Springer Texts in Business and Economics

Darren Grant

Methods of Economic Research

Craftsmanship and Credibility in Applied Microeconomics

Springer

Springer Texts in Business and Economics

More information about this series at http://www.springer.com/series/10099

Darren Grant

Methods of Economic Research

Craftsmanship and Credibility in Applied Microeconomics

Springer

Darren Grant
Sam Houston State University
Huntsville, TX, USA

ISSN 2192-4333 ISSN 2192-4341 (electronic)
Springer Texts in Business and Economics
ISBN 978-3-030-01733-0 ISBN 978-3-030-01734-7 (eBook)
https://doi.org/10.1007/978-3-030-01734-7

Library of Congress Control Number: 2018959843

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To Gordon Stanton: veteran, middle-school teacher, hero. I didn't know it then, but I do now: you turned me into a social scientist. And to everyone who believed in me, all six or eight or ten of you. Thank you.

Preface: Where I Lived, and What I Lived For

In life epiphanies are few. Certainly none were involved in the writing of this book. It was just a gradual working out of many small ideas. As they became clearer, the connections between them began to appear.

So whatever arrogance you find in the pages that follow should be ascribed only to the infirmities of my temper and to my professional isolation. Among economists, I am not a glitterati or even a paparazzi, just someone who sees the pictures in the next morning's paper. This has given my mind so much room in which to wander that it scarcely recognizes where it started from. What was once familiar now seems less so and is greeted more brusquely than before.

This isolation meant that many of my formative experiences occurred outside the confines of our profession. You will hear little of these. Still, I should mention the most formative experience of all: simply residing in Walker County, Texas. At the crossroads of three cultures—Deep South, Midwestern, and Gulf Coast, with a whiff of the West thrown in—it is large enough to spawn a variety of institutions, yet so small and open that you can take in almost the whole of it. If there is anywhere better for a social scientist to experience American life as it is routinely lived, I have not encountered it. The tendrils of politics, sociology, economics, anthropology, geography, history, and criminology extend so far into private and public life that you cannot go a day without recognizing the effects of each. I perceive everything through this lens.

Toward the end of my journey, I wanted to know if anyone else had gotten to my destination before me. It was clear by that point what the journal of their travels would be called. So I looked for it: four words, two titles, neither taken. All right then.

I wrote this book for a reason many of us share—a latent discontent that develops early in our economics training, a queasy feeling that something is just not right. I know that feeling …

Huntsville, TX, USA — Darren Grant
June, 2018

Acknowledgments

Several people deserve recognition and my heartfelt appreciation for their contributions to this book. First is the team at Springer, including Lorraine, Nick, Kelly, and the rest. They most affected the manuscript itself via a key strategic decision that I dimly perceived was correct long before I understood why.

Several reviewers provided thoughtful, helpful feedback on the manuscript: Mark Anderson, Richard Cox, Craig Depken, John Garen, Daniel Henderson, Barry Hirsch, Venoo Kakar, Katherine Keisler, Daniel Kling, Steve Koch, Jason Lindo, Charlie Sawyer, Rosanna Smart, Tino Sonora, Peter Swann, and Jadrian Wooten. In addition, my de facto copy editor, Jennifer Shirk, cleaned up all manner of sloppiness.

My employer, Sam Houston State University, provided support in two important ways. A faculty development leave in the first half of 2016 disconnected me from the hustle and bustle and let me start writing in earnest. Also, several student assistants helped with figures, supporting material, and other aspects of preproduction: Femi Babalola, Brent Hines, Kevin Southerland, Elizabeth Stokes, Anubhav Thakur, and, especially, Jared Zbranek.

I also mustn't neglect to mention my steadfast partner in this process, the delete key. Oh delete key, I had known of your existence, yet I knew you not. What a boon you have been—what a friend to man!

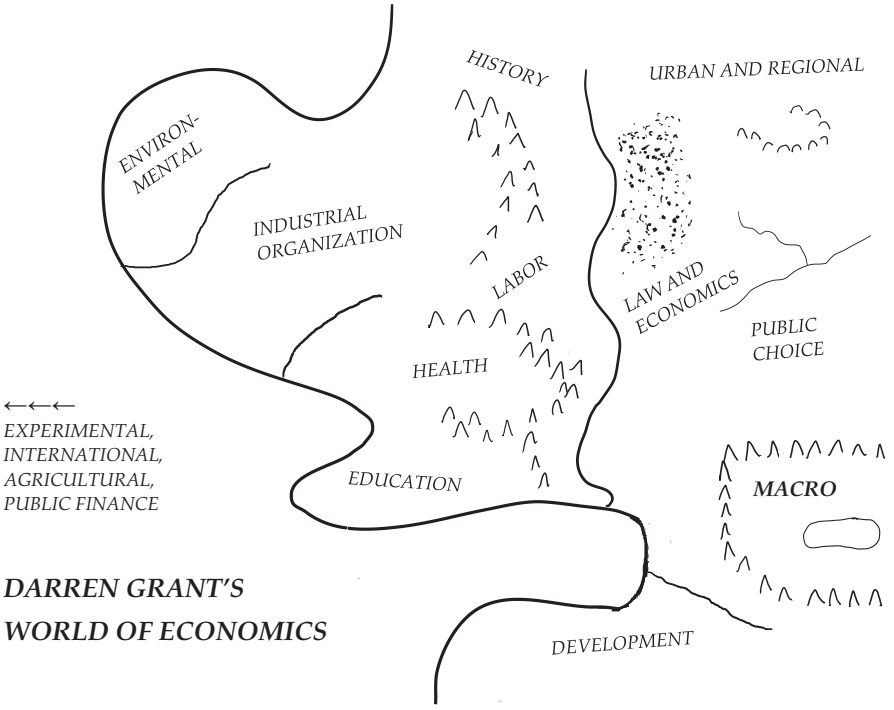
my old friend, delete
what you giveth is also
what you take away

For all the virtues of restraint, however, the most vital element of writing this book, and the hardest for me to come by, was the nerve required to get up yet another day and fight. I wish to credit various artists, musical, visual, and, otherwise, with helping to supply that nerve. I had not before realized that role of art, but I realize it now.

That source of inspiration was surpassed by only one other: my family. My wife, Marsie, is a model of professional dedication. A thousand times I have seen her, papers spread all around, pen in hand. For her grading homework is an act of love.

My son, Sheridan, beams with intellectual passion. He never met a math problem he didn't like. For fearlessness I look to my daughter Tammany, whose first dive off the diving board was a flip, who has the courage of her convictions—both the courage and the convictions. Each of you was a beacon of fortitude impelling me to say what I had to say, not what I ought.

Finally, I must acknowledge my office mates. This book was written mostly in my outdoor workshop, with open air the year round. Each day I had company there. Hummingbirds in the spring at the feeder, and other birds carefully checking the figs, always to eat them the day they were ready to be picked. In the summer a turtle or two, bashfully crunching its way through the leaves outside my window, and spiders setting up shop in the bushes by the door. And, always, a squirrel, who every afternoon would perch on a branch outside and peer down at me, as if to say, "Not finished yet?"—or maybe, now that this book is done: "What's next?"



