
The effectiveness of labor market programs in Europe, 1983–2007.

Type of program (omitted: mixed and other)	Short-term treatment effect		Medium-term treatment effect	
	Coefficient estimate	*t*-stat	Coefficient estimate	*t*-stat
Public job creation	–.31	(−.67)	−.46	(–.62)
Private-sector subsidy	–.14	(−.33)	.79	(.86)
Placement	.72	(1.63)	1.16	(1.36)
Training	.22	(.57)	1.14	(1.68)
Observation	181		92	

Note: Models are ordered probits. The dependent variable is a categorical variable indicating whether the estimate of the program effect is negative (−1), insignificant (0), or positive (+1). Controls include the square root of the size of the sample used, the duration of the program, the type of measurement in the study (e.g., time in registered unemployment, other type of duration, postprogram earning), the age and gender of participants when they are pooled, a dummy for experimental designs, a dummy for published studies, dummies for the type of participants (registered unemployed, long-term unemployed, or any other disadvantaged group), and country group dummies (one for English-speaking countries, one for Nordic countries, one for Austria, Germany, and Switzerland, and one for the other countries represented in the sample). *T*-stats of the coefficient estimates are reported in adjacent columns based on standard errrors (clustered by study).

Source: Card et al. (2010, tables 7 and 8, pp. F468–F469).

TABLE 14.8
The effectiveness of labor market programs in Europe, 1983–2007.

Type of program (omitted: training)	Negative treatment effect		Positive treatment effect	
	Marginal effect at sample mean	*t*-stat	Marginal effect at sample mean	*t*-stat
Public job creation	.17	(1.99)	−.25	(−2.25)
Private-sector subsidy	−.15	(−4.00)	.31	(3.34)
Placement and sanctions	−.20	(−3.69)	.44	(4.29)
Young workers programs	.16	(2.19)	−.24	(−2.39)
Observation	137		137	

Note: Models are ordered probits. The dependent variable is a categorical variable indicating whether the estimate of the program effect is negative (−1), insignificant (0), or positive (+1). Table entries document the marginal effects (evaluated at the sample mean) from the corresponding ordered probit regression for the negative and positive outcomes respectively, i.e., the difference in the predicted probability for achieving a negative (positive) treatment effect which arises from changing an indicator among the explanatory factors from 0 to 1. Controls include the type of research design (experimental, etc.) and timing of study, labor market institutions, marcoeconomic context (unemployment, GDP growth, ALMP spending), and country dummies. *T*-stats of the marginal effects are reported in adjacent columns. The underlying standard errors adjust for clustering by study.

Source: Kluve (2010, table 4, p. 911).

et al. (1999). Among Europe-based studies, evaluations of public employment programs are around 25 percentage points less likely to estimate a significant positive impact than training programs and 17 points more likely to report a negative impact (see table 14.8).

- Job search assistance, sometimes associated with sanctions for noncompliance, has a favorable impact, notably in the short term. This is also true, although to a lesser extent, for private-sector subsidies, notably among Europe-based studies. Kluve (2010) finds that evaluations of placement assistance and private-sector subsidies have a probability higher by 50 and 30 percentage points, respectively, of estimating a significant positive impact than do training programs.
- Long-run impacts are generally more often positive and significant than short-run impacts (Card et al., 2010). Indeed, many programs with insignificant or even negative impacts after only a year have significantly positive impact estimates after two or three years. This is particularly true for classroom and on-the-job training programs, which appear to be particularly likely to yield more favorable medium-term than short-term impact estimates.
- The context of programs matters less than their type when it comes to explaining their effectiveness (Kluve, 2010). Stricter employment protection may be associated with worse performances by labor market programs, while higher unemployment rates are usually associated with better impacts from these programs, but the sizes of these effects are small. The fact that high unemployment is associated with more positive effects does not hold if studies from the 1980s are eliminated. This association could reflect a creaming effect, by which the most employable unemployed persons are the earliest entrants into programs when unemployment is high. As well, the macroeconomic situation (measured by GDP growth for instance) seems to have no impact on the effectiveness of labor market programs.
- Evaluations (including randomized experiments) that measure outcomes based on time spent in registered unemployment appear to show more positive short-term results than evaluations based on employment or earnings (Card et al., 2010).

These meta-analyses supply an overview of the evaluation impact results, but they must be complemented by information gleaned from research carried out on each of the measures falling into the category of active labor market policy.

4.2 Job Search Assistance and Monitoring

Job search assistance programs have been the object of numerous evaluations in the United States and in Europe. Research of a more narrowly focused kind has been done on the impact of threats and sanctions as specific components of job search assistance.

4.2.1 North American Evaluations

Job search assistance programs generally consist of interviews with job seekers to guide them in their efforts to find work. Starting in the 1980s in the United States, these programs have been evaluated independently of other interventions on the basis of social experiments targeted at unemployment insurance recipients. The help given to the job seekers in the treatment groups is briefly summarized in table 14.9. Its impact on the average duration of unemployment is presented in table 14.10. It is apparent that this help significantly reduces the duration of unemployment. Further, Meyer (1995) stresses

TABLE 14.9
Experiments with help in job searching carried out in the United States in the 1970s and 1980s.

Place	Type of help
Nevada (1977–78)	Weekly interviews, checks on eligibility
Charleston (1983)	2 in-depth interviews and a 3-hour session on job searching
New Jersey (1986–87)	Obligation to contact the employment agency regularly, offer of training
Washington (1986–87)	Intensive job search activities
Wisconsin (1983)	6-hour job search workshops

Source: Meyer (1995, tables 4a, 4b, pp. 111–112).

TABLE 14.10
Effects of job search experiments on weeks of benefits, measured as treatments minus control.

Place	Weeks of benefits
Nevada (1977–78)	−3.90 (0.41)
Charleston (1983)	−0.70 (0.39)
New Jersey (1986–87)	−0.50 (0.25)
Washington (1986–87)	−0.47 (0.28)
Wisconsin (1983)	−0.62 (0.43)

Note: Standard error in parentheses.

Source: Meyer (1995, tables 5a, 5b, pp. 115–116).

that it generally leads to a reduction in the total expenditure of the bodies administering unemployment insurance, inasmuch as the benefits that flow from this help outweigh its costs. It should nevertheless be noted that these experimental situations mingle help for the unemployed with surveillance of the search effort they are making. The contribution of each of these components is generally difficult to isolate.

The study of Black et al. (2003) on the program of job search help which the state of Kentucky set up in 1993 confirms the results of Meyer (1995) and those of several studies of European programs. This program lends itself to a natural experiment, since participation is in principle compulsory for all unemployed persons, but because of the limited capacities of the employment agencies only a portion of the unemployed are actually enrolled. More precisely, in Kentucky, unemployment insurance claimants are assigned profiling scores predicting unemployment spell duration. Among those with higher scores (predicting long spell duration), random assignment allocates only a fraction to mandatory program participation, while other unemployed persons with the same scores but who have been excluded from participation constitute a control group. Black et al. find that on average the treatment group receives unemployment benefits for a period shorter by 2.2 weeks than the control group does. This study also shows that the rate at which the treatment group returns to work rises sharply during the interval between notification of (compulsory) participation in the program and the date at which it actually begins. In other words, the disagreeable prospect of having to

have regular contact (two to three hours per week) with the employment agencies, and of having them check on one's job search effort, is enough to quickly force those who are not experiencing any real difficulty in finding work out of the unemployment insurance system. As a general rule, the establishment of surveillance and counseling programs has the effect of exerting pressure on a percentage of the eligible unemployed.

4.2.2 European Evaluations

In an experimental setting with Danish job seekers, Graversen and van Ours (2008) confirm the results of Black et al. (2003). They highlight a perceptible rise in exits into employment prior to enrollment in the program. This spike generally follows the dispatch of a letter reminding job seekers that the program is obligatory, on pain of having their benefit payments suspended.

Table 14.11 gives some partial indications of the effect of active labor market policies in Europe. The study of Björklund and Regnér (1996) looks at a social experiment in which the services delivered to the unemployed in 1975 in a small city in central Sweden were intensified. For three months the 216 unemployed persons in the treated group received intensive job search assistance of 7.5 hours per week, while the 194 in the control group received normal assistance of around 1.5 hours per week. Nine months

TABLE 14.11
Estimated effects of labor market policies in Europe.

Studies	Country	Type	Responses	Impact
Social experiments				
Björklund and Regnér (1996)	Sweden	JSA	Employment rate	13*
			Monthly wage	6
Dolton and O'Neill (1996)	UK	JSA	Employment rate	4
Torp et al. (1993)	Norway	Training	Employment rate	3
van den Berg and van der Klaauw (2006)	Netherlands	JSA	Exit rate to employment	6
Observational data				
Westergard-Nielsen (1993)	Denmark	Training	Male hourly wage	1
Dolton et al. (1994)	UK	Training	Male hourly wage	26
			Female hourly wage	–8
Main and Shelly (1990)	UK	Training	Youth employment rate	11*
			Wage	32
Björklund (1994)	Sweden	Training	Youth employment rate	8*
			Wage	10*
Sianesi (2002)	Sweden	Subsidies	Employment rate	40*
Jaenichen and Stephan (2011)	Germany	Subsidies	Share in regular employment	25–50*

Note: JSA: Job search assistance. Estimated variations of consecutive responses to the program expressed in percentage points for employment rates and percentages for wages; an asterisk indicates a significant result at the threshold of 5%.

Source: Heckman et al. (1999, table 25, pp. 2070–2075) and Kluve (2010, table 2, pp. 908–909).

after the experiment, the percentage of persons in the treated group who had found a job was higher by 13 points than that of persons in the control group.

Dolton and O'Neill (1996) studied the impact of the Restart placement program in the United Kingdom in 1989. This program had been introduced in 1987 with the purpose of helping the long-term unemployed. Individuals unemployed for six months were contacted and given six monthly interviews, each lasting about 15 to 25 minutes, with a counselor who attempted to improve their job search strategies and who could initiate contacts with possible employers. Persons who refused this program lost their unemployment benefits. Dolton and O'Neill have experimental data, for in 1989 the authorities set up a random sample of individuals summoned to the interviews. Individuals not summoned form the control group, but they could ask to take part in these interviews. The method adopted by Dolton and O'Neill is to compare the performance of the beneficiaries of the Restart program with that of individuals belonging to the control group. They find that the exit rate from unemployment of the control group is from 20% to 30% lower than that of the treatment group during the six months subsequent to the missed interview. After one year, the beneficiaries had an average employment rate 4% higher than that of the nonbeneficiaries.

In another analysis of the Restart program, Dolton and O'Neill (2002) examine the long-term effects of this program. They find that five years after the program began, the unemployment rate for men in the treated group was 6 percentage points lower than it was for men in the control group. In contrast, they find no significant long-term improvement for women. From this they conclude that for the men, the sanctions triggered by nonparticipation in the interview count for more in the short-term effects of the program, whereas the services to help them in their job searches, of which they are informed when they attend the interviews, have more weight in the long term.

It should be noted that empirical studies do not systematically find a positive impact of counseling on the entry rate into employment, at least not for all groups. For instance, in examining the impact of a randomized experiment, van den Berg and van der Klaauw (2006) find that counseling and monitoring did not affect the exit rate to work in the Dutch unemployment insurance system at the end of the 1990s. The monitoring of relatively well-qualified individuals in favorable macroeconomic conditions leads to substitution of search methods and small net effects on the exit rate to work: these individuals resorted less to the sort of informal methods that were not observable by the agency (social networks, checking the newspapers) and more to formal methods of the sort that the agency could observe, especially the use of the services offered by the agency itself. However, individuals with worse prospects may have less scope for such substitution, and monitoring of their search activity may lead to an increase in the exit rate to work.

Fougère et al. (1999) have studied the impact of job search assistance in France in the period 1986–1988, using a job search model with endogenous search effort. In the theoretical model, job search assistance exerts an ambiguous effect on search effort and on the exit rate from unemployment, since the intensity with which personal searches are carried out declines when job search assistance plays a larger part. Nonetheless, econometric estimates suggest that public placement services do have a positive impact on the exit rates from unemployment of disadvantaged individuals, in other words poorly trained youth, and women. Crépon et al. (2005) study the effects of intensive counseling schemes introduced in France in 2001 that are provided to about

20% of the unemployed. Using duration models and a very rich data set, they find significant favorable effects on both unemployment duration and recurrence, but the impact on unemployment recurrence is stronger than on unemployment duration. In particular, the program shifts the incidence of recurrence, one year after employment, from 33% to 26%.

Studies evaluating job search assistance programs are now becoming numerous. These studies suggest that the specific activity of counseling the unemployed exerts a positive effect on the employment rate and, more weakly, on the hiring wages of those who benefit from it.

4.2.3 Threat Effect and Sanctions

When participation in labor market programs is compulsory, for instance after a number of months of registered unemployment, they can incite workers to search more actively for a job to avoid having to attend the scheduled sessions. Neglecting this preprogram "threat effect" may lead to underestimating the true effect of active labor market policies on unemployment exit rates because the counterfactual situation is altered when the threat effect is taken into account. Rosholm and Svarer (2008) were able to estimate specifically the impact of the risk of program participation in Denmark over the years 1992–2002. They conclude that mandatory participation in active labor market policy programs does shorten unemployment duration, even if actual program participation does not! The magnitude of this effect is quantified at three weeks on average. Within the OECD, only five countries have compulsory program participation after a defined spell of unemployment (Australia, Denmark, the Netherlands, Sweden, and the United Kingdom). But participation can become compulsory in many other countries upon referral by the counselor in charge (see OECD, 2007, table 5.5). These results are confirmed by Geerdsen (2006), who finds effects on exit rates comparable in size to the effect of benefits exhaustion found in studies on the U.S. unemployment insurance system.