DARREN GRANT'S
WORLD OF ECONOMICS

Introduction: Let's Get Some Things Out of the Way Right Now

Let's get some things out of the way right now.

Every book is defined by what it leaves out as much as by what it puts in, and there's no reason to be coy about either. Topically, this book concerns applied microeconomics: the analysis of real-world data to answer real-world microeconomic questions. Sometimes this term's scope is more limited, referring only to "the economics of people"—health, labor, and the like. My world is bigger than that, as you can see. Still, some things are beyond my range and largely left out, agricultural, international, and experimental economics among them.

In terms of content, this book is about maximizing the credibility of empirical research findings—only. There are other, more practical objectives to which economists could adhere. Where these conflict with mine, you now know how I will choose.

The book does not teach standard graduate-level econometrics and economic theory. This I assume you already know. If you don't, go put together your toolkit and then come see me. I will help you employ these tools and others in pursuit of the objective stated above.

This approach accords with the principle of comparative advantage. Existing texts do a good job laying out common econometric models and the conditions under which they apply, while the well-developed formalizations of pure theory are represented effectively in any number of journal articles. The main omission, a big one, concerns what makes a good model, in the sense that the term is used below, and when to use various modeling strategies. I regret not being able to do this topic full justice in this book, without having somewhere else to refer you.

Another limitation derives from my inability to fully navigate the large literatures surrounding some of the papers dissected below. For each of these suns, I did my darnedest to fully explore the corona, the larger planets, but rarely made it beyond that. I may have missed something relevant, especially if it was published after the first draft of these chapters was written, in 2016, give or take a few months. It is hard enough to hit a still target, much less a moving one, so the number of post-2016 updates to the material is quite limited. A section of the web site identified below is set aside for you to point out any important omissions.

There critiques of all types are welcomed, to which I will respond when merited. But I will say something now to anyone affronted by my lack of propriety. Perhaps you feel I have omitted something I was supposed to say, or placed a well-received paper in too harsh of light, or otherwise said something altogether unexpected. Fine. Just remember, this book isn't about you. If you're unhappy, write your own book. You can call it *I Only Do What Everyone Says to Do, and Right Now You're Making Me Incredibly Nervous*—if that title isn't already taken.

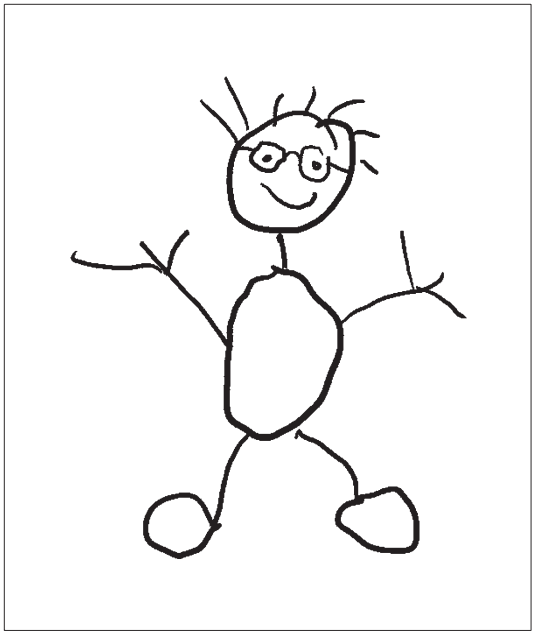
This book excerpts from a wide variety of source material, as you will see. For alacrity and space, these excerpts have often been reworked, with text removed, reordered, or rephrased, but—I hope—the original meaning retained. If the passage is in quotes, it is verbatim. If it is set off from the main text instead, indented without quotes, it is almost always adapted from the original. If a phrase is borrowed from literature (not *the* literature) or film, with nothing to identify it as such, neither quotes nor offset, it is an allusion, like the one in the previous paragraph. I am not trying to sneak it by you and pass it off as my own.

Finally, this book is supplemented by a glossary and a web site. I know firsthand how hard it can be to follow colloquial expressions or unusual words in a nonnative language—and sometimes even in a native one! None of this need remain inscrutable. Thus the glossary.

The web site, www.worldofeconomics.com, is intended to make this book a "living document." It contains ancillary materials, corrections, links of interest, and discussion boards for the "Food for Thought" questions and for key topics broached in this book. I will participate in the discussions on this web site, and I hope you will too. We both still have much to learn.

Contents

About the Author



Darren Grant, *Self Portrait*. Reprinted by permission of the Museum of Modern Art, New York.

Darren Grant is currently an economics professor at Sam Houston State University in Texas. He has previously served as a Navy contractor, a health management program director, and a high school teacher and coach, in various towns and cities across the South. He has published applied research in leading field journals in policy analysis, health economics, the economics of education, industrial organization, public choice, behavioral economics, and labor economics, including the *Journal of Human Resources*, the *Journal of Health Economics*, the *Journal of Policy Analysis and Management*, and the *Journal of Economic Behavior and Organization*. The empirical models used in this research run the gamut of those found in the profession: parametric and nonparametric, linear and nonlinear, reduced form and structural, and everything in between. Much of what he has learned about doing economic research he has learned the hard way.

Chapter 1
Craftsmanship and Credibility in Economic Research

Abstract This introductory chapter argues that applied microeconomic research should be considered a craft, because it cannot be reduced to a fixed set of procedures that guarantee the "right answer" if performed competently. It then takes stock of the various elements, or inputs, that go into economic research, each of which is explored over the course of this book.

The central problem of the economics profession is the problem of credibility. Too often, economists analyzing similar data draw opposing conclusions; celebrated results turn out to be wrong; seemingly well-supported policies turn out to be ineffectual or destructive when put into practice; or the models, findings, and theories on an issue are so divided or impractical that no clear policy can be advanced in the first place.

The consequences are not merely academic. Within the last 20 years, high-profile failures in Russian privatization, electricity deregulation, the onset of and response to the recent financial crisis, and the use of stock options as executive compensation have all been abetted by economic research that has not held up in practice. And for each incident that is visible outside the profession there are many others that are visible only within it.

These problems are sufficiently cemented into the public consciousness that economics is one of the few professions with a negative reputation among the general public. This perception is directly tied to our lack of credibility. The First Law of Economists: For every economist, there exists an equal and opposite economist. The Second Law of Economists: They're both wrong. This joke would not be considered funny if it had no grain of truth. Try it for yourself: substitute any other profession—doctor, astronomer, sociologist—and see if it is still funny. If you don't consider this to be a problem, you should.

This state of affairs is not foisted on the profession by an angry public, it is not an accident or an inevitable consequence of the nature of economic research. It is not a cause for hand wringing (Rosenberg 1992), philosophical meditation (Reder 1999), or calls to abandon economics' central mission of generating empirically supported models that can explain human behavior and guide policy formation (McCloskey 1983). It is a problem to be solved, that is all.

D. Grant, *Methods of Economic Research*, Springer Texts in Business and Economics, https://doi.org/10.1007/978-3-030-01734-7_1

It is a problem that economists have brought on themselves, in ways both small and large. They utter pompous statements reflecting ignorance of, or callousness about, the profession's credibility problems. Paul Samuelson: "I don't care who writes a nation's laws–or crafts its advanced treaties–if I can write its economics textbooks." They prognosticate far beyond the range of their data, as in Thomas Piketty's *Capital in the Twenty-First Century* (2013). They are more insular than are other kinds of social scientists (Fourcade et al. 2015).

But the heart of the matter is this: economists labor under a misconception of how credible economic research is produced. This misconception governs everything about the profession: its over-reliance on formalization and under-reliance on contextualization, its preference for certain types of methods and data, its publishing patterns, its social structure (Baron and Hannan 1994), even its language. Correct this misconception, replace it with something better and more true, and increased credibility will follow.

1.1 Economic Research and the Scientific Method

This misconception is rooted in a place that makes it especially difficult to dislodge: the scientific method. This method utilizes controlled experiments to evaluate hypotheses about phenomena of interest. A simple formalization of the method is as follows. A scientific hypothesis is formed that predicts a cause-effect relationship between a "treatment" and an "outcome." Two groups of people are then assembled, with individuals randomly assigned to each group. One group is given the treatment and the other group is not. The outcome is then measured to determine if it differs in the expected way across the two groups. If not, the scientific hypothesis is rejected.

The scientific method is used for two reasons. The first is that researchers, like everyone else, are subjective, self-serving, weak-willed people with limited knowledge and life experience and imperfect judgement. Not you of course—you're great. But everyone else is inclined to overlook or ignore other potential causes for observed relationships, intuit causal relations where none exist, see patterns in noise, and make methodological choices that subtly increase the chances of things working out a certain way. The scientific method provides the structure needed to prevent these traits from contaminating behavioral research, just as dirty lab equipment would contaminate chemical research. Because of this, above all else, the scientific method works: when properly implemented, it reliably yields accurate conclusions. That is the second reason it is used.

Of course, in economic research it is often infeasible to implement both lynchpins of the scientific method—randomization and application of a treatment by the researcher. Generally neither can be implemented. Then the only remaining structure consists of the formalisms of theory and econometrics—the two "pillars of the profession." Accordingly, economics relies heavily upon these pillars. This contributes greatly to the sense that our field is the "most scientific" of the social sciences, and has done a lot of good in numerous areas. One has only to compare, for example, economists'

research in health policy and education policy with that of more traditional scholars in those fields to appreciate the value of that structure and formalization.

But the heavy reliance on these pillars also contributes to the misconception of how credible economic research is produced. The scientific method performs a miracle. It generates *credibility*, which is so often elusive in this world, so difficult to establish or maintain, from a far simpler substance: *competence*. When the scientific method is employed, the validity of a scientific conclusion depends only on the competence of the researcher in designing and implementing an experiment in accordance with the scientific method. Absent fraud, this competence can often be assessed through peer review, and the scientific community is born.

But without the two lynchpins of the scientific method, credibility cannot be reduced to competence. It just isn't so simple. Competence in what? Economists have proved—empirically if not mathematically—that a given phenomenon can be ascribed to an almost infinite variety of mechanisms. Despite your carefully derived econometric model, you cannot know all the "unknown unknowns" that permeate virtually all economic phenomena. Are the only interesting research questions those in which the data truly allow you to rule everything else out? The credibility of economic research does not and can never devolve merely to competence in the development and implementation of theory and econometric models. Competence becomes a necessary, but not sufficient, condition for credibility.

Some economists would object to this characterization of the profession. They would acknowledge the ill-fit of our subject with the scientific method and the primacy of theory and econometric models, but would raise two counterclaims. The first is that theory, econometric models, and the "identification strategies" employed in empirical work have gotten much better over time. The second is that there is already a solution on hand to address the problem: demonstrating robustness. After a primary result is declared, supplementary results are presented to support it, estimates using alternative estimators, instruments, or controls, along with falsification tests intended to ferret out potential biases.

On a factual basis, these critics are correct. Methods have improved, a little here, a lot there. Robustness checks are now reasonably common, though far from universal. There is nothing wrong with improved methods or robustness checks, and much that is good. And yet…

The first claim doubles down on the characterization above. The methods that are being improved are economic theory and econometric modeling: the pillars of the profession. Few people have claimed that the problem is with these tools per se, but rather with their intrinsic limitations in the absence of the scientific method. And to the extent that the profession emphasizes analyses that conform to these new methods and identification strategies, we inhibit the study of equally important phenomena that do not so conform.

The second claim, on the other hand, acknowledges the limits of econometric modeling and addresses them to a reasonable extent. Robustness checks are a practical, ad hoc appendage to this pillar that doesn't so much assure its faultlessness as rule out some potential infirmities, such as biases that can arise from certain types of omitted variables. If this were all we could do to address the problem, we'd be done.

But it's not. It's not even the only way to look at the problem. What would it be like, instead, if we re-examined everything about economic research in the light of its contribution to credibility? If we stepped back to square one, put everything that we do or could be doing on the table, and forced it to prove its worth? If we approached the creation of credibility systematically, instead of ad hoc, mulled over each piece of the process to see what it can and should contribute? Where would that lead us? The answer is: a long way from where we are now.

1.2 Taking Inventory

So let us inventory everything that could possibly influence the credibility of an empirical study. If credible conclusions can be considered the output of our study, these are all potential inputs. They certainly include the economic tools on which we tend to focus, the pillars of the profession, but also much more.

The first such input to consider is personal credibility, which is embodied in the individual conducting the study. Personal credibility means that people believe something is so because you said it was so. In the world writ large, this very important source of credibility is usually produced gradually, though the accuracy of previous judgements that are retrospectively confirmed. It is usually ascertained just as gradually, through repeated personal experience with the individual making those judgements. Fortunately, in academia there are short cuts, such as the individual's pedigree, citation count, personal magnetism, and connectedness within the field. Just kidding! You'd be silly to pay attention to any of this stuff. Our profession is supposed to be scientific, remember? Attaching the credibility of a conclusion to the qualities of its author instead of the qualities of the study is fashion, not science.

A much better starting point is the real world situation you intend to analyze. The facts of this situation—institutional specifics, social and technical details, etc.—should influence both the scope and content of your analysis. To properly reveal its subject, a microscope needs to focus in the right amount—not too little, not too much. A similar principle applies here. Your research should be well-grounded in these social, technical, and institutional details, without getting overwhelmed or bogged down by them. It should prune these details and shape them into something manageable and useful.

Next there is your data. It only partially illuminates the phenomenon you are analyzing, yet it's all you've got. It whispers, "Believe everything I say!"—yet how can you know if you should? Your research should respect the infirmities of its data, while still using the descriptive facts embodied within to inform the analysis.

After that comes the analytical model that forms the centerpiece of your study. Sure, its backbone is composed of theory and econometrics. But if competence in these pillars can't guarantee us credibility, how are they best employed in the service of this goal? And what role is to be played by the study preliminaries I've just mentioned? We can't just throw everything in a pot, like gumbo, and hope it comes out all right. Models should be developed according to a clear set of guiding principles that answer these questions and more.