Sanctions might also have an impact on social assistance recipients, not exclusively on the beneficiaries of unemployment insurance. For instance, in the Netherlands welfare recipients can be sanctioned if they do not comply with job search requirements (for those who are able to work). Van den Berg et al. (2004) identified the impact of such sanctions for the city of Rotterdam when they were introduced in the 1990s. They estimate a duration model and find that a reduction of 20% in the welfare payment to job seekers over a two-week period, imposed as a sanction for not adhering to job search rules, doubles the exit rates from unemployment of the individuals thus sanctioned. Further, they find that these effects persist beyond the two-week period. They also conclude that the earlier the sanctions kick in over the course of the welfare spell, the lower the probability of becoming long-term dependent. The impact of sanctions, ex ante (the threat) and ex post (the effective application of the sanctions) are also confirmed in an experimental setup by Boone et al. (2009). They find that the ex ante effect of programs is stronger than the ex post effect. Van der Klaauw and van Ours (2013) also find that sanctions were more effective than reemployment bonuses in shifting the unemployed from welfare to work more quickly in Rotterdam.

The threat effect is closely linked to the possibility of sanctioning benefit recipients in case of noncompliance, as seen in chapter 5. The contribution of Lalive et al. (2005) relies on Swiss data covering the ensemble of those who entered into

unemployment between September 1997 and March 1998. Job seekers are observed up until May 1999. Lalive et al. were able to identify specifically the effect of warning and administering sanctions in Switzerland, where job search requirements are monitored (through the number of job applications) and where program participation can be mandatory. The available data are sufficiently precise to allow them to distinguish the impact, on the ensemble of the unemployed, of the threat of being sanctioned—the ex ante effect—from the impact of the sanction on the unemployed persons effectively sanctioned—the ex post effect. Lalive et al. find that the two effects exert similar pressure on search efforts, and that both contribute to reducing the duration of unemployment.

The positive impact of sanctions on the exit rate from unemployment is the object of a relative consensus in the literature, but two recent studies suggest that the quality of the jobs found might be affected by them in some cases. Using the Swiss data just mentioned, Arni et al. (2009) find that the threat of being sanctioned (the ex ante effect) reduces the wages of future jobs, while sanctions effectively applied (the ex post effect) reduce both the wages and the duration of the jobs accepted. On data covering job seekers in Sweden between 1999 and 2004, van den Berg and Vikström (2009) find that the hourly wages and the number of hours worked are, on average and all other things being equal, weaker for the unemployed who were sanctioned than for others. They also hold less highly qualified jobs.

4.3 Training Programs

Among the active labor market policy programs for the unemployed, training programs are probably the most costly and the most difficult to gauge in terms of impact because the gains can be significant only in the medium and longer term. The long-run impact on the probability of employment depends on the adequacy and quality of the training, while the short-run impact is usually lowered by "lock-in" effects (participants stop looking for a job during the program and thus remain unemployed longer than nonparticipants). The long-run impact on wages appears ambiguous and most often not significant, except when the training program takes a long time to complete and is targeted at particularly disadvantaged categories of worker.

4.3.1 Short-Term vs. Long-Term Effects on the Employment Rate

In general, European studies find that training programs have a significant positive effect on the employment rate of the beneficiaries. With observational data, Main and Shelly (1990) and Björklund (1994) arrive at high figures, whereas the study of Torp et al. (1993), which reports on a social experiment carried out in Norway in 1991, with training periods of around five months, finds that this training had no more than a very slight effect on the probability of being employed twelve months later. More recently, Crépon et al. (2012) evaluate the impact of training programs for the unemployed in France and find that they have no impact on the exit rate from unemployment in the short run. Using a longitudinal data set from the unemployment insurance system, and controlling for both observed and unobserved heterogeneity, they also find a significant and favorable effect of training on the duration of the subsequent employment spells. This confirms results obtained by Winter-Ebmer and Zweimüller (2003) in Austria for men.

Training programs generally have a negative effect on exits into employment in the short term, since those being trained are not immediately available to take up paid work. The empirical analysis of Lechner et al. (2011), focused on Germany and making use of a very detailed administrative database that extends over 10 years, finds negative employment effects in the short term for all program types, effects whose magnitude and persistence increase with program duration. In the longer term, these negative effects are offset, and training seems to increase employment rates by 10 to 20 percentage points; this impact appears to be sustainable over time. Lechner and Wunsch (2009) show further that the lock-in effects of training programs are significantly weaker in periods of recession, since even if the participants do engage in an active job search while enrolled, they have fewer opportunities to exit from unemployment into employment. In addition, the returns to training programs over the medium term are higher when the training is delivered during periods of recession, since the positive effects of training are amplified when the economy does revive and begin to climb out of the trough. The research of Forslund et al. (2011), focused on Sweden, sheds complementary light on the question. It compares the impact on employment of training programs for persons who are in work versus training for those seeking work. The training of employed persons has a larger impact in the short run, but that result is reversed beyond a period of 100 days. The lock-in effect of training programs on the unemployed is also lessened in regions where unemployment is higher.

However, training programs may exert no impact if they are used merely as a means to extend benefits beyond their maximum normal duration (some programs comprise training allowances or allow the prolongation of unemployment insurance entitlements). For instance, Richardson and van den Berg (2002) quantify the individual effect of training in Sweden (where training programs did permit such benefit extensions before 2001) over the period 1993–2000. They find positive treatment effects, but only in the very short run, and if they include the time spent on training (the "lock-in" effect), this treatment effect disappears.

4.3.2 The Effects on Wages

Research in Europe, like its counterpart in North America, fails to detect any important impact of brief periods of training on the future wages of the trainees.

European Studies

The effect of training on wages appears more ambiguous than its effect on the employment rate. For example, Björklund (1994) finds that the active labor market policies of the late 1970s in Sweden were the cause of a very strong rise in wages. With English data, Dolton et al. (1994) estimate a very large positive effect on the wages of men but a negative one on the wages of women. The research of Westergard-Nielsen (1993) reports on a sample of more than 30,000 observations covering a period of 8 years in Denmark. The aim here was to assess the effects of a "vocational classroom training" program delivered for 2–4 weeks. The authors find an increase of around 1% in the wages of men.

In the United Kingdom, Blundell et al. (1996) use the National Child Development Survey (NCDS) to estimate the impact of participation in training on wages over the period 1981–1991. Controlling for selection factors and for transitory shocks capable

of affecting incomes, Blundell et al. find that participation in training delivered within the firm increases the wage of men by 3.6% on average but has no effect on the wage of women. Participation in training outside the firm generates stronger returns: around 7% for men and 5% for women. These impacts appear to be limited for adults. But they might be larger for youth. Fersterer et al. (2008) focus on the effect of apprenticeship. They use the shutting-down of firms as an instrumental variable, in other words a source of exogenous variation in the duration of the training imparted to apprentices. Use of this instrumental variable permits them to reduce selection bias and leads them to estimate high wage returns to apprenticeship, on the order of 15% to 20%.

The wide spread of these estimates should make us cautious in drawing conclusions about the effect of training policies on wages. Selection biases might lead to overestimates of this impact. It is quite possible that it is the most efficient individuals who apply for and are admitted to these training programs. If that is the case, we will observe that the individuals who get the training have better results than others, even if the programs themselves did nothing to improve the efficiency of the enrollees. The case of "Formation continue" in France is a good example of this.

To obviate the risk of underinvestment in training (highlighted by the theoretical analysis above), France set up a system in 1971 that obliges firms to spend a figure currently set at 1.5% of their total payroll on training for employees. Using a survey of training and skills upgrading, Goux and Maurin (2000) show that employer-sponsored training does not have a large effect on the wages of those who receive it but that it does increase the length of time these recipients remain with the firm. To be precise, they show that the apparent wage premium of employees enrolled in training (on the order of 5% for a week of training!) comes solely from unobserved characteristics. In other words, it is likely that those whom the firm regards as its "best" employees are the ones who benefit from extra training and higher wages. This study also notes that firms predominantly finance *specific* training only, which accounts for the extended careers of the recipients with the same firms and the observed absence of further wage premiums for those who change firms after having been trained (and who are, as it happens, very few in number).

Leuven and Oosterbeek (2008) confirm the importance of selection bias in the case of the Netherlands. These authors reduce the control group to nonparticipants in training who were barred from taking part by some aleatory event, independent of their will and independent of the result variable constituted by the wage. It was possible to construct this control group thanks to a survey carried out on a sample of Dutch wage earners, which included questions about their personal work histories. Leuven and Oosterbeek (2008) show that when the control group is reduced to wage earners who missed out on training for involuntary reasons, the impact of training on increased wages shrinks from 12.5% to 0.6%.

North American Studies

Table 14.12 contains several illustrations which sum up the conclusions that emerge from research based on nonexperimental American data. The assessed measures mainly concern training for economically disadvantaged populations. In the first place, readers will note the great divergence that may exist between studies utilizing identical data. For example, estimates of the annual gains for male participants in the Comprehensive Employment and Training Act (CETA) program in 1976 range from $-1,553 to $+1,638.

TABLE 14.12
Nonexperimental estimates of the effects of federal government programs in the United States.

Study	Program	Δ wage $M^{(2)}$	Δ wage $W^{(3)}$
Economically disadvantaged adults			
Cooley et al. (1979)	1969–1971 MDTA	1,395	2,038
Dickinson et al. (1986)	1976 CETA	−1,553	24
Geraci (1984)	1976 CETA	0	2,026
Ashenfelter and Card (1985)	1976 CETA	1,638	2,220
Economically disadvantaged youth			
Gay and Borus (1980)	1969–1972 Job Corps	−261	−1,555
Dickinson et al. (1986)	1976 CETA	−1,347	449
Bassi et al. (1984)	1977 CETA	−1,225	97

Note: MDTA refers to programs set up under the Manpower Development and Training Act of 1962; CETA refers to programs set up under the Comprehensive Employment and Training Act of 1973. (2) and (3) = Annual wage increase after the program for white men (M) and white women (W), expressed in 1997 dollars.

Source: Heckman et al. (1999, table 24, p. 2065).

For the women in the same cohort, the estimates of average gains are positive, but they nevertheless range from $24 to $2,038. According to Heckman et al. (1999), these wide spreads are generated by the difficulty of constructing a control group in a coherent manner using the matching method, which, as we noted above, does not automatically take unobserved heterogeneity into account. Still, if we set aside the studies most affected by this type of bias, the results obtained from nonexperimental data are very close to those obtained from experimental data. One highly general point is that training programs focusing on disadvantaged populations benefit adult women especially. Conversely, the effects of these programs on the wages of adult men are not always positive, and when they are, the extent of the effect is less than it is with women. The figures in the lower part of table 14.12 confirm what nonexperimental studies tell us about the impact of training programs on economically disadvantaged youth—an impact that often proves to be negative for young, white males (it is sometimes slightly positive for young males from ethnic minorities) and at best slightly positive for young females.

Overall, evaluations of training programs in the United States do not produce an impressive balance sheet when it comes to the efficiency of these programs. Only the group of economically disadvantaged adult women appears to derive a real benefit for an acceptable cost from these programs. Conversely, the effects on other categories of the population, in particular young people, are most often very modest and sometimes even negative. Upon reflection, these conclusions are not at all surprising, for as we saw in chapter 4, a year of extra education raises income by between 6% and 10%. It would have been astonishing if the gains from training programs, which are generally of short duration and cost much less than a year of education in school or college, were to exceed these figures.

The reason for the relative inefficiency of training programs might be the fact that state intervention is also subject to disfunctionalities. Given the existence of information asymmetries between the private sector and the public authorities, problems arise

regarding the verifiability of investments in training which limit the efficiency of subsidies paid to firms and workers. Public institutions can obviously take the place of the private sector in training workers directly. This will be general training only, for the know-how specific to a firm can only be gained "on the job." In this sense, the training supplied by public institutions, since it is not closely related to production, is often less efficient than that acquired within firms (Acemoglu and Pischke, 1999a, 1999b). Moreover, the quality of the production of public institutions providing training itself proves difficult to verify.

4.3.3 An Iconic Example: The Job Corps Program in the United States

Table 14.13 illustrates, with several examples, the conclusions of social experiments concerning training programs carried out in the United States on groups of economically disadvantaged youth (men and women). It turns out that the programs tested have high costs and do not really improve the situation of these young people, in terms of either employment or wages, except to a very modest degree for women. Social experiments carried out in the United States also find that it is the least skilled individuals who derive the least advantage from training programs. Temporary job creation (*WE*) seems to benefit them, however. One possible interpretation of this result is that this type of measure gives persons in this category the chance to acquire work habits that persons in more skilled categories already possess.

However, for youth, one program stands out: Job Corps is the United States' largest education and job training program for disadvantaged youth between the ages of 16 and 24. It offers an intensive one-year program which includes vocational training and academic education but also counseling, social skills training, health care, and health education, as well as placement. Most students reside at the Job Corps center while training. Each year, more than 60,000 new participants enroll at a cost per student of about $16,000. Schochet et al. (2006) analyzed results from an experiment carried out from late 1994 to early 1996, which selected at random nearly 81,000 eligible applicants nationwide to be assigned either to a program group or to a control group. They show

Table 14.13
The results of some social experiments in the United States, on economically disadvantaged youth.