Then there is what you do with the estimates you obtain. This includes designing hypothesis tests, which need not be limited to the simple, standard null. It includes examining the implications of your findings, which are as important as the findings themselves. It also includes relating these findings to those of others, which may or may not accord with your own. Once in hand, your estimates should be thoroughly worked over, not coddled.

Finally, there is the narrative by which you articulate your findings and justify your conclusions. Do they lie there, devoid of context, alone like a fencepost under a desert moon? Do they soar in the ethereal heights of abstraction, unable to land on the terra firma of a grim and messy reality? Are they immobile, frozen stiff, or are they kinesthetic like the beings that presumably populate your economic model? This stuff isn't window dressing. It very reasonably affects the credibility of your conclusions, as well as their application to policy—your ability to shape the real world. Thus we end where we began.

That is not all. While each of these components is important, so is their assembly into the final product: how they all fit together to form and support your ultimate conclusions. If one focuses only on the two pillars of the profession, this concern may seem superfluous. Theory makes a prediction and you specify an econometric model to test it—this is obvious. But with all the inputs we have just mentioned, there are now many more parts that must be joined together, and many more options for doing so.

These connections include the transition between data description and econometric analysis, which can be simple and abrupt or more carefully hued. And the link between theory and model, which sometimes are fastened at the hip, as in a structural model, and sometimes are barely acquainted, as in many reduced form studies. They include the way that theory and econometrics are used to enhance hypothesis testing and assure its integrity. And how the facts of the situation relate to your analytics, and your final conclusions relate back to those facts.

Thinking about these connections is not at all superfluous. Remember Dava Sobel's *Longitude* (1995), the story of how accurate timekeeping solved the problem of navigation at sea? The pieces in John Harrison's clocks were of the highest quality, and they fit together perfectly. To tell the time accurately, you could not afford to have anything less. Here the same principle holds true.

1.3 Economic Research and Craftsmanship

Except for personal credibility, we would be unwise to take any of these inputs off the table. Given the profession's credibility problems, we should use every resource at our disposal. But then how shall we think about the process of doing research in this way? If credibility is formed from the deft assembly of many diverse research elements into a cohesive conclusion, is what we are doing still science?

To a reasonable degree, it is. Plenty of "hard scientists," including many geologists, astrophysicists, and evolutionary biologists, also cannot use the scientific

method per se. Like them, we have scientific structure and standardization. In fact, we will have more of it, because the chapters that follow add structure to many research decisions economists now make reflexively or intuitively. But there is also something more. The multifaceted, interlaced research process envisioned here is clearly distinct from the strict application of the scientific method. It is closer to the process of "making or manufacturing an object with careful attention to detail," as one dictionary puts it, with "skill in clever or artistic work" and "skilled workmanship," as put by another. This isn't mere competence. It is craftsmanship.

When you think about it, isn't it odd how rarely economics is called a craft? This term fits what we do quite naturally. Perhaps it is because focusing on craftsmanship puts us closer to plumbers and the makers of furniture than to Ernest Rutherford and Louis Pasteur. I hope you are OK with that. This characterization is foisted upon us not because of our own stupidity, but because of the exigencies of the occasion—the existence of many potential theoretical and empirical causes for an observed relationship, some of which we cannot recognize or account for. If we approach our craft in a stilted and overly rigid manner in the guise of being "scientific," we are misleading ourselves, but we will not mislead most other people.

The thesis of this book is that craftsmanship is essential to the credibility of applied microeconomic research. Just as intimate knowledge of music theory alone does not produce a wonderful symphony, as an encyclopedic vocabulary and grammatical rectitude do not a novelist make, neither does technical proficiency guarantee that your conclusions have a high probability of being correct. High levels of craftsmanship—the tempered blending of theory, data, institutional knowledge, and statistical analysis—are also required. In large part, the profession's credibility problems stem from insufficient attention to craftsmanship.

1.4 Plan of the Book

Our examination of craftsmanship proceeds in four parts. Part I of this book introduces a basic framework for thinking about the phenomenon that you intend to study. To this end, it reappropriates two words that economists have hijacked for other uses: scale and system. These concepts can shape your analysis from top to bottom.

Part II examines how information about this phenomenon is integrated into your study. General, qualitative information about it should matter, even when the study's technical aspects aren't directly affected—but how? Quantitative information—that is, data—surely matters, yet how should we think about its role? Clarifying our thinking on these matters can enhance our use of this information and prevent its misuse.

Theory and econometrics take center stage in Part III, which concerns the development of your analytical model. I do not attempt to teach either of these pillars, but rather assume a working knowledge of each. If you aren't competent in these technical aspects, craftsmanship won't redeem you. But if you are, it can inform a variety of practical questions: how to best put these pillars to use, how to recognize and address their limitations, and how to connect them to each other, to the scientific

method, and to facts about the phenomenon of interest. These are the fundamentals of good research design and execution.

Part IV focuses on the construction of your final conclusion, its integration into the literature, and its presentation to the reader. It envisions an ultimate goal to which your research should aspire, in which your findings achieve not only credibility but also improved utility among policymakers and the public. These are the fundamentals of an effective narrative with which you convey the results of your study.

Within each chapter, I develop key aspects of craftsmanship, explain their relevance, and give them some structure. This helps make craftsmanship more systematic and less haphazard and inscrutable. "I know it when I see it" may be a reasonable way to identify pornography, but for economic research we can aspire to more. Then examples are provided, both good and bad, from the literature. Some ideas are new; some are old but explained differently; some are downright hoary. Together, they try to construct a vision of craftsmanship as clear and cohesive as your research should aspire to be.

After finishing this book, I hope that you will look at applied microeconomic research differently. Some things that are considered routine you will consider strange; some that are overlooked you will consider essential; some things that impress others will not impress you, and vice versa. You will have less reverence for clever things that don't matter much in the end, and more interest in mundane details that matter a lot. Ultimately, I hope you will achieve an understanding of your research topic and of the research process that is, perhaps, less "traditionally scientific," but more accurate and more satisfying. So let's get started. All we have to lose are the shackles that we have strapped ourselves into.

References

Baron JN, Hannan MT (1994) The impact of economics on contemporary sociology. J Econ Lit 32(3):1111–1146

Fourcade M, Ollion E, Algan Y (2015) The superiority of economists. J Econ Perspect 29(1):89–114

McCloskey D (1983) The rhetoric of economics. J Econ Lit 21:481–517

Piketty T (2013) Le Capital au XXIe Siècle. Le Seuil, Paris

Reder MW (1999) Economics: the culture of a controversial science. University of Chicago Press, Chicago, IL

Rosenberg A (1992) Economics—mathematical politics or science of diminishing returns? University of Chicago Press, Chicago, IL

Sobel D (1995) Longitude: the true story of a lone genius who solved the greatest scientific problem of his time. Walker Books, London

Part I
Ways of Thinking

Chapter 2
Systems

Abstract This chapter provides a general definition of a system in economics, and shows how to go about defining such a system in practice. It argues that the ideal scope of an economic analysis is the system, and explains the types of problems that can occur by ignoring the system and focusing narrowly on the relationship of interest. These problems are illustrated in applications to labor supply in Europe, worker motivation in the American steel and aircraft industries, and children's rights in Nigeria.

A vast, under-explored world lies beneath the more traditional realm of theory, data, and estimation. Some economists hardly recognize its existence. But ignore this world and you risk building your study on a weak foundation. So our discussion of craftsmanship must begin here, with the development of two concepts central to this underworld: system and scale. You have heard these terms before. The profession has appropriated them for its own narrow uses. We're taking them back.

The concept of a system is used in many disciplines: there are ecosystems, electrical systems, political systems, and star systems. Each denotes a set of elements that are closely linked together. The plants in an ecosystem nourish its herbivores, who provide food for the carnivores, who in turn limit the number of herbivores, allowing the plants to flourish. The stars in a star system orbit each other, so that it is natural to think of this collection of stars as a unit rather than as a set of individual entities.

The most common uses of this term in economics accord with this meaning without doing it full justice. "System" often refers to a set of equations that are linked in estimation, such as those describing the supply and demand relations in a competitive market. It also refers to the philosophical underpinnings, for lack of a better phrase, of a nation's economy—for example, the capitalist system.

This limited usage is ironic, since economies are themselves systems. In the macroeconomy, everything is linked. Policies that lower inflation will affect real and nominal wages, profits, investment, output, exchange rates, and so on. Macroeconomists clearly recognize this as a central fact of their field, and their estimation methods are shaped accordingly. Vector autoregressions (VARs), for example, are systems of equations that mimic the macroeconomic systems they are

D. Grant, *Methods of Economic Research*, Springer Texts in Business and Economics, https://doi.org/10.1007/978-3-030-01734-7_2

modeling. On a regional basis, computable general equilibrium analysis serves the same purpose, specifying systems of equations for the production and consumption of many intermediate and final goods, in which all markets clear. Then the effect of a tax on some good, a localized demand shock, etc., can be traced throughout the regional economy.

These two systems are "traditionally" economic, in the sense that both the thinking and the mathematics wholly involve the prices and quantities of inputs and outputs. But most microeconomic systems are far more general than that. They include traditional economic variables, other quantities that are not traditionally economic, and institutional factors that can't be quantified at all. That's just fine, because systems aren't a mathematical tool, but a conceptual tool.

2.1 Thinking in Terms of Systems

To better understand the role of systems in your own research, let's contrast two portrayals of the effect of industrialization on the working man, one in the Old World, one in the New. The first, Upton Sinclair's *The Jungle* (1906), follows a hapless Lithuanian immigrant, Jurgis, through one misadventure after another while living and working in Chicago's Packingtown. The most devastating part of the book involves the sanitary conditions, or lack thereof, in the meatpacking industry:

> There was never the least attention paid to what was cut up for sausage; there would come all the way back from Europe old sausage that had been rejected, and that was moldy and white—it would be doused with borax and glycerin, dumped into the hoppers, and made over again for home consumption. There would be meat that had tumbled out on the floor, in the dirt and sawdust, where the workers had tramped and spit uncounted billions of consumption germs. There would be meat stored in great piles in rooms, and thousands of rats would race about on it. These rats were nuisances, and the packers would put poisoned bread out for them; they would die, and then rats, bread, and meat would go into the hoppers together.

The Old World equivalent, Emile Zola's *Germinal* (1885), focuses instead on coal miners in northern France:

> The four pikemen had spread themselves one above the other over the whole face of the cutting. This seam was so thin that they seemed to be flattened between the roof and the wall, dragging themselves along by their knees and elbows, unable to turn without crushing their shoulders. Maheu suffered most. At the top the temperature rose to thirty-five degrees Celsius, and the air was stagnant. His torment was aggravated by the moisture. The rock just above him streamed with water, which fell in large continuous rapid drops with a sort of obstinate rhythm, always on his face. In a quarter of an hour he was soaked, smoking as with the hot steam of a laundry.

And the worst thing is that he's having sausage for lunch.

Both books share the same purpose, social activism, but there is a key difference. While *The Jungle* is more or less confined to Jurgis and his family, *Germinal* expands to take in the wider world that surrounds Maheu. Each element of this world gets its due: not just the workers and their families but the shopkeepers, managing credit as much as inventory; the capitalists who own the mine, provide its

enormous physical capital, and accept its enormous financial risks; the managers, balancing speed with safety, the owners' complaints with those of the workers; the politicians who influence events remotely, in more ways than one. Each sees a piece of the whole, and we as readers see it with them. What these actors don't see—but we do—is how it all fits together: the system, which pins each actor, like Maheu, between a rock and a hard place.

It is very natural for us, like Sinclair, to focus on our object of interest and orient our study accordingly. Our conclusions can be simpler, more definitive, less nuanced. For social activism, this might be good. For social science, it's not. We need systems thinking.

2.2 Identifying Systems

To delineate a system, start with the phenomenon of interest, then work your way outward to everything that is integrally related to the dependent or key independent variables, whether or not it is strictly economic, and whether or not it can be measured. Many of these factors will themselves interrelate. Circumscribe the collection of interlocking relations that govern the phenomenon of interest, and then you have your system.

Thus, in human resource management, a compensation system includes not just base pay, benefits, and bonuses, but also the method by which workers are evaluated, the scope of decision-making authority that they have, and so on (see Ouchi 1977, or Milgrom and Roberts 1992). For a common currency, such as the Euro or the CFA, the system incorporates monetary and fiscal policies, political cohesion, and labor market mobility—all of which help maintain stability in the face of local shocks (as is well known). For price setting, the system would include costs and demand, the strategic reactions of competitors, the amount of advertising, and potential entrants (as suggested by Porter's five forces and micro theory).

In thinking about systems, be practical, not philosophical. Don't let the concept run amok, as it does at times elsewhere in social science, so that the flowchart of possible interactions and causal mechanisms resembles a Rube Goldberg machine. The temptation to do this isn't hard to understand: in many human endeavors, everything is related to something else to at least some small degree. So the list of relevant factors expands like a supernova, gobbling up everything in its path. Any predictions that your model might make are reduced to mush. It is, as they say, not even wrong.

But there is plenty of space between the two extremes. You need not list everything related to the phenomenon of interest, however tangentially. Just identify those things that have a substantive impact. A sense of magnitude—of scale, discussed in the next chapter—distinguishes the one from the other.

As a guide, think about people who really understand a cultural, economic, or political phenomenon—who live it, perhaps, and know it in the small and in the large. How do they talk about it? Typically, they concentrate on the major actors, beliefs, motivations, and environmental factors that influence the phenomenon.

Their understanding of the phenomenon is embodied within this system. The system doesn't radiate infinitely outward. It's not a list or mere collection of factors. Instead, it contains a discrete number of highly relevant, interlocking elements. There is no ambiguity in the meaning of digestive system or immune system, even for non-biologists. Similarly, we all understand that a loan officer is part of the financial system, while a seller of piggy banks is not. Approach your definition of systems the same way.