Measure	Cost(1)	Δ employment(2)	Δ wages(3)
NSW	9,314	0.3	−79
JOBSTART	6,403	−0.9	−721
NJS (JTPA)			
Women	1,116	—	133
Men	1,731	—	−553

Note: The programs tested combine training and subsidy. JTPA = Job Training Partnership Act; NJS = National JTPA Study; NSW = National Supported Work demonstration. (1) Marginal cost of treatment for one person for one year in 1997 dollars. (2) Difference in employment rates between the treated group and the control group in the last quarter of the year subsequent to the experiment. (3) Difference in annual average wages between the treated group and the control group in the first or second year subsequent to the experiment, in 1997 dollars.

Source: Heckman et al. (1999, table 22, p. 2058).

TABLE 14.14
Benefits and costs of Job Corps. All values are in 1995 dollars.

	All Job Corps	Those 20 to 24 years old
Earnings following program exit	119	34,896
Output produced during vocational training	220	250
Reduced use of other programs/services	2,186	937
Reduced crime	1,240	−3,787
Total benefits	3,544	32,045
Program cost	−16,205	−17,755
Transfers	2,361	2,562
Total costs	−13,844	−15,193
Net benefits	−10,300	16,853

Note: Allowances received by participants are considered as "transfers" from taxpayers to Job Corps students. As these items have intrinsic value to the students irrespective of their contribution to the participants' future, they are subtracted from total program costs when calculating the cost of Job Corps to society.

Source: Schochet et al. (2006, table 10).

that the program does yield earnings gains in years 3 and 4 after random assignment (a gain of about 12% in year 4). But table 14.14 shows that these gains are not sustainable except for the older enrollees, those aged 20 to 24 at program application.

Table 14.14 shows that the Job Corps also significantly reduces involvement with crime: "According to the survey data, the arrest rate was reduced by 16 percent (about five percentage points), and similar reductions were found also for conviction and incarceration rates" (Schochet et al., 2006, p. 3). The program also has small but beneficial impacts on receipt of social assistance and on self-assessed health status. Overall, the cost per student exceeds benefits for the full sample but appears to be cost-effective for the 20- to 24-year-olds. This experiment shows that the returns to training of disadvantaged teenagers or young adults are quite uncertain and may actually be negative, even with such a comprehensive program. The study also highlights the need for very early interventions (see chapter 4).

4.3.4 The Benefits of Education and Training over the Life Cycle

Training policies can have effects that vary widely according to the populations concerned. Figure 14.13, taken from Heckman (2000), sums up the main lessons to be learned from studies in this field (see also chapter 4, section 5.4). The figure displays net returns to education as a function of age for two types of individual. A battery of criteria (social background, IQ test score, etc.) makes it possible to distinguish persons with high innate capacities for learning and socialization from those with low ones. Figure 14.13 shows, first of all, that the returns to education diminish with age for all categories of the general population, as retirement draws nearer. It also shows that the net return to education is greater for very young children with low capacities than for very young children with high ones. Conversely, this return falls off more rapidly for those with low capacities, since the boost given by special education in terms of intellectual development and socialization declines quickly as individuals grow older.

Figure 14.13 suggests that educational assistance should be specifically targeted at young children from socially disadvantaged backgrounds and/or ones whose capacities

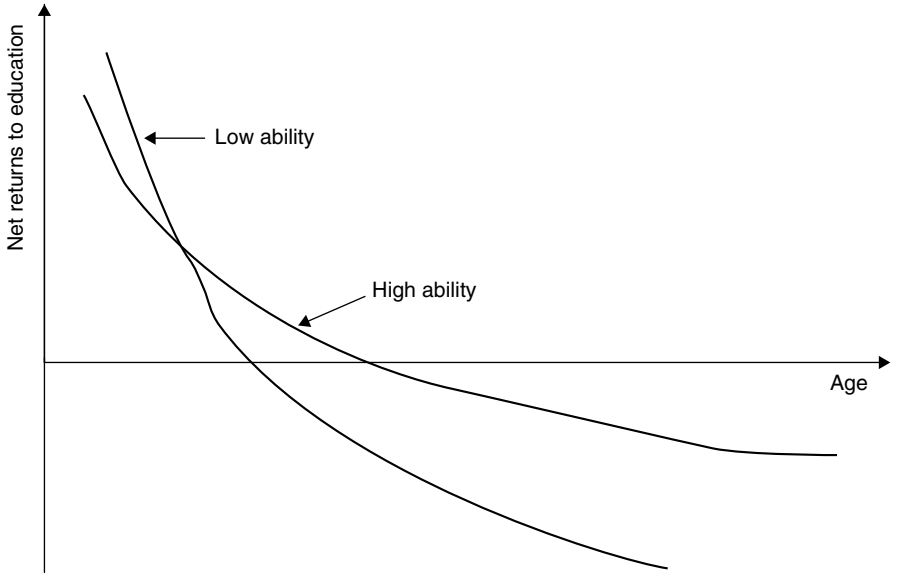


Figure 14.13
The relationship between age and net returns to education for two types of individual.

Source: Heckman et al. (2000).

for social integration are low. Expenditure of this type brings a much higher return than educational assistance to adults. This does not mean that nothing should be done to help the most disadvantaged adults. The conclusion to be drawn is rather that education is not the most suitable way to assist such persons: the return to society is inadequate, and the boost to the earning power of the beneficiaries insignificant. Hence Heckman (2000) suggests that it is preferable to help them by subsidizing their jobs through lower payroll taxes or reductions in income tax.

4.4 Hiring Subsidies

Hiring subsidies are temporary wage subsidies. They are usually targeted at low-wage jobs or hard-to-place workers, and they aim to bridge the gap between the worker's productivity and the minimum wage that prevails in the sector. Even though they are temporary, they can have a durable impact by obviating long-term unemployment and unemployment recurrence, giving the unemployed opportunities to acquire valuable experience in the private sector, which in turn might improve their job prospects. So, when targeted at specific groups, one important impact of these programs is to "reshuffle" the queue of job seekers.

4.4.1 European Studies: A Large Impact

The aggregate impact of hiring subsidies on unemployment might be lessened due to the well-known deadweight effect (some of these workers would have been hired anyway) but also by the substitution effect (nontargeted workers and unemployed persons may

be hired less often) and the displacement effect (firms benefiting more from the program get an advantage over other firms in the same sector).

A large majority of the evaluations from micro data report a positive impact of hiring subsidies on unemployment exit rates and employment prospects. These evaluations are typically based on difference-in-differences methods, and as such allow deadweight effects to be taken into account. For instance, Jaenichen and Stephan (2011) analyzed the impact of the main hiring subsidy in Germany, targeted at hard-to-place persons and which can cover up to 50% of the wage cost over 12 months. They find that wage subsidies may increase the employment prospects of supported workers by significant amounts: 3 years after entry into the program (i.e., after it was over for the individuals in question), the share in regular employment is 25% to 42% higher in the treatment group than in the control group. Bernhard et al. (2008) find similar results for beneficiaries of social assistance who enter a hiring subsidy program (20 months after taking up a subsidized job, the employment rate is almost 40 percentage points higher among participants). In Sweden, similar results were obtained by Sianesi (2002), who estimated the impact of several types of active programs on the probability of employment over time. While she finds no impact of training and public jobs, entry into a job subsidy program does pay off significantly in terms of persistently higher employment rates (up to 40 percentage points) soon after the program ends and for several years subsequently.

These subsidies can be especially beneficial to youth, when targeted at the low-skilled or at low-wage jobs. In the United Kingdom, one of the active options in the New Deal for Young People (NDYP) after 6 months of unemployment was a job subsidy of up to £60 per week that lasted 6 months. Dorsett (2006) evaluated this program. His result is that the subsidized employment option is more effective at increasing the chances of exiting unemployment and securing unsubsidized employment after the program than the other available options (notably education/training or temporary public employment). This result is confirmed by Van Reenen (2003): unemployed young men are 20% more likely to get jobs as a result of the NDYP program, and much of this effect is likely to be due to the wage subsidy option (and also, but to a smaller extent, to enhanced job search). In France, Fougère et al. (2000) studied several programs targeted at the young unemployed and concluded that the reduction of labor costs is the only program to have a significant impact on the employment probabilities of low-wage workers, even though the effect appears to be stronger for workers between 25 and 30.

4.4.2 North American Studies: A Significant Impact But Less Than in Europe

A number of hiring subsidy programs exist in the United States. An example of a targeted program is the Work Opportunity Tax Credit (WOTC), introduced in 1996, which offers generous subsidies to firms that hire disadvantaged workers, including certain welfare recipients and the disabled. The similar Welfare-to-Work (WtW) tax credit, implemented in 1998, offers firms potentially larger subsidies for hiring long-term welfare recipients. It replaced a similar program called the Targeted Jobs Tax Credit (TJTC), in effect from 1979 through 1994. The hiring of disadvantaged welfare recipients can be subsidized at a rate ranging between 35% and 50% of the wage (with a ceiling) for one or two years. Using a difference-in-differences method, Hamersma (2005) finds evidence

of short-run improvement in employment levels (but no evidence of impact on tenure in the longer run) and also a positive impact on earnings.

The New Job Tax Credit (NJTC) was a nontargeted federal program set up in the middle of the period 1977–1978 to counteract the recession. It was not tailored to any particular population, but it did target the earners of low wages (the subsidy amounts to 50% of the first $4,200 of wages per hire up to a maximum of $100,000 per firm in a year). Using survey data, Perloff and Wachter (1979) compare firms declaring that they know about the NJTC and firms which do not. They conclude that employment grew 3% faster thanks to the NJTC. Perloff and Wachter concede that their result is an upper bound of the true effect, as serious endogeneity bias affects their comparison. Using aggregate time series, Bishop (1981) finds that the NJTC had significant positive employment effects (between 0.66% and 2.95%) and negative effects on prices.

The NJTC was the only U.S. hiring subsidy implemented at the federal level until 2010. At the state level, there were many more Job Creation Tax Credits. Chirinko and Wilson (2010) construct a large data set documenting all these policies. They conduct an event study (difference-in-differences across U.S. states) and estimate that the tax credit induced a 0.1% increase in employment in the month when firms both know and can qualify for the tax credit. They pay particular attention to dynamic effects (documenting an Ashenfelter dip between the signing and qualifying dates).

In a large survey on hiring subsidies, Neumark (2013) concludes that they do not have significant effects on *total* employment when they are targeted at specific disadvantaged groups. Such targeted policies stigmatize their beneficiaries and entail substitution effects. However Neumark also concludes that nontargeted hiring subsidies to specific populations, such as the NJTC, may have significant effects on employment.

4.5 Temporary Public Jobs

In some countries, notably in Europe, the unemployed can be referred directly to temporary public jobs. The main rationale for this type of policy is to offer some labor market experience to workers who are usually disadvantaged and have a low probability of finding a job in the regular labor market. Hence, like temporary hiring subsidies, temporary public jobs can "reshuffle" the queue of unemployed persons. This policy is also sometimes used on a larger scale to lower the number of the administratively registered unemployed during slumps. In that setting it appears largely as a short-run strategy to mechanically lower unemployment, without consideration of the impact on the odds of employment in the nonsubsidized private sector at the end of the contract.

There is now strong evidence from various countries that this type of policy has no positive postprogram impact on the probability of being employed in the regular labor market after the end of the program. For instance, in Sweden, where these jobs have been widely used in the past, there is absolutely no evidence of postprogram impact (Sianesi, 2002). The same conclusion holds for similar programs in Germany: Caliendo et al. (2004) show that at the beginning of the 2000s in Germany, two years after the start of public employment programs participants have a significantly higher probability of being registered as unemployed at the labor office in comparison to matched nonparticipants. The main reasons for this effect may be that the nature of the accrued work experience is not valued in the private sector, the existence of "stigma" effects, and the

fact that participants are "locked-in" during the program and do not look for a regular job (see van Ours, 2004, on the case of the Slovak Republic).

This measure can be adopted as a social policy tool, if not as an employment policy tool, to improve the welfare of some very disadvantaged groups who have no chance of getting a job even in the subsidized private sector or whose chances of benefiting from training are deemed small. This, though, implies tight targeting by the authorities. Temporary public jobs may also be used within the framework of an activation strategy, in which a logic of rights and duties is invoked so as to incentivize certain unemployed persons to intensify their search effort (the threat effect). For instance, programs such as "Work-for-the-Dole" in Australia and the "One-Euro" mini-jobs in Germany, which oblige certain unemployed individuals to work part-time while pursuing their job search, are partly designed to enforce the logic of "rights and duties" within a system where the unemployed can lose part of their benefits if they refuse to enter the program.

From a macroeconomic perspective, the conclusions are similar. The creation of temporary or permanent jobs in the nonmarket sector may entail powerful crowding-out effects: the boost in total labor demand that results, particularly when the policy is applied on a large scale, tends to push wages up, and this may create a drag on job creation in the market sector, to the point of totally offsetting the initial effect of the measure on unemployment and employment (Algan et al., 2002, for the OECD countries; Calmfors et al., 2004, for the Swedish case).