2.3 Uses of Systems Thinking

Systems thinking helps in three ways. First, it helps define the scope of your study. The system identifies all relevant factors that, ideally, your analysis should consider. Second, it helps identify potential estimation problems. Third, it facilitates the application of your analysis to policy, which is almost always developed and implemented within a system of diverse actors with diverse motivations.

This first reason, the scope of the study, is why this chapter comes right after the introduction. The idea of delimiting a study's scope is rarely formally discussed in our field, but we should do so thoughtfully. Rarely can we run controlled experiments and implement our policies exactly as intended, so that everyone behaves just as they are supposed to. Because of this, the first step in your analysis should be to carefully construct and populate the imaginary world that stands in for real thing. That is the system.

The next reason, estimation problems, comes in many flavors. Simultaneity occurs when two variables in the system influence each other. Omitted variables bias occurs when an important variable in the system, related to the dependent and independent variables, is not accounted for. Specification problems arise if the model enters each independent variable in isolation when they operate most effectively in concert, as systems are wont to do.

Since we already know about these types of estimation problems, and have methods to address them, what's the big deal? But that is thinking about it the wrong way. Systems aren't a mathematical tool, but a conceptual tool. They help you identify problems that, if the data exists, you already know how to solve. Systems thinking helps you understand the process that generates the values of everything: the dependent variable, the independent variables, and the error term. Then you can strive for harmony between this process and the process embodied in your theoretical and econometric models.

The final reason for systems thinking concerns the application of your results to policy. Often, the attitudes of various stakeholders are vital to making policy changes happen or making them work. Thus, if systems help you understand the distributional effects of a policy, you have identified a potential impediment to its adoption. If they help you recognize how various individuals will respond to the policy, you may identify unanticipated consequences, ways to implement the policy more effectively, or a second-best solution when the first-best is ruled out. These concepts all have long currency in our field.

Thus, in the end, systems thinking is not at all at odds with the economic way of thinking. Instead, it complements it, forcing us to consider some tempting blind spots and providing a guide with which to do so. When the phenomenon of interest is highly structured, amenable to accurate mathematical expression, and easily measured, there may be little need to think about systems. Otherwise, systems help you look outside of the box that our training sometimes puts us in. They help you understand the scope of your analysis, the major actors that populate it, and how your research question connects to the big picture. These are all essential elements of good craftsmanship. Using them should enhance our credibility, just as ignoring them should diminish it, as we will now see.

2.3.1 Labor Supply in Europe

Prescott's (2004) missive "Why Do Americans Work So Much More Than Europeans?" wants to "determine the importance of tax rates in accounting for… differences in labor supply for the major advanced industrial countries." So he builds a simple intertemporal optimization model in which labor supply responds to the marginal tax rate and the utility of leisure. Calculating the former, calibrating the latter, and incorporating data on capital's cost share yields predictions for the average person's hours of work that are then compared to the actual values across the G-7 countries. Changes in marginal tax rates, the only factor that varies across the simulations, can explain almost all of the variation in hours worked.

Look out—it's a trap! While Prescott might be interested in one factor in particular, that doesn't relieve him of the responsibility of considering others as well, any more than you can run a univariate regression because you really don't care about the controls. Abdicating this responsibility makes the analysis neat and self-contained, but reduces its credibility and sets the literature to debating the accuracy of its conclusions, rather than focusing on the question that spawned it to begin with.

This is the antithesis of the systems way of thinking. Not only has the analysis focused on a single potential explanation for the phenomenon of interest, it also hasn't attempted to articulate the system of social choices that underlies decisions about marginal tax rates and work effort. Reading the paper, you could be excused for wondering whether Prescott has ever even been to Europe, which clearly differs from America in far more than marginal tax rates.

The progenitor of this system would be the tradeoff between market, home, and community activities that every society must contend with. If there are unpriced positive externalities to time spent raising healthier children or building stronger communities, then societies may find it worthwhile to enhance the value of nonmarket activities, decrease the value of market activities, or both. This probably supports Europeans' persistence in having high tax rates, though they know it affects the amount of work. If so, pointing out the potential efficiency implications of these tax rates, as Prescott does at length, misses the point.

What would this system look like? We don't have to guess. The classic reference, untouched by Prescott, is Gosta Esping-Andersen's *Three Worlds of Welfare Capitalism* (1990). It identifies three types of modern welfare states, so classified by their labor legislation, support for the family, and "de-commodification," which decouples individual welfare from market success:

> Contemporary advanced nations cluster not only in terms of how their traditional social-welfare policies are constructed, but also in terms of how these influence employment and general social structure. To talk of 'a regime' is to denote the fact that in the relation between state and economy a complex of legal and organizational features are systematically interwoven (p. 2).

There's your system. Burgoon and Baxandall (2004) show that each type of welfare state has distinct implications for work hours, which can be fully understood only by examining two dependent variables, not one: hours per person and hours per employed person.

Translating Esping-Andersen's sociological jargon into useful terms is easier than it seems. De-commodification and family support involve redistributive social spending, which increases tax rates and the value of leisure together, making the two complements of a sort. If so, some of the differences in labor supply that Prescott attributes to tax rates stem instead from differences in the value of leisure, which he leaves static. Labor legislation pertains to unionization, which also must be accounted for if it is "systematically interwoven" with everything else. The systems-oriented analysis of Huberman and Minns (2007) places the hours-of-work decision in the context of culture, history, and labor relations, and finds that unionization plays a major role in explaining why Americans work more than Europeans do. Alesina et al. (2006) finds that it plays *the* major role.

2.3.2 Worker Motivation in the Steel Industry

A second paper simultaneously illustrates the power of thinking about systems and the problems that result from ignoring them.

In the 1990s it was all the rage to talk of redefining the employer-employee relationship via enlightened human resource management (HRM) practices.

> During the last decade, the media have proclaimed the death of the old employment contract and a new emphasis on flexibility and external, not internal, labor markets (Levine 2002).

It was the era of Saturn—a "new kind of car company."

Accordingly, Ichniowski et al. (1997) examine the effects of various HRM practices on the productivity of a homogenous type of steel finishing line. The mechanized process feeds in and spits out coils of steel, so productivity equates to uptime. The authors hook up with the steel industry—one is from Pittsburgh, that helps—and go to visit each line they study, talking with the engineers there and documenting 26 different HRM practices that are used.

The authors group various combinations of these practices into a total of four different systems. The most traditional of these, System 4, has strict work rules, close supervision, and little training, as in *Germinal* or Devo's poignant ballad "Working in a Coal Mine." The others add teamwork, information sharing, training, job flexibility, job security, etc., in stages, so that the top system, System 1, has all "innovative" practices, as the authors put it. The claim is that these innovative practices work best when coupled together: teamwork improves when employees are more carefully screened, incentives are more effectual when task assignment is more flexible, and so on. That is, they are sets of complements, or, for System 1, one large "supermodular" set of complements.

Sure enough, the authors find that uptime increases substantially as you move from System 4 to System 1, and once you account for the system, individual practices matter little. The study thus provides empirical support for thinking about HRM policies in terms of systems, and in fact this is common to do. If you refer to an HRM system, people will know what you mean.

Yet I never refer to this article when I teach this material to labor economics students. The system that it renders is balanced by the system that it omits. HRM policies are, so to speak, equilibrium outcomes, chosen unilaterally by firms, mindful of their effect on workers, or determined through collective bargaining (which is common in the steel industry). Thus, the mere fact that these practices tend to occur together—a fact poorly documented in the paper—supports their complementarity (in what is now a reasonably competitive industry). But it also raises the question of why some firms adopt no innovative practices, others some, and still others all. Without answering this question, our understanding of the role and effect of these systems is woefully incomplete. Imagine pitching factory managers on the value of these practices for their firms. Wouldn't they would want to know the answer to the question I have just posed? Then how can we be satisfied in its absence?

The authors provide some information about these practices' adoption, noting their ubiquity in "greenfield" mills that have been built from scratch and their scarcity on older lines where labor trusts management less. A more compelling explanation would embed these decisions into the firm's operating environment, which could include competition, product demand, input supply, and the production technology. Describing how these practices fit into this broader system would add insight and enhance the credibility of the authors' conclusions.

This deficiency is remedied in the *Wall Street Journal* article that I teach from, White (1996), which comes from the same era, and focuses on the same transition in HRM practices. It describes how an aircraft engine plant that had operated on the Devo model was then transformed by the use of incentive pay, teams, job flexibility, and increased training:

> For Mr. Ponchak, fifty-four years old, survival meant learning to take risks. Against long odds, amid dispiriting waves of layoffs, his seventeen-year-old plant slashed the time it takes to set up metal-grinding machines. It broadened job descriptions, so today eighteen inspectors do more work than twenty-eight did five years ago. And it found novel ways to pay people, turning them into an army of cost-cutters.

Productivity increased substantially, costs fell accordingly, and the plant survived and even grew.

Unlike Ichniowski et al., this article also shows us why this happened: the product market changed. Previously the Cold War was hot and competition was limited, so the firm's primary concern wasn't minimizing costs, but ensuring a large supply of workers equipped with the necessary firm-specific human capital. Then—

> Mr. Ponchak was named manager of the Maine plant in 1992, just as "the end of the world happened." Actually it was the end of the Cold War, combined with a cyclical dive in the commercial-jet engine business. Pratt engines make up 55% of those flying in Western fleets, but global price competition is increasingly cutthroat.

Reduced demand and increased competition forced the firm to become more cost-competitive, motivating the changes in HRM policy that saved the factory. Thus, while this article is less formal and less general than its counterpart, it is also more insightful.

Note also the change in perspective adopted by thinking in terms of this larger system. The non-traditional HRM practices extolled by Ichniowski et al. are no longer "enlightened" or "innovative"—merely appropriate in some situations and not in others. They increase productivity, at least in the short term, but carry attendant costs in other dimensions, such as employee turnover. These costs are real. Do you remember what happened to Saturn? In 2010, it became a "defunct kind of car company." Labor troubles were part of the reason.

2.3.3 Children's Rights in Nigeria

Systems often imply the presence of endogeneity, the econometric consequence of multiple directions of cause and effect. Thus, since we don't know what gives rise to each finishing line's HRM policies, we can't be certain that their use is independent of the line's productivity.

The archetypal conundrum of this sort arises when trying to determine laws' effects on social outcomes. The system here includes the social forces giving rise to the law, the law itself, and its acceptance and enforcement, which are all interconnected (Axelrod 1986):

> Social norms often precede laws but are then supported, maintained, and extended by laws. The two are often mutually supporting, because norms can become formalized into laws and because laws provide external validation of norms.

We know this. "The law is endogenous," we say. But what this hackneyed phrase really says is that we lack a satisfactory way to describe the problem. Is it that behavioral changes make the law's existence more likely? Or that public sentiment affects both the outcome and the probability the law will be passed? Or that social acceptance of the law influences its effectiveness or enforcement? Or some combination of the above? This phrase uses narrow, imprecise econometric terminology

to describe a more general problem: the inability of an ordinary regression to adequately represent the system that generates these laws and makes them work.

The resulting ambiguity makes it easier to blast ahead anyway with some variant of the following basic panel model:

$$Y_{s,t} = \alpha + \beta X_{s,t} + \gamma L_{s,t} + \varepsilon_{s,t} \tag{2.1}$$

where Y is the outcome of interest, L is a dummy for the law in question, and X is a set of control variables. (The book follows convention throughout by letting Greek letters represent parameters and ε the error term, and by indexing states or other cross-sectional units with s and time with t.) We check for problems with typical techniques—robustness tests, state trends, etc.—and then act as if our estimate of γ applies everywhere by default. I have seen countless such studies in my time, which analyze legislation without a word about the processes that create it or affect its implementation.

How much better it would be to understand the nature of the system! Consider Nigeria's Child Rights Act, an expansive law prohibiting child labor, child marriage, and more. Nigeria uses an interesting type of federalism: though the federal government adopted the Act in 2003, it remained unenforceable in any state until also adopted by that state. Figure 2.1 maps the states that had adopted the Child Rights Act as of 2007. Just the sight of it makes you want to pull up Stata, doesn't it? There are plenty of adopters and non-adopters, reasonably dispersed. Time to estimate this law's effect, so the other states can see how much they have to gain by adopting it, too!

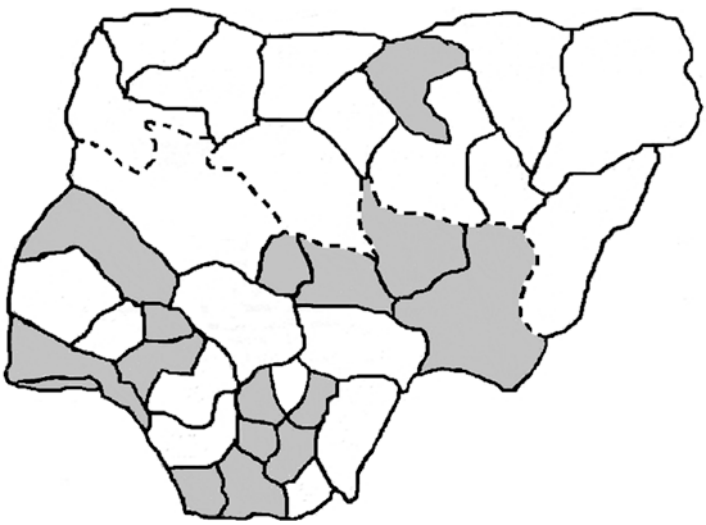


Fig. 2.1 A map of the states of Nigeria. The shaded states had adopted the Child Rights Act as of 2007

Just like that, we are thinking in terms of Eq. (2.1). In doing so, we are overlooking other major factors influencing children's roles in family and communal life: the system governing child outcomes and this law's influence on them.