4.6 Equilibrium Effects

Should programs or experiments that prove workable on a small scale then be scaled up? The analysis in section 2.4 stresses the need to take equilibrium effects into account when estimating the impact of labor market programs. Unfortunately, very few studies are able to do so. Indeed, most experiments or actual programs do not allow this identification. Besides, it requires the researcher to dispose of information on nonparticipants who could not have been influenced by the program. There are too few studies available to draw firm conclusions on the relative importance of these effects.

As for job search assistance targeted at youth, the study of Crépon et al. (2013) in France concludes, for instance, that equilibrium effects can be substantial, notably for young men, whereas Blundell et al. (2004) in the United Kingdom do not identify any significant effect for a similar program. Gautier et al. (2012) analyze a Danish randomized evaluation of a job search assistance program. They compare individuals in experimental counties to job seekers in some similar nonparticipating counties and find substantial negative treatment externalities. As for training, Ferracci et al. (2010) find that in France, the impact of a training program on employment for unemployed workers is lower in areas where the percentage of treated workers is higher. They also show that the employment rate of individuals who did not have this training presents a U-shaped profile, falling and then rising with the proportion of unemployed persons who did have training. In labor markets where the size of programs tends to be substantial, as in Europe, this issue is particularly important. But more empirical research will be needed to draw firm conclusions on the importance of externalities.

We saw above that, according to the bulk of the assessments carried out, training programs in the United States appear to have little effect, except when they are targeted

at the group of economically disadvantaged adult women. These evaluations, however, were made in a partial equilibrium framework and thus register only a part of the impact of training programs. The existence of positive externalities linked to training suggests that these studies likely underestimate the gains from these programs. Yet on the other hand, we cannot rule out the possibility of negative effects being induced when the program demands large investments and concerns a high proportion of the population. The study of Heckman, Lochner et al. (1998) suggests that these effects are not negligible, according to their analysis of the consequences of an extra subsidy of $500 to those who enroll in college in the United States, financed by a proportional tax on income. The estimates show that college enrollments increase by 5.3% at *partial* equilibrium, on the assumption that the structure of wages is not affected by the increase in the subsidy, and leaving aside the effects of taxes. But when this policy is assessed at *general* equilibrium, the estimated effect falls to 0.46% on account of the decline in the wage of a college graduate with respect to that of a high school graduate, a decline itself due to the rise in the number of those enrolled in colleges.

Overall, job search assistance is the least costly of the active policies and probably one of the most effective: social experiments carried out in a number of countries (Sweden, Canada, the United Kingdom, and the United States) yield convincing results, though it remains an open question whether checking up on job search effort or helping the unemployed while they look (or whatever combination of these two) is the most important factor. Of all the measures aimed at young people, only employer wage subsidies give much reason for satisfaction. Training entails lock-in effects in the short run but may have positive effects in the longer term. In all cases, including intensive placement and monitoring, policies have heterogeneous effects across population groups.

5 SUMMARY AND CONCLUSION

- In considering public expenditures on labor market policies, a distinction is made between *active policy* measures, which aim to improve the functioning of the labor market, and *passive policy* measures, which seek instead to improve the living conditions of workers. As a general rule, the amount spent on passive measures exceeds that spent on active ones.
- Public agencies occupy an important place in the array of institutions that manage job offers in many countries. From the standpoint of the social optimum, placement agencies (public or private) are only justified if they guarantee a better matching of unemployed persons to vacant jobs than the "natural" process would and if operating them does not incur excessively high fixed costs. Decentralized equilibrium with private agencies is likely inefficient, on account of congestion effects and the potentially oligopolistic structure of the placement market. Empirical studies suggest that public employment services have a significant effect, at a reasonable cost, on the exit rate from unemployment of the individuals concerned.
- General training improves the productivity of an individual for all jobs, while specific training increases only his productivity for a particular job. In a

perfectly competitive economy, the investment in general training would be entirely financed by workers, since they would benefit exclusively from the investment. Individual choices would then be socially optimal. The same does not hold true if the matching process governing the labor market is imperfect. In this context, decentralized equilibrium is characterized by underinvestment in general training, even if firms and workers can commit themselves to complete contracts, since it is impossible for agents to bargain over the amount of this type of training with *future* employers, who will benefit tomorrow from the investment made today.

- When it comes to specific training, decentralized equilibrium is socially efficient when the employer and the worker can commit themselves to complete contracts. This result is independent of any possible imperfection in the matching process, since the amount of time spent looking for work does not play a part in decisions regarding investment in specific training. But as we know (see chapter 6), agents most often cannot sign complete contracts. In the presence of incomplete contracts, decentralized equilibrium leads to underinvestment in this type of training.
- Employment subsidies in the form of reduced labor costs for the employer generate upward pressure on the negotiated wages. When unemployment benefits are perfectly indexed to wages, the employee captures the *whole* subsidy initially granted to the firm in the form of a wage rise, and at equilibrium subsidies have no effect on employment. Conversely, when unemployment benefits are imperfectly indexed or wages are rigid, employment subsidies reduce the unemployment rate.
- The creation of public-sector jobs, by exerting upward pressure on wages, can crowd out private-sector jobs. Its effect on unemployment is thus a priori ambiguous. Empirical assessments suggest that nontargeted employment subsidies, or the creation of public-sector jobs, are costly measures that should only find marginal application.
- All labor market policies may exert externalities on nonparticipants and thus lead to equilibrium effects that may diminish or enhance the total effect of the measures on employment and wages. For instance, intensive placement strategies for some groups may reduce the chances of finding a job for those who do not benefit from this special treatment, notably when the program is relatively small. Subsidies offered to some firms to hire disadvantaged workers or create certain types of jobs may put pressure on wages, which in turn can weigh down job creation among firms that cannot benefit from such subsidies but still compete on the same market.
- To evaluate the impact of employment policies, we must compare the performances of the individuals who benefit from measures with those of individuals who do not. This kind of assessment poses problems, since the characteristics of the individuals who do benefit from employment policies are generally particular, which creates a potential selection bias. It is possible to deal with this problem, on the basis of observational data gathered from surveys or administrative data sets, by assessing the performance of policies for groups

of individuals possessing identical characteristics (the matching method). The existence of unobserved characteristics nevertheless constitutes an unavoidable limitation on this type of approach. Social experiments, which consist of choosing the beneficiaries of employment policies at random within the guidelines of a precisely defined protocol and comparing their performances with those of nonbeneficiaries, make it possible to deal with this problem.

- The appraisal of active employment policies yields mixed results. Studies carried out in the United States conclude that only adult, economically disadvantaged women appear to derive any real benefit, for an acceptable cost, from measures to promote training. Overall, job search assistance is the least costly of the active policies and probably one of the most effective. Some programs of job subsidies in the private sector may also enhance the chance of employment. Evaluations of temporary public employment creation programs point to insignificant or even negative effects on the chances of holding a job in the regular labor market at the end of the program. There is also evidence that the threat of having to enter mandatory programs, with the presence of sanctions against half-hearted job searches, increases unemployment exits. Finally, all empirical research dedicated to assessing employment policies generally neglects their macroeconomic effects.

6 RELATED TOPICS IN THE BOOK

- Chapter 2, section 2.2: Main results on labor demand elasticity
- Chapter 3, section 1.2: The question of tax incidence
- Chapter 4, section 2: The theory of human capital
- Chapter 5, section 2: Basic job search theory
- Chapter 5, section 3: Empirical aspects of job search
- Chapter 9, section 3: The matching model
- Chapter 9, section 4: The efficiency of market equilibrium
- Chapter 12, section 1.2: The effect of taxes on the labor market
- Chapter 13, section 1.4: Optimal unemployment insurance in a dynamic environment

7 FURTHER READINGS

Blundell, R., & Costa Dias, M. (2009). Alternative approaches to evaluation in empirical micro-economics. *Journal of Human Resources*, *44*(3), 565–640.

Blundell, R., Costa Dias, M., Meghir, C., & Van Reenen, J. (2004). Evaluating the employment impact of a mandatory job search program. *Journal of the European Economic Association*, *2*(4), 569–606.

Card, D., Kluve, J., & Weber, A. (2010). Active labour market policies: A meta-analysis. *Economic Journal*, *120*, F452–F477.

Crépon, B., Duflo, E., Gurgand, M., Rathelot, R., & Zamora, P. (2013). Do labor market policies have displacement effects? Evidence from a clustered randomized experiment. *Quarterly Journal of Economics*, *128*(2), 531–580.

REFERENCES

Acemoglu, D. (1997). Training and innovation in an imperfect labour market. *Review of Economic Studies, 64*, 445–467.

Acemoglu, D., & Pischke, J.-S. (1998). Why do firms train? Theory and evidence. *Quarterly Journal of Economics, 113*, 79–119.

Acemoglu, D., & Pischke, J.-S. (1999a). The structure of wages and investment in general training. *Journal of Political Economy*, *107*, 539–572.

Acemoglu, D., & Pischke, J.-S. (1999b). Beyond Becker: Training in imperfect labour markets. *Economic Journal*, *112*, 112–142.

Aghion, P., & Howitt, P. (1998). *Endogenous growth theory*. Cambridge, MA: MIT Press.

Algan, Y., Cahuc, P., & Zylberberg, A. (2002). Public employment and labour market performances. *Economic Policy*, *17*(34), 9–64.

Angrist, J. (1998). Estimating the labor market impact on voluntary military service using social security data on military applicants. *Econometrica*, *66*, 249–288.

Angrist, J., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Arni, P., Lalive, R., & van Ours, J. (2009). How effective are unemployment benefit sanctions? Looking beyond unemployment exit (IZA Discussion Paper 4509). Institute for the Study of Labor, Bonn, Germany.

Ashenfelter, O. (1978). Estimating the impact of training programs on earnings. *Review of Economics and Statistics*, *6*(1), 47–57.

Ashenfelter, O., & Card, D. (1985). Using the longitudinal structure of earnings to estimate the effects of training programs. *Review of Economics and Statistics*, *67*(3), 648–660.

Autor, D., & Duggan, M. (2003). The rise in disability rolls and the decline in unemployment. *Quarterly Journal of Economics*, *118*(1), 157–206.

Bassi, L., Simms, M., Burnbridge, L., & Betsey, C. (1984). Measuring the effect of CETA on youth and the economically disadvantaged (Report for the US Department of Labor 20-11-82-19). Washington, DC: The Urban Institute.

Becker, G. (1964). *Human capital*. Chicago, IL: University of Chicago Press.

Benabou, R. (1996). Heterogeneity, stratification, and growth. *American Economic Review*, *86*, 584–609.

Bernhard, S., Gartner, H., & Stephan, G. (2008). Wage subsidies for needy job-seekers and their effect on individual labour market outcomes after the German reforms (IZA Discussion Paper No. 3772). Institute for the Study of Labor, Bonn, Germany.

Bergemann, A., & van den Berg, G. (2008). Active labor market policy effects for women in Europe: A survey. *Annales d'Economie et de Statistique*, *91–92*, 385–408.

Bishop, J. (1981). Employment in construction and distribution industries: The impact of the New Jobs Tax Credit. In S. Rosen (Ed.), *Studies in labor markets* (pp. 209–246). Chicago, IL: University of Chicago Press.

Björklund, A. (1994). Evaluations of Swedish labor market policy. *International Journal of Manpower*, *15*(5), 16–31.

Björklund, A., & Regnér, H. (1996). Experimental evaluation of European labour market policy. In G. Schmid, J. O'Reilly, & K. Schömann (Eds.), *International handbook of labour market and evaluation* (pp. 89–114). Adelshot, U.K.: Edward Elgar.

Black, D., Smith, J., Berger, M., & Noel, B. (2003). Is the threat of reemployment services more effective than the services themselves? Evidence from random assignment in the UI system. *American Economic Review*, *93*, 1313–1327.

Blundell, R., Costa Dias, M., Meghir, C., & Van Reenen, J. (2004). Evaluating the employment impact of a mandatory job search program. *Journal of the European Economic Association*, *2*(4), 569–606.

Blundell, R., Dearden, L., & Meghir, C. (1996). The determinants and effects of work-related training in Britain. Institute for Fiscal Studies, London.

Blundell, R., & MaCurdy, T. (1999). Labor supply: A review of alternative approaches. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 3A, chap. 27). Amsterdam: Elsevier Science.

Boone, J., Sadrieh, A., & van Ours, J. (2009). Experiments on unemployment benefit sanctions and job search behavior. *European Economic Review*, *53*(8), 937–951.

Cahuc, P., & Fontaine, F. (2009). On the efficiency of job search with social networks. *Journal of Public Economic Theory*, *11*(3), 411–439.

Cahuc, P., & Le Barbanchon, T. (2010). Labor market policy evaluation in equilibrium: Some lessons of the job search and matching model. *Labour Economics*, *17*, 196–205.

Caliendo, M., Hujer, R., & Thomsen, S. (2004). New evidence on the effects of job creation schemes in Germany—A matching approach with threefold heterogeneity. *Research in Economics*, *58*(4), 257–302.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching.*Journal of Economic Surveys*, *22*(1), 31–72.

Calmfors, L. (1994). Active labour market policy and unemployment. A framework for the analysis of crucial design features. *OECD Economic Studies*, *22*, 7–47.

Calmfors, L., Forslund, A., & Hemström, M. (2004). The effects of active labor market policies in Sweden: What is the evidence? In J. Agell, M. Keen, & A. Weichenreider (Eds.), *Labor market institutions and public regulation*. Cambridge, MA: MIT Press.

Calmfors, L., & Lang, H. (1995). Macroeconomic effects of active labor market programs in a union wage-setting model. *Economic Journal*, *105*, 601–619.

Card, D., Kluve, J., & Weber, A. (2010). Active labour market policies: A meta-analysis. *Economic Journal*, *120*, F452–F477.

Chang, C., & Wang, Y. (1996). Human capital investment under asymmetric information: The Pigovian conjecture revisited. *Journal of Labor Economics*, *14*, 505–519.

Chirinko, R., & Wilson, D. (2010). Job creation tax credits and job growth: Whether, when, and where? (Working Paper Series 2010-25). Federal Reserve Bank of San Francisco.

Coleman, J., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfelde, F., & York, R. (1966). *Equality of educational opportunity*. Washington, DC: U.S. Government Printing Office.

Cooley, T., McGuire, T., & Prescott, E. (1979). Earnings and employment dynamics of manpower trainees: An exploratory econometric analysis. In R. Ehrenberg (Ed.), *Research in labor economics* (vol. 4, suppl. 2, pp. 119–147). Greenwich, CT: JAI Press.

Cooper, R., & John, A. (1988). Coordinating coordination failures in Keynesian models. *Quarterly Journal of Economics*, *103*, 441–465.

Crépon, B., Dejemeppe, M., & Gurgand, M. (2005). Counseling the unemployed: Does it lower unemployment duration and recurrence? (IZA Discussion Paper 1796). Institute for the Study of Labor, Bonn, Germany.

Crépon, B., Duflo, E., Gurgand, M., Rathelot, R., & Zamora, P. (2013). Do labor market policies have displacement effects? Evidence from a clustered randomized experiment. *Quarterly Journal of Economics 128*(2), 531–580.

Crépon, B., Ferracci, M., & Fougère, D. (2012). Training the unemployed in France: How does it affect unemployment duration and recurrence? *Annales d'Economie et de Statistique*, *107–108*, 175–199.

Dickinson, K., Johnson, T., & West, R. (1986). An analysis of the impact of CETA on participants' earnings. *Journal of Human Resources*, *21*, 64–91.

DiNardo, J., & Lee, D. (2011). Program evaluation and research designs. In O. Ashenfelter & O. Card (Eds.), *Handbook of labor economics* (vol. 4A, pp. 463–536). Amsterdam: Elsevier.

Dolton, P., & O'Neill, D. (1996). Unemployment duration and the restart effect: Some experimental evidence. *Economic Journal*, *106*, 387–400.

Dolton, P., & O'Neill, D. (2002). The long-run effects of unemployment monitoring and work-search programs: Experimental evidence from the United Kingdom. *Journal of Labor Economics*, *20*, 381–403.

Dolton, P., Makepeace, G., & Treble, J. (1994). The wage effect of YTS: Evidence from YCS. *Scottish Journal of Political Economy*, *41*(4), 444–453.

Dorsett, R. (2006). The new deal for young people: Effect on the labour market status of young men. *Labour Economics*, *13*, 405–422.

Ferracci, M., Jolivet, G., & van den Berg, G. (2010). Treatment evaluation in the case of interactions within markets (IZA Discussion Paper 4700). Institute for the Study of Labor, Bonn, Germany.

Fersterer, J., Pischke, S., & Winter-Ebmer, R. (2008). Returns to apprenticeship training in Austria: Evidence from failed firms. *Scandinavian Journal of Economics*, *110*(4), 733–753.

Fisher, R. (1935). *Design of experiments*. New York, NY: Hafner.

Forslund A., Fredriksson, P., & Vikström, J. (2011). What active labor market policy works in a recession? (IFAU Working Paper No. 2011:2). Institute for Evaluation of Labor Market and Education Policy, Uppsala, Sweden.

Fougère, D., Kramarz, F., & Magnac, T. (2000). Youth employment policies in France. *European Economic Review*, *44*, 928–942.

Fougère, D., Pradel, J., & Roger, M. (1999). The influence of the state employment service on the search effort and on the probability of leaving unemployment (Working Paper 9904). CREST-INSEE, Paris, France.

Gautier, P., Muller, P., van der Klaauw, B., Rosholm, M., & Svarer, M. (2012). Estimating equilibrium effects of job search assistance (IZA Discussion Paper 6748). Institute for the International Study of Labor, Bonn, Germany.

Gay, R., & Borus, M. (1980). Validating performance indicators for employment and training programs. *Journal of Human Resources*, *15*, 29–48.

Geerdsen, L. (2006). Is there a threat effect of labour market programmes? A study of ALMP in the Danish UI system. *Economic Journal*, *116*, 738–750.

Geraci, V. (1984). Short-term indicators of job training program effects on long-term participant earnings (Report for the U.S. Department of Labor 20-48-82-16). Washington, DC: The Urban Institute.

Goux, D., & Maurin, E. (2000). Returns to firm-provided training: Evidence from French worker-firm matched data. *Labour Economics*, *7*(1), 1–20.

Graversen, B., & van Ours, J. (2008). How to help unemployed find jobs quickly: Experimental evidence from a mandatory activation program. *Journal of Public Economics*, *92*, 2020–2035.

Greenberg, D., Michalopoulos, C., & Robins, P. (2003). A meta-analysis of government-sponsored training programs. *Industrial and Labor Relations Review*, *57*(1), 31–53.

Hamersma, S. (2005). The work opportunity and welfare-to-work tax credits. Urban-Brookings Tax Policy Center Brief, No. 15, October, Washington, DC.

Heckman, J. (2000). Policies to foster human capital. *Research in Economics, 54*, 3–56.

Heckman, J., Ichimura, H., & Todd, P. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training program. *Review of Economic Studies*, *64*, 605–654.

Heckman, J., Ichimura, H., & Todd, P. (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies*, *65*, 261–294.

Heckman, J., Lalonde, R., & Smith, J. (1999). The economics and econometrics of active labor market programs. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 3A, chap. 31, pp. 1865–2097). Amsterdam: Elsevier Science.

Heckman, J., Lochner, L., & Taber, C. (1998). General equilibrium treatment effects: A study of tuition policy. *American Economic Review*, *88*, 381–386.

Heckman, J., & Smith, J. (1998). The sensitivity of experimental impact estimates: Evidence from the National JTPA study. In R. Freeman & L. Katz (Eds.), *Youth employment and unemployment in the OECD countries*. Chicago, IL: University of Chicago Press.

Holmlund, B., & Linden, J. (1993). Job matching, temporary public employment, and equilibrium unemployment. *Journal of Public Economics*, *51*, 329–343.

Ioannides, Y., & Topa, G. (2010). Neighborhood effects: Accomplishments and looking beyond them. *Journal of Regional Science*, *50*, 343–362.

Jaenichen, U., & Stephan, G. (2011). The effectiveness of targeted wage subsidies for hard-to-place workers. *Applied Economics*, *43*, 1209–1225.

Katz, E., & Ziderman, A. (1990). Investment in general training: The role of information and labour mobility. *Economic Journal*, *100*, 1147–1158.

Kluve, J. (2010). The effectiveness of European active labor market programs. *Labour Economics*, *17*, 904–918.

Lalive, R., van Ours, J., & Zweimüller, J. (2005). The effect of benefit sanctions on the duration of unemployment. *Journal of the European Economic Association*, *3*(6), 1386–1417.

Layard, R., & Nickell, S. (1986). Unemployment in Britain. *Economica*, *53*, 121–169.

Lechner, M., & Wunsch, C. (2009). Are training programs more effective when unemployment is high? *Journal of Labor Economics*, *27*(4), 653–692.

Lechner, M., Miguel, C., & Wunsch, C. (2011). Long-run effects of government-sponsored training programs. *Journal of the European Economic Association*, *9*(4), 742–784.

Leuven, E., & Oosterbeek, H. (2008). An alternative approach to estimate the wage returns to private-sector training. *Journal of Applied Econometrics*, *23*, 423–434.

Lewis, H.-G. (1963). *Unionism and relative wages*. Chicago, IL: University of Chicago Press.

Lippoldt, D., & Brodsky, M. (2004). Public provision of employment services in selected OECD countries: The job brokerage function. In D. Balducchi, R. Eberts, & C. O'Leary

(Eds.), *Labor exchange policy in the United States* (pp. 211–248.) Kalamazoo, MI: W. E. Upjohn Institute.

Lucas, R. (1988). On the mechanics of economic development. *Journal of Monetary Economics*, *22*(1), 3–42.

Main, B., & Shelly, M. (1990). The effectiveness of the Youth Training Scheme as a manpower policy. *Economica*, *57*(228), 495–514.

Meyer, B. (1995). Lessons from the U.S. unemployment insurance experiments. *Journal of Economic Literature*, *33*, 91–131.

Neumark, D. (2013). Spurring job creation in response to severe recessions: Reconsidering hiring credits. *Journal of Policy Analysis and Management*, *32*(1), 142–171.

OECD. (2007). *Employment outlook*. Paris: OECD Publishing.

OECD. (2013). *Employment outlook*. Paris: OECD Publishing.

Patacchini, E., & Zenou, Y. (2011). Neighborhood effects and parental involvement in the intergenerational transmission of education. *Journal of Regional Science*, *51*, 987–1013.

Perloff, J., & Wachter, M. (1979). The new jobs tax credit: An evaluation of the 1977–78 wage subsidy program. *American Economic Review*, *69*(2), 173–179.

Quandt, R. (1972). Methods for estimating switching regressions. *Journal of the American Statistical Association*, *67*(338), 306–310.

Regnér, H. (1997). *Training at the job and training for a new job: Two Swedish studies*. Swedish Institute for Social Research, Stockholm, Sweden.

Richardson, K., & van den Berg, G. (2002). The effect of vocational employment training on the individual transition rate from unemployment to work (Working Paper Series 2002:8). IFAU—Institute for the Evaluation of Labour Market and Education Policy.

Rosenbaum, P., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effect. *Biometrika*, *70*(1), 41–55.

Rosholm, M., & Svarer, M. (2008). The threat effect of active labour market programmes. *Scandinavian Journal of Economics*, *110*(2), 385–401.

Roy, A. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers*, *3*, 135–146.

Rubin, D. (1974). Estimating the causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*, 688–701.

Rubin, D. (1980). Comment on "Randomization analysis of experimental data in the Fisher randomization test" by Debrabata Basu. *Journal of the American Statistical Association*, *75*(371), 591–593.

Rubin, D. (1990). Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, *5*, 472–480.

Schochet, P., Burghart, J., & McConnell, S. (2006). National Job Corps study and longer term follow up study. Mimeo, Mathematica Policy Research Inc., Princeton, NJ.

Sianesi, B. (2002). Swedish active labour market programmes in the 1990s: Overall effectiveness and differential performance. The Institute for Fiscal Studies, WP02/03, London.

Sianesi, B. (2004). An evaluation of the Swedish system of active labor market programs in the 1990s. *Review of Economics and Statistics*, *86*(1), 133–155.

Snower, D. (1995). The low-skill, bad-job trap. In A. Booth & D. Snower (Eds.), *Acquiring skills* (chap. 6). Cambridge, U.K.: Cambridge University Press.

Stevens, M. (1994). A theoretical model of on-the-job training with imperfect competition. *Oxford Economic Papers*, *46*, 537–562.

Torp, H., Raaum, O., Heraes, E., & Goldstein, H. (1993). The first Norwegian experiment. In K. Jensen & P. Masden (Eds.), *Measuring labour market measures* (pp. 97–140). Copenhagen: Danish Ministry of Labour.

Ulph, D. (1995). Dynamic competition for market share and the failure of the market for skilled workers. In A. Booth & D. Snower (Eds.), *Acquiring skills* (chap. 5). Cambridge, U.K.: Cambridge University Press.

van den Berg, G., & van der Klaauw, B. (2006). Counseling and monitoring of unemployed workers: Theory and evidence from a controlled social experiment. *International Economic Review*, *47*(3), 895–936.

van den Berg, G., van der Klaauw, B., & van Ours, J. (2004). Punitive sanctions and the transition rate from welfare to work. *Journal of Labor Economics*, *22*(1), 211–241.

van den Berg, G., & Vikström, J. (2009). Monitoring job offer decisions, punishments, exit to work, and job quality (IZA Discussion Paper 4325) Institute for the Study of Labor, Bonn, Germany.

van der Klaauw, B., & van Ours, J. (2013). Carrot and stick: How reemployment bonuses and benefit sanctions affect job finding rates. *Journal of Applied Econometrics*, *28*(2), 275–296.

van Ours, J. (2004). The locking-in effect of subsidized jobs. *Journal of Comparative Economics*, *32*, 37–55.

Van Reenen, J. (2003). Active labour market policies and the British new deal for the young unemployed in context (Working Paper No. 9576). NBER, Cambridge, MA.

Westergard-Nielsen, N. (1993). The effects of training: A fixed effect model. In K. Jensen & P. Masden (Eds.), *Measuring labour market measures* (pp. 167–200). Copenhagen: Danish Ministry of Labour.

Winter-Ebmer, R., & Zweimüller, J. (2003). On-the-job-training, job search and job mobility. *Swiss Journal of Economics and Statistics*, *139*(4), 563–576.

Wooldridge, J. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). Cambridge, MA: MIT Press.

Yavas, A. (1994). Middlemen in bilateral search markets. *Journal of Labor Economics*, *12*, 406–429.

Mathematical Appendices

The purpose of these appendices is to set out in detail the main mathematical materials the reader needs in order to be able to follow the technical reasoning in certain chapters of this book. They deal with static and dynamic optimization, random variables, and Poisson processes.

1 APPENDIX A: STATIC OPTIMIZATION

In this appendix, we establish heuristically the results that must be determined in order to solve a problem of static optimization. For a more rigorous exposition, readers are advised to consult works such as Hoy et al. (2011) and Carter (2001).

1.1 Unconstrained and Constrained Maximum

In economics, many optimization problems occur in the form:

$$\max_{(C_1,\ldots,C_n)} U(C_1,\ldots,C_n) \tag{1}$$

subject to constraint:

$$\Phi(C_1,\ldots,C_n) \leq R \tag{2}$$

In this problem, U and Φ are twice continuously differentiable real-valued functions defined on $\mathbb{R}^n$. Criterion U, for example, represents the utility of a consumer, and the variables $(C_1,\ldots,C_n)$ are then his consumption of different goods. In this interpretation, parameter R designates the income of the consumer, and the inequality (2) is identified as his budget constraint.

In a first phase, let us set the constraint (2) to one side and simply consider the unconstrained maximum of the problem (1). Its solutions, denoted C_i^* for $i = 1,\ldots,n$, satisfy equations:

$$\frac{\partial U}{\partial C_i} = 0 \quad \text{for } i = 1,\ldots,n \tag{3}$$

For vector $(C_1^*, \ldots, C_n^*)$ to be a solution of problem (1) subject to the budget constraint (2), it is necessary that $\Phi(C_1^*, \ldots, C_n^*) \leq R$. If this inequality is not satisfied, it is certain that the constraint (2) will be binding at the optimum of problem (1) and so will be written $\Phi(C_1, \ldots, C_n) = R$. Let us assume that using this last equality, we can express variable C_1 as a function of the vector $(C_2, \ldots, C_n)$, that is, $C_1 = \Psi(C_2, \ldots, C_n)$. Problem (1) thus becomes:

$$\max_{(C_2,\ldots,C_n)} U\left[\Psi(C_2, \ldots, C_n), C_2, \ldots, C_n\right]$$

The solutions $(\bar{C}_2, \ldots, \bar{C}_n)$ of this problem are then implicitly defined by the equations:

$$\frac{\partial \Psi}{\partial C_i}\frac{\partial U}{\partial C_1} + \frac{\partial U}{\partial C_i} = 0 \quad \text{for } i = 2, \ldots, n \tag{4}$$

with:

$$\bar{C}_1 \equiv \Psi(\bar{C}_2, \ldots, \bar{C}_n) \Longleftrightarrow \Phi\left[\Psi(\bar{C}_2, \ldots, \bar{C}_n), \bar{C}_2, \ldots, \bar{C}_n\right] \equiv R \tag{5}$$

The derivation of the second equality appearing in (5) gives $\partial\Psi/\partial C_i = -(\partial\Phi/\partial C_i)/(\partial\Phi/\partial C_1)$, and if we bring this last relation into (4) we find that the vector $(\bar{C}_1, \ldots, \bar{C}_n)$ is characterized by:

$$\frac{\partial U}{\partial C_i}\Big/\frac{\partial U}{\partial C_1} = \frac{\partial \Phi}{\partial C_i}\Big/\frac{\partial \Phi}{\partial C_1} \quad \forall i = 1, \ldots, n \quad \text{with} \quad \Phi(\bar{C}_1, \ldots, \bar{C}_n) = R \tag{6}$$

Relations (3) and (4) are called the *first-order conditions* of the maximization problem (1) subject to constraint (2). These are the *necessary* conditions for vector $(\bar{C}_1, \ldots, \bar{C}_n)$ to be a local maximum of problem (1). They become sufficient when functions U and Φ are concave.

1.2 THE TECHNIQUE OF THE LAGRANGIAN

The *Lagrangian L* relative to problem (1) subject to constraint (2) is defined by:

$$L(C_1, \ldots, C_n, \lambda) = U(C_1, \ldots, C_n) + \lambda\left[R - \Phi(C_1, \ldots, C_n)\right]$$

Variable λ is called the Lagrange (or Kuhn and Tucker) *multiplier* associated with constraint (2). We will show that we return to the first-order conditions (3) and (4) if we set the partial derivatives of the Lagrangian to zero with respect to variables C_i, that is, $(\partial L/\partial C_i) = 0$ for all $i = 1, \ldots, n$, and take into account the *complementary-slackness* condition:

$$\lambda\left[R - \Phi(C_1, \ldots, C_n)\right] = 0 \quad \text{with} \quad \lambda \geq 0 \tag{7}$$

We thus have:

$$\frac{\partial L}{\partial C_i} = 0 \Longleftrightarrow \frac{\partial U}{\partial C_i} = \lambda \frac{\partial \Phi}{\partial C_i} \quad \forall i = 1, \ldots, n \tag{8}$$

If the budget constraint is not binding, we have $R > \Phi(C_1, \ldots, C_n)$ and the complementary-slackness condition (7) then dictates $\lambda = 0$. That being so, equation (8) is identical to the first-order condition (3) for an "unconstrained" maximum of problem (1). Conversely, if constraint (2) is binding, we have $R = \Phi(C_1, \ldots, C_n)$ and (8) entails $(\partial U/\partial C_1) = \lambda(\partial \Phi/\partial C_1)$. Eliminating the multiplier λ between this last equality and relation (8) for $i \neq 1$, we come back to the first-order conditions (6) for a constrained optimum.

1.3 THE INTERPRETATION OF THE LAGRANGE MULTIPLIERS

Multiplier λ is very easy to interpret by considering the variations of the optimal value of criterion $U(C_1, \ldots, C_n)$ when parameter R changes. Let us assume that the budget constraint (2) is binding; we then have:

$$\sum_{i=1}^{n} \frac{\partial \Phi}{\partial C_i} \frac{\partial C_i}{\partial R} = 1$$

Using this last equality and the first-order conditions (8), we get:

$$\frac{\partial U}{\partial R} = \sum_{i=1}^{n} \frac{\partial U}{\partial C_i} \frac{\partial C_i}{\partial R} = \sum_{i=1}^{n} \lambda \frac{\partial \Phi}{\partial C_i} \frac{\partial C_i}{\partial R} = \lambda$$

The Lagrange multiplier λ thus represents the increase in the criterion $U(C_1, \ldots, C_n)$ when constraint (2) is "relaxed" by one unit. In a sense, it measures the "weight" of this constraint, which is why it is also called the *shadow price*, or the *shadow value*, of budget constraint (2). If the latter is not binding, its shadow value is null, since the complementary-slackness condition (7) dictates $\lambda = 0$.

1.4 SUMMARY AND PRACTICAL GUIDE TO STATIC OPTIMIZATION

When faced with a problem of the form:

$$\max_{(C_1, \ldots, C_n)} U(C_1, \ldots, C_n) \tag{9}$$

subject to constraints:

$$\Phi_j(C_1, \ldots, C_n) \leq R_j, \quad j = 1, \ldots, m \tag{10}$$

these are the steps to follow.

1. Attribute a multiplier λ_j to every constraint (10) and write the Lagrangian:

$$L = U(C_1, \ldots, C_n) + \sum_{i=1}^{n} \lambda_j \left[R_j - \Phi_j(C_1, \ldots, C_n) \right]$$

2. Set the derivatives of the Lagrangian to zero with respect to choice variables C_i:

$$\frac{\partial L}{\partial C_i} = \frac{\partial U}{\partial C_i} - \sum_{j=1}^{m} \lambda_j \frac{\partial \Phi_j}{\partial C_i} = 0 \quad \text{for } i = 1, \ldots, n \tag{11}$$

3. Write the complementary-slackness condition:

$$\lambda_j \left[R_j - \Phi_j(C_1, \ldots, C_n) \right] = 0 \quad \text{with } \lambda_j \geq 0 \quad \forall j = 1, \ldots, m \tag{12}$$

4. The first-order conditions of problem (1) are found by eliminating the Lagrange multipliers λ_j between relations (11) and (12).

5. Relations (11) and (12) are *necessary* conditions of optimality. The solution must also satisfy the *second-order conditions* in order to be a maximum. The second-order conditions are satisfied if functions $U(C_1, \ldots, C_n)$ and $\Phi_j(C_1, \ldots, C_n)$ are concave. More detail about second-order conditions will be found in Hoy et al. (2011) and Carter (2001).

1.5 THE ENVELOPE THEOREM

Let us return to problem (1) subject to constraint (2) and let us assume that the criterion U and the function Φ appearing in the constraint both depend on a parameter a. Problem (1) is then written:

$$\max_{(C_1, \ldots, C_n)} U(C_1, \ldots, C_n, a) \tag{13}$$

subject to constraint:

$$\Phi(C_1, \ldots, C_n, a) \leq R \tag{14}$$

Let us denote by $C_i(a)$ the solutions of this problem for $i = 1, \ldots, n$, and let us designate by $V(a)$ the value of criterion U at the optimum, that is, $V(a) = U\left[C_1(a), \ldots, C_n(a), a\right]$. We then have:

$$V'(a) = \sum_{i=1}^{n} \frac{\partial U}{\partial C_i} \frac{\partial C_i}{\partial a} + \frac{\partial U}{\partial a} \tag{15}$$

Let us suppose in a first stage that problem (13) is not subject to constraint (14). That being so, the first-order conditions are given by $\frac{\partial U}{\partial C_i} = 0$ for $i = 1, \ldots, n$, and equation (15) is written:

$$\frac{\partial U}{\partial a} = V'(a)$$

This equality is known as the envelope theorem. It signifies that in order to find the variations of the value function $V(a)$ of problem (13) with respect to parameter a, it suffices to focus on the *partial* derivative of criterion U with respect to this parameter.

The envelope theorem takes a noticeably different form when there exists a constraint of the type (14). The Lagrangian then takes the place of criterion U. The Lagrangian relative to problem (13) reads:

$$L[C_1(a), \ldots, C_n(a), a, \lambda] = U(C_1, \ldots, C_n, a) + \lambda [R - \Phi(C_1, \ldots, C_n, a)] \quad \textbf{(16)}$$

At the optimum, the multiplier λ is also a function of parameter a. Let us assume that function $\lambda(a)$ is derivable (piece-wise at least). Deriving the Lagrangian (16) with respect to a, we arrive at:

$$\frac{\mathrm{d}L}{\mathrm{d}a} = \sum_{i=1}^{n} \left(\frac{\partial U}{\partial C_i} - \lambda \frac{\partial \Phi}{\partial C_i} \right) \frac{\partial C_i}{\partial a} + \frac{\partial U}{\partial a} + \lambda'(a)(R - \Phi) - \lambda \frac{\partial \Phi}{\partial a}$$

The first-order conditions are again defined by (8), which entails:

$$\frac{\mathrm{d}L}{\mathrm{d}a} = \frac{\partial U}{\partial a} + \lambda'(a)(R - \Phi) - \lambda \frac{\partial \Phi}{\partial a}$$

At the optimum of the problem (13), the complementary-slackness conditions entail:

$$\lambda \{R - \Phi [C_1(a), \ldots, C_n(a), a]\} = 0 \quad \text{with} \quad \lambda \geq 0$$

For the values of parameter a for which $\lambda(a) > 0$, we have $R - \Phi = 0$. When the constraint is not binding, we have $\lambda(a) = 0$ and thus $\lambda'(a) = 0$ for these values of parameter λ. Consequently, we always have $\lambda'(a)(R - \Phi) = 0$ and we can therefore deduce that:

$$\frac{\mathrm{d}L}{\mathrm{d}a} = \frac{\partial U}{\partial a} - \lambda \frac{\partial \Phi}{\partial a} = \frac{\partial L}{\partial a} \quad \textbf{(17)}$$

Moreover, at the optimum of the problem (13) the complementary-slackness conditions are satisfied. Thus we have $L = U[C_1(a), \ldots, C_n(a), a]$, which entails, using (15):

$$\frac{\mathrm{d}L}{\mathrm{d}a} = \sum_{i=1}^{n} \frac{\partial U}{\partial C_i} \frac{\partial C_i}{\partial a} + \frac{\partial U}{\partial a} = V'(a)$$

With the help of (17), we finally arrive at:

$$\frac{\partial L}{\partial a} = V'(a)$$

This equality constitutes the (generalized) envelope theorem. It signifies that in order to find the variations in the value function of problem (1) with respect to parameter a, it suffices to focus on the *partial* derivative of the Lagrangian with respect to this parameter.

2 APPENDIX B: DYNAMIC OPTIMIZATION

As with the preceding appendix, we do not provide an exhaustive account of this matter here. But we do present, in an intuitive fashion, the results and techniques with which one must be familiar in order to work through a problem of dynamic optimization. For a more rigorous approach, readers may turn to Gandolfo (2010) and Hoy et al. (2011).

2.1 THE OPTIMAL CONTROL PROBLEM

In economics, problems of dynamic optimization in continuous time most often occur in the form:

$$\max_{C(t)} \int_0^T U\,[K(t), C(t), t]\,\mathrm{d}t \tag{18}$$

subject to constraints:

$$\dot{K}(t) = G\,[K(t), C(t), t] \tag{19}$$

$$K(0) = K_0 \text{ given} \tag{20}$$

$$K(T) \geq 0 \tag{21}$$

Parameter T represents the terminal date, which may be infinite. Variable $K(t)$ is the *state variable*, serving to describe the evolution of the system under scrutiny. Variable $C(t)$ is the *control variable*, and in the majority of problems it is identified with the decisions made by an agent. The instantaneous criterion U is generally a function describing the utility of a consumer, or the profit of a firm, or a social welfare function. Since program (18) consists of finding control variables which maximize a well-specified intertemporal objective, this program is also called the *optimal control* problem. Equation (19) describes the interactions between the control variables and the state variables and is known as the *transition equation*, or the equation of motion. It may, for example, describe the accumulation of capital within a firm. Equality (20) specifies the *initial condition*, declaring that the value $K(0)$ of the state variable at the initial date $t = 0$ is a known datum K_0. Finally, inequality (21) is a *terminal condition* which dictates that the final value $K(T)$ of the state variable is either positive or null. It means, for example, that an agent does not have the right to leave his debts to his descendants.

2.2 THE FIRST-ORDER CONDITIONS

We will establish, in a manner more intuitive than rigorous, the first-order conditions of problem (18). For that, we will rely on the technique of Lagrange multipliers developed in appendix A on static optimization. Let us, at every date t, link a multiplier $\lambda(t)$ to the transition equation (19). Let us also link a multiplier μ to the terminal condition (21). In this context, $\lambda(t)$ is called a *dynamic multiplier*, or *costate variable*. The Lagrangian of problem (18) is then written as follows:

$$L = \int_0^T U[K(t), C(t), t]\,\mathrm{d}t + \int_0^T \lambda(t)\left\{G[K(t), C(t), t] - \dot{K}(t)\right\}\mathrm{d}t + \mu K(T)$$

This expression is distinguished from a "static" Lagrangian by the appearance of the derivative $\dot{K}(t)$ of the state variable. It is possible to eliminate this derivative by integrating by parts[1] the term in which $\dot{K}(t)$ is found. We thus have:

$$\int_0^T \lambda(t)\dot{K}(t)\mathrm{d}t = [\lambda(t)K(t)]_0^T - \int_0^T K(t)\dot{\lambda}(t)\mathrm{d}t$$

After regrouping terms, the Lagrangian takes the form:

$$L = \int_0^T \{U[K(t), C(t), t] + \lambda(t)G[K(t), C(t), t]\}\,\mathrm{d}t + \int_0^T K(t)\dot{\lambda}(t)\mathrm{d}t + \lambda(0)K_0 - [\lambda(T) - \mu]K(T)$$

Function $H = U + \lambda G$ appearing in the first integral of the Lagrangian is called the *Hamiltonian* of problem (18). By analogy with the static problem studied in appendix A, the first-order conditions are found by setting the derivatives of the Lagrangian L to zero with respect to variables $C(t)$ and $K(t)$ for all t comprised between 0 and T. Thus we have:

$$\frac{\partial L}{\partial C(t)} = 0 \Longleftrightarrow \frac{\partial H}{\partial C(t)} = 0 \tag{22}$$

$$\frac{\partial L}{\partial K(t)} = 0 \Longleftrightarrow \frac{\partial H}{\partial K(t)} + \dot{\lambda}(t) = 0 \tag{23}$$

$$\frac{\partial L}{\partial K(T)} = 0 \Longleftrightarrow \frac{\partial H}{\partial K(T)} + \dot{\lambda}(T) + \lambda(T) - \mu = 0 \tag{24}$$

Condition (23) is called the *maximum principle*. It indicates that, at the optimum, the derivative of the Hamiltonian with respect to the control variable must be null for

[1] Readers are reminded that the integration by parts formula is:

$$\int_a^b u\mathrm{d}v = [uv]_a^b - \int_a^b v\mathrm{d}u$$

all t. The set formed by transition equations (19) and condition (23) is known as the *Euler equations*. Finally, equality (24) expresses the terminal condition of the optimization problem. Now, as we saw in appendix A, the optimal solutions must satisfy the complementary-slackness conditions (7). These conditions here dictate $\mu K(T) = 0$ in particular. By continuity, relation (23) is true in $t = T$. Using (24), we thus obtain the *transversality condition*:

$$\lambda(T)K(T) = 0 \tag{25}$$

By analogy with the static case, the multiplier $\lambda(t)$ is interpreted as the shadow price, assessed at date $t = 0$, of an extra unit of the state variable at date t. The transversality condition (25) thus means that if the terminal date $K(T)$ is strictly positive, its shadow price is necessarily null. Conversely, if $\lambda(T) > 0$, the final stock $K(T)$ is equal to 0.

2.3 INFINITE HORIZON

We move from problem (18), where the horizon is finite, to one with an infinite horizon by making the terminal date T tend to infinity. The transition equation (19) and the initial condition (20) remain unchanged, but the terminal condition (21) is now written:

$$\lim_{t \to +\infty} K(t) \geq 0$$

The first-order conditions (22) and (23) remain unchanged, but we make $T \to +\infty$ in (25), so the transversality condition now takes the form:

$$\lim_{t \to +\infty} \lambda(t)K(t) = 0 \tag{26}$$

If, for example, $K(t)$ represents a stock of capital increasing at constant rate g, relation (26) entails that the costate variable—the shadow price of capital—must tend to 0 at a rate greater than g. In fact, notwithstanding the intuitive nature of this result, Michel (1982) has shown that the solutions of the dynamic optimization problem with an infinite horizon are not obliged to satisfy equality (26). The "real" transversality condition would be $\lim_{t \to +\infty} H(t) = 0$, with equation (26) being a sufficient condition. In the majority of problems dealt with in economics, it is quite easy to ensure that condition (26) is satisfied.

2.4 CALCULUS OF VARIATIONS AND THE EULER EQUATION

We sometimes encounter problems of dynamic optimization having the particular form:

$$\max_{K(t)} \int_0^T U\left[K(t), \dot{K}(t), t\right] \mathrm{d}t \tag{27}$$

Here the only constraints are the initial and terminal conditions (20) and (21). This might be a case, as in chapter 2 for example, of intertemporal profit maximization in a firm bearing adjustment costs linked to variations $\dot{K}(t)$ in the state variable.

Program (27) is often referred to as a problem of "calculus of variations." Formally, we move from the optimal control problem (18) to the calculus of variations problem (27) by taking the transition equation (19) as being simply written $\dot{K}(t) = C(t)$. That being so, the Hamiltonian of problem (18) is given by $H = U + \lambda C$, and the maximum principle (22) entails:

$$\frac{\partial H}{\partial C(t)} = \frac{\partial U}{\partial C(t)} + \lambda(t) = 0 \tag{28}$$

The Euler equation (23) is here written:

$$\frac{\partial H}{\partial K(t)} + \dot{\lambda}(t) = \frac{\partial U}{\partial K(t)} + \dot{\lambda}(t) = 0$$

Deriving relation (28) with respect to t and bearing in mind that $C(t) = \dot{K}(t)$, we get:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{\partial U}{\partial \dot{K}(t)}\right] + \dot{\lambda}(t) = 0$$

Eliminating $\dot{\lambda}(t)$ between the last two equations, in the end we find:

$$\frac{\partial U}{\partial K(t)} = \frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{\partial U}{\partial \dot{K}(t)}\right] \tag{29}$$

This condition, which is likewise known as the *Euler equation*, yields a differential equation characterizing the optimal trajectory of the variable $K(t)$. The transversality conditions (25) and (26) remain valid.

2.5 Summary and Practical Guide to Optimal Control

Let us consider the dynamic optimization problem with n control variables $C_1(t), \ldots, C_n(t)$, and m state variables $K_1(t), \ldots, K_m(t)$, and with the form:

$$\max_{\{C_1(t),\ldots,C_n(t)\}} \int_0^T U\left[K_1(t), \ldots, K_m(t); C_1(t), \ldots, C_n(t), t\right] \mathrm{d}t \quad \text{with} \quad T \leq +\infty$$

subject to constraints:

$$\dot{K}_j(t) = G_j\left[K_1(t), \ldots, K_m(t); C_1(t), \ldots, C_n(t), t\right] \quad \forall j = 1, \ldots, m \tag{30}$$

$$K_j(0) = K_{j0} \text{ given} \quad \forall j = 1, \ldots, m$$

$$K_j(T) \geq 0 \quad \text{or} \quad \lim_{t \to +\infty} K_j(t) \geq 0 \quad \forall j = 1, \ldots, m$$

Readers are advised to follow these steps (the index t is most often omitted in order to simplify the notation):

1. Attribute a costate variable $\lambda_j(t)$ to each transition equation (30) and write the Hamiltonian:

$$H = U(K_1, \ldots, K_m; C_1, \ldots, C_n, t) + \sum_{j=1}^{m} \lambda_j G_j(K_1, \ldots, K_m; C_1, \ldots, C_n, t)$$

2. Apply the maximum principle, which amounts to setting the partial derivatives of the Hamiltonian to zero with respect to the control variables, that is:

$$\frac{\partial H}{\partial C_i} = 0 \quad \forall i = 1, \ldots, n \tag{31}$$

3. Write the Euler equations:

$$\frac{\partial H}{\partial K_j} = -\dot{\lambda}_j \quad \text{with} \quad \dot{K}_j = G_j(K_1, \ldots, K_m; C_1, \ldots, C_n, t), \ \forall j = 1, \ldots, m \tag{32}$$

4. Relations (31) and (32) make it possible to arrive at a system of differential equations in λ_j and K_j. The resolution of this system gives the optimal trajectories of the state variables K_j.

5. Do not forget to verify the transversality conditions, which, according to whether the horizon is finite or infinite, are written:

$$\lambda_j(T)K_j(T) = 0 \quad \text{or} \quad \lim_{t \to +\infty} \lambda_j(t)K_j(t) = 0 \quad \forall j = 1, \ldots, m$$

6. The maximum principle (31) and the Euler equations (32) are necessary conditions of optimality. They become sufficient if functions U and G_j are concave.

3 APPENDIX C: BASIC NOTIONS CONCERNING RANDOM VARIABLES

For appendices C and D, supplementary information can be found in Ross (2010).

3.1 Random Variables and Probability Densities

A *discrete* random variable (henceforth r.v.) X is characterized by the set of all its possible realizations $(x_1, \ldots, x_i, \ldots, x_n)$ where n is extendable to infinity, and the probabilities $(p_1, \ldots, p_i, \ldots, p_n)$ linked to its realizations. These probabilities are evidently such that

$\sum_{i=1}^{n} p_i = 1$. The *mathematical expectation* (or the *mean*), denoted $\mathbb{E}(X)$, of this r.v. is defined by:

$$\mathbb{E}(X) = \sum_{i=1}^{n} p_i x_i$$

The *variance* $V(X)$ and the *standard deviation* $\sigma(X)$ are rudimentary indicators of the dispersion of the values of r.v. X around its average. They are given by the formulas:

$$V(X) = \sum_{i=1}^{n} p_i \left[x_i - \mathbb{E}(X)\right]^2 = \left(\sum_{i=1}^{n} p_i x_i^2\right) - \mathbb{E}^2(X) \quad \text{and} \quad \sigma(X) \equiv \sqrt{V(X)}$$

A *continuous* r.v., still denoted X, is defined over an interval $[a, b]$ of the set of real numbers; bounds a and b can be infinite. A continuous r.v. is characterized by its *probability density*, denoted $f(x)$, which is a function greater than or equal to 0 defined over $[a, b]$. Let us consider a small interval $[x, x + \mathrm{d}x]$ belonging to segment $[a, b]$: intuitively, quantity $f(x)\mathrm{d}x$ is equivalent to probability p_i for a discrete variable; it represents the probability that the realizations of the continuous r.v. X lie in the interval $[x, x + \mathrm{d}x]$. The probability density is such that $\int_a^b f(x)\mathrm{d}x = 1$ and the mathematical expectation is defined by the formula:

$$\mathbb{E}(X) = \int_a^b x f(x)\mathrm{d}x$$

The *cumulative distribution function*, denoted $F(x)$, measures the probability of event $\{X \leq x\}$ for a given value of x. We thus have:

$$F(x) = \Pr\{X \leq x\} = \int_a^x f(\xi)\mathrm{d}\xi \Longleftrightarrow F'(x) = f(x)$$

Finally, the variance $V(X)$ and the standard deviation $\sigma(X)$ of a continuous r.v. are again defined by:

$$V(X) = \sigma^2(X) = \mathbb{E}\left[X - \mathbb{E}(X)\right]^2 = \mathbb{E}(X^2) - \mathbb{E}^2(X)$$

3.2 Independence and Correlation

Let us consider two discrete r.v., with realizations and probability distributions respectively denoted $\{x_i;\, i = 1, \ldots, n\}, \{y_j;\, j = 1, \ldots, m\}$ and $\{p_i;\, i = 1, \ldots, n\}, \{q_j;\, j = 1, \ldots, m\}$. Intuitively, these r.v. are *independent* if the observation of the realization of one of them gives no indication about the realization of the other. More formally, this means that events $\{X = x_i\}$ and $\{Y = y_j\}$ are disjunct $\forall(i, j)$. That being the case, we can write:

$$\Pr\{X = x_i \text{ and } Y = y_j\} = \Pr\{X = x_i\} \cdot \Pr\{Y = y_j\} \quad \forall(i, j) \qquad \textbf{(33)}$$

By definition, the expectation of product XY is given by:

$$\mathbb{E}(XY) = \sum_{i,j} x_i y_j \Pr\{X = x_i \text{ and } Y = y_j\}$$

Taking account of (33), we get:

$$\begin{aligned}\mathbb{E}(XY) &= \sum_{i,j} x_i y_j \Pr\{X = x_i\} \cdot \Pr\{Y = y_j\} \\ &= \left(\sum_i x_i \Pr\{X = x_i\}\right)\left(\sum_j y_j \Pr\{Y = y_j\}\right) = \mathbb{E}(X)\mathbb{E}(Y) \end{aligned} \quad (34)$$

Hence, when two discrete r.v. are independent, the expectation $\mathbb{E}(XY)$ of the product is equal to the product $\mathbb{E}(X)\mathbb{E}(Y)$ of the expectations. This property holds true for continuous r.v. Conversely, when two r.v. are not independent, the properties (33) and (34) are no longer verified. The *covariance* $\text{Cov}(X, Y)$ and the *correlation coefficient* $\rho(X, Y)$ allow us to assess the direction and degree of the dependence between two r.v.; they are defined by:

$$\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \quad \text{and} \quad \rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma(X)\sigma(Y)}$$

Note that if $\text{Cov}(X, Y) = 0$, the random variables are not necessarily independent (except if they are normal variables). Coefficient $\rho(X, Y)$ takes its values over the interval $[-1, +1]$.

Given two r.v., X and Y, and parameters a, b, and c, the expectation and variance operators satisfy the following properties:

$$\mathbb{E}(aX + bY + c) = a\mathbb{E}(X) + b\mathbb{E}(Y) + c$$

$$V(aX + bY + c) = a^2 V(X) + b^2 V(Y) + 2ab\text{Cov}(X, Y)$$

3.3 Some Common Probability Distributions

Uniform Distribution

The probability density and the cumulative distribution function of a uniform r.v. X defined over the interval $[a, b]$ are given by:

$$f(x) = \frac{1}{b-a} \quad \text{and} \quad F(x) = \frac{x-a}{b-a}$$

We then easily calculate:

$$\mathbb{E}(X) = \frac{a+b}{2} \quad \text{and} \quad V(X) = \frac{(b-a)^2}{12}$$

EXPONENTIAL DISTRIBUTION

We say that a r.v. X follows an exponential distribution with parameter $\lambda > 0$ over the interval $[0, +\infty)$, when it has the probability density:

$$f(x) = \lambda e^{-\lambda x}$$

Its cumulative distribution function is then given by:

$$F(x) = \int_0^x \lambda e^{-\lambda \xi} \mathrm{d}\xi = 1 - e^{-\lambda x}$$

with:

$$\mathbb{E}(X) = \frac{1}{\lambda} \quad \text{and} \quad V(X) = \frac{1}{\lambda^2}$$

The exponential distribution comes into the definition of the Poisson process in particular (see appendix D below).

NORMAL DISTRIBUTION

A r.v. X follows a normal distribution with mean μ and standard deviation σ; we use the notation $X \rightsquigarrow \mathcal{N}(\mu, \sigma)$, when its probability density is defined over $(-\infty, +\infty)$ by the function:

$$f(x) = \frac{1}{\sigma\sqrt{2\Pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{35}$$

Readers may, as an exercise, verify that the average and the standard deviation of a r.v. having the function (35) for its density are effectively equal to μ and σ.

LOG-NORMAL DISTRIBUTION

The r.v. X follows a log-normal distribution with parameters (x_0, μ, σ) over the interval $[x_0, +\infty)$ if the r.v. $\ln(X - x_0)$ follows the normal distribution $\mathcal{N}(\mu, \sigma)$. In other words, if $Z \rightsquigarrow \mathcal{N}(\mu, \sigma)$, X is also defined by the equality $X = x_0 + e^Z$. Its probability density is then given by:

$$f(x) = \frac{1}{\sigma(x - x_0)\sqrt{2\Pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln(x - x_0) - \mu}{\sigma}\right)^2\right], \quad \forall x \geq x_0$$

We can then calculate the expectation and the standard deviation; they come to:

$$\mathbb{E}(X) = x_0 + \exp\left(\mu + \frac{\sigma^2}{2}\right) \quad \text{and} \quad \sigma(X) = \sqrt{1 - \exp(-\sigma^2)}\exp\left(\mu + \frac{\sigma^2}{2}\right)$$

4 APPENDIX D: THE POISSON PROCESS AND THE VALUE OF AN ASSET

In models in continuous time, we often assume that certain random events follow a Poisson process. With this hypothesis, the probability of these events occurring (or enduring) depends on a set of parameters having a precise economic significance. Moreover, it turns out that the equation describing the evolution of the value of an asset whose states change according to a Poisson process takes a simple analytical form.

4.1 THE POISSON PROCESS

Given a series of parameters $\lambda(t) \geq 0$, defined for $t \in [0, +\infty)$, we say that an event X (for example, the occurrence of a productivity shock) follows a Poisson process with parameters $\{\lambda(t)\}$ if the duration $T(t)$, starting from date t, that it is necessary to wait for X to occur is a random variable having an exponential cumulative distribution function defined by:

$$F_t(y) \equiv \Pr\{T(t) \leq y\} = 1 - e^{-\int_t^{t+y} \lambda(\xi)\mathrm{d}\xi}$$

The probability density of the random variable $T(t)$ then takes the form:

$$f_t(y) = F_t'(y) = \lambda(t+y)e^{-\int_t^{t+y} \lambda(\xi)\mathrm{d}\xi} \tag{36}$$

Making y tend to 0 in this relation, we see that parameter $\lambda(t)$ is interpreted as the *instantaneous* probability of the realization of event X at date t. When the parameters take the same value at every date, which amounts to setting $\lambda(t) = \lambda$ for all $t \geq 0$, the r.v. $T(t)$ no longer depends on date t. The Poisson process is "stationary" and the cumulative distribution function and the probability density are then written simply:

$$F(y) = 1 - e^{-\lambda y} \quad \text{and} \quad f(y) = \lambda e^{-\lambda y}$$

The unconditional expectation $\mathbb{E}[T(t)]$ of the r.v. $T(t)$ is identifiable as the *average* duration which it is necessary to wait, starting from date t, for event X to occur. This expression takes a particularly interesting form when the parameter of the Poisson process is constant. With this hypothesis, let T simply be the r.v. $T(t)$; it comes to:

$$\mathbb{E}(T) = \int_0^{\infty} y\lambda e^{-\lambda y}\mathrm{d}y = \frac{1}{\lambda}$$

The ratio $(1/\lambda)$ thus represents the average duration of the event studied. If, for example, λ represents the instantaneous probability (assumed constant) that an unemployed person finds a job every week, the ratio $(1/\lambda)$ represents the average duration of unemployment, measured in weeks.

4.2 EVOLUTION OF THE VALUE OF AN ASSET

We will determine the value of an asset (for example, a filled job) which, at every date x, can either bring in an instantaneous income $\omega(x)$ or change state (become vacant for example). This change of state is a random event which follows a Poisson process with parameters $\{\lambda(t)\}$. The duration $T(t)$ which it is necessary to wait, starting at date t, for this change of state to occur, is thus a r.v. the probability density of which is the function $f_t(.)$ defined by relation (36). We will assume further that if the asset changes state at instant $(t+y)$, its present discounted value at that date is a known quantity denoted by $\bar{\Pi}(t+y)$. Assuming that the interest rate is an exogenous constant r, the present discounted value at date t of the asset, $\Pi(t)$, is written:

$$\Pi(t) = \mathbb{E}\left\{\int_t^{t+T(t)} \omega(x)e^{-r(x-t)}\mathrm{d}x + e^{-rT(t)}\bar{\Pi}\left[t+T(t)\right]\right\}$$

In this equality, the symbol $\mathbb{E}$ designates the mathematical expectation operator. As the sole r.v. that comes into the term between braces is the duration $T(t)$ with probability density $f_t(.)$, we get:

$$\Pi(t) = \int_0^{\infty}\left\{\left[\int_t^{t+y}\omega(x)e^{-r(x-t)}\mathrm{d}x + e^{-ry}\bar{\Pi}(t+y)\right]\lambda(t+y)e^{-\int_t^{t+y}\lambda(\xi)\mathrm{d}\xi}\right\}\mathrm{d}y \qquad (37)$$

This expression of $\Pi(t)$ can be simplified using the integration by parts formula, $\int u\mathrm{d}v = uv - \int v\mathrm{d}u$. Let us set $u = \int_t^{t+y}\omega(x)e^{-r(x-t)}\mathrm{d}x$ and $\mathrm{d}v = \lambda(t+y)e^{-\int_t^{t+y}\lambda(\xi)\mathrm{d}\xi}\mathrm{d}y$, we then have $\mathrm{d}u = \omega(t+y)e^{-ry}\mathrm{d}y$ and $v = -e^{-\int_t^{t+y}\lambda(\xi)\mathrm{d}\xi}$, and so:

$$\int_0^{\infty}\left[\int_t^{t+y}\omega(x)e^{-r(x-t)}\mathrm{d}x\right]\lambda(t+y)e^{-\int_t^{t+y}\lambda(\xi)\mathrm{d}\xi}\mathrm{d}y = \left[-e^{-\int_t^{t+y}\lambda(\xi)\mathrm{d}\xi}\int_t^{t+y}\omega(x)e^{-r(x-t)}\mathrm{d}x\right]_0^{\infty} + \int_0^{\infty}\omega(t+y)e^{-ry}e^{-\int_t^{t+y}\lambda(\xi)\mathrm{d}\xi}\mathrm{d}y$$

Assuming that the discounted value of incomes $\int_t^{t+y}\omega(x)e^{-r(x-t)}\mathrm{d}x$ is bounded when y tends to infinity, the term between square brackets is null, and equation (37) is rewritten as follows:

$$\Pi(t) = \int_0^{\infty}\left[\omega(t+y) + \lambda(t+y)\bar{\Pi}(t+y)\right]e^{-\int_t^{t+y}[r+\lambda(\xi)]\mathrm{d}\xi}\mathrm{d}y$$

With the change of variable $x = t+y$, we then have:

$$\Pi(t) = \int_t^{\infty}\left[\omega(x) + \lambda(x)\bar{\Pi}(x)\right]e^{-\int_t^{x}[r+\lambda(\xi)]\mathrm{d}\xi}\mathrm{d}x \qquad (38)$$

Deriving this last equation with respect to t, we get:

$$\dot{\Pi}(t) = -\left[\omega(t) + \lambda(t)\bar{\Pi}(t)\right] + \left[r+\lambda(t)\right]\int_t^{\infty}\left[\omega(x) + \lambda(x)\bar{\Pi}(x)\right]e^{-\int_t^{x}[r+\lambda(\xi)]\mathrm{d}\xi}\mathrm{d}x$$

where $\dot{\Pi}(t)$ designates the time derivative of $\Pi(t)$. In the last part of the right-hand side of this equality, we recognize the expression of the discounted value of the asset $\Pi(t)$ given by relation (38). Finally, the evolution of the value of the asset is completely described by the following equation:

$$r\Pi(t) = \omega(t) + \lambda(t)\left[\bar{\Pi}(t) - \Pi(t)\right] + \dot{\Pi}(t) \tag{39}$$

Thus we obtain the asset-value functions or the arbitrage equations used throughout this book.

4.3 AN ALTERNATIVE PROOF

It is possible to arrive at formula (39) in an intuitive manner, proceeding by approximation. Assuming that the asset brings in a flow of income $\omega(t)\mathrm{d}t$ over a small interval of time $\mathrm{d}t$, and that this asset may be destroyed over this small interval of time with a probability $\lambda(t)dt$, the value of the asset is written:

$$\Pi(t) = \frac{1}{1 + r\mathrm{d}t}\left\{\omega(t)\mathrm{d}t + \lambda(t)\mathrm{d}t\bar{\Pi}(t + \mathrm{d}t) + [1 - \lambda(t)\mathrm{d}t]\Pi(t + \mathrm{d}t)\right\}$$

Rearranging the terms of this equality, we get:

$$r\Pi(t) = \omega(t) + \lambda(t)\left[\bar{\Pi}(t + \mathrm{d}t) - \Pi(t + \mathrm{d}t)\right] + \frac{\Pi(t + \mathrm{d}t) - \Pi(t)}{\mathrm{d}t}$$

We have arrived exactly at relation (39) by making $\mathrm{d}t$ tend to 0.

REFERENCES

Azariadis, C. (1993). *Intertemporal macroeconomics*. Hoboken, NJ: Wiley Blackwell.

Carter, M. (2001). *Foundations of mathematical economics*. Cambridge, MA: MIT Press.

Gandolfo, G. (2010). *Economic dynamics* (4th ed.). Paris: Springer Verlag.

Hoy, M., Livernois, J., McKenna, C., Rees, R., & Stengos, T. (2011). *Mathematics for economics*. Cambridge, MA: MIT Press.

Michel, P. (1982). On the transversality condition in infinite horizon optimal problems. *Econometrica*, *50*(4), 975–985.

Ross, S. (2010). *Introduction to probability models* (10th ed.). Salt Lake City, UT: Academic Press.

Name Index

Subject Index