An obvious factor is religion. Nigeria can be bifurcated into predominantly Muslim states, in the north, and predominantly Christian states, in the south. Their differences, regarding the role of women especially, far exceed those in the countries of *The Three Worlds of Welfare Capitalism*. In the north, the Act's minimum age of marriage—eighteen, higher than in most states of the U.S.[1]—conflicts with religious tradition (Braimah 2014):

> While the Act sets a child to be someone under eighteen, in Islam there is no age that marks childhood. A child's maturity is established by signs of puberty, such as menstruation.

Sure enough, all of the states below the dashed line in Fig. 2.1 have now adopted the Child Rights Act. Except for lonely Jigawa State, the states above the line have not.

But religion, it turns out, is not what the dashed line demarcates! Majority-Muslim states lie on both sides of this line. A cursory understanding of the system is not good enough. Rather, the top states alone operate on Shari'a Law. This suggests, and legal scholarship confirms (Akinwumi 2009; Braimah 2014), that the type of law is a key part of the system, more important even than religion.

Reading these studies, and appreciating the depth of the discord between the Act and Shari'a Law, the idea of using the northernmost states as "controls" for the others (as in a difference-in-difference framework) seems increasingly odd, as does the idea of treating γ as a timeless, one-size-fits-all prediction of the law's causal effect. This is confirmed by the experience of Jigawa State, the exception that proves the rule. Dig a little further, and you will discover that it has struggled to implement the Act in the face of local opposition (Akor 2011). To study the Child Rights Act without knowing the system that surrounds it would be a fool's errand.

Altogether, these applications not only demonstrate how systems thinking is used, but also show us where to start our study: at the beginning, with a good understanding of the phenomenon of interest. To understand it, we must get a good, careful look at it, and know what to look for. This is the task for which systems thinking is designed.

Food for Thought

Because they extend the themes developed in each chapter, these questions should always be read, even if they are not all worked.

Only sometimes will they admit a precise, deductive answer. Other times, finding the best answer will require contemplation and discussion with others. This is as it should be, since craftsmanship cannot be learned from a "cookbook," a step-by-step

[1] And far higher than in some. (Parental consent is generally required at these younger ages, but this is no different in Nigeria.) Unlike Nigeria, the U.S. has not ratified the U.N. Convention on the Rights of the Child, which first impelled this Act into motion.

manual of what to do. Discussion boards for these problems can be found on the book's website, www.worldofeconomics.com.

For some questions in this chapter, some facts have been oversimplified to make things more clear-cut.

1. Two large groups of Native Americans are the Apache, located primarily in Arizona, and the Sioux, located primarily in South Dakota.

   Centuries ago the Apache subsisted on agriculture, raising sheep, and food-gathering. While land was plentiful, it, and the crops it grew, were private property; nevertheless, chiefs taxed families and redistributed income (food) to the less fortunate, though income was not fully equalized. Chiefs were also responsible for deciding how many animals (sheep) should be eaten in any given year. Chiefs were vested with a great deal of decision making power, and generally served for life.

   The Sioux, on the other hand, subsisted on bison. They were plentiful but mobile, so the band constantly migrated in search of food. Hunting bison is difficult and dangerous, requiring the effort and coordination of several individuals; food obtained in the hunt is shared completely within the band. The Sioux were loosely organized into small bands whose chiefs had limited power and operated primarily through persuasion. Chiefs served at the pleasure of the band, and families could leave one band and join another if they wished.

   (a) The Sioux system has the following features: (1) small bands, (2) free movement from one band to another, and (3) limited power on the part of the chief. Describe how these features work together to help the bands operate most effectively.
   (b) The Apache system has the following features: (1) land ownership and free enterprise, (2) redistribution, and (3) an authoritarian chief with a life reign. Describe how these features work together to help the tribe operate most effectively.
   (c) Explain how the organization of each tribe is shaped by the environmental conditions under which it operates. To do this, articulate the primary HRM objective of each tribe, relate it to the tribe's environment, and then show how the tribe's organization best achieves that objective.

2. The Beveridge health system model, as implemented in the United Kingdom, has the following key features:

   (a) full government financing of doctor's visits and hospital care,
   (b) free choice of physician,
   (c) an organization, the ironically-named "NICE," that sets guidelines for when procedures are and are not "indicated" for treatment, and
   (d) a "global budget" for hospitals, that is, a fixed amount of money that has to last the year.

   Consider the objective of the health system to be to maximize political support via an appropriate combination of cost (via taxes) and quality. Discuss how these features work together to help the NHS achieve this objective.

3. Three factors relevant to the efficiency of market functioning are the degree of competition, the amount of information participants have relevant to proposed or actual transactions, and the degree to which incentives faced by market participants are appropriately aligned with social costs and benefits—in short, competition, information, and incentives.

   (a) Discuss the complementarity of competition and incentives with regard to the Coase theorem. Is there an efficiency difference if property rights are granted to a monopolist, as opposed to a set of competitors (as in cap and trade systems)?
   (b) Discuss the complementarity of information and competition with regard to Diamond's (1971) simple search model. In the absence of information about product price, how does an increase in the number of competitors affect the market price for a homogenous product?
   (c) Discuss the complementarity of information and incentives with respect to adverse selection.
   (d) Comment on the general complementarity of these three factors in promoting market efficiency. Can this fact help you define the system associated with the functioning of any given market?

4. The concept of resiliency, as it applies to systems, refers to the degree to which the system can combat the effects of an adverse shock or the speed and ease with which it recovers from it. Some of the factors that affect resiliency are considered below.

   (a) Networks are systems that are composed of similar parts and the connections between them. These systems' resiliency will depend on their redundancy: the multiplicity of ways in which one can connect any two points. With that in mind, which of the following networks is likely to be most resilient: an airline network, a wireless network, or a telegraph network?
   (b) Another factor affecting the resiliency of networks and other social systems is centrality: the degree to which all connections or decisions pass through or stem from a single focal point. (The bad guys in science fiction movies like *Star Wars* or *Independence Day* always seem to have highly centralized networks.) Do you think bureaucracies with greater centrality are more or less resilient? How so?
   (c) The diversity of ecosystems is known to enhance their resilience. Do you think the same is true of economic systems? Why or why not?

5. Prescott's simple analysis is able to account for the U.S./European difference in labor supply, despite omitting other contributing factors such as unionization. This suggests that "two wrongs made a right": that Prescott made a second error that corrected for this omission, allowing him to fully account for the difference in labor supply using tax rates alone.

(a) Speculate on what this error might be, then check your work by consulting Alesina et al. (2006).
(b) How would correcting the first error, and accounting for the other components of the system, help you uncover the second error?

References

Akinwumi D (2009) Legal impediments on the practical implementation of the Child Right Act 2003. Int J Leg Inf 37(3):10
Akor O (2011) Why Jigawa deferred Child Rights Act. Daily Trust. http://allafrica.com/stories/201112280470.html
Alesina A, Glaeser E, Sacerdote B (2006) Work and leisure in the U.S. and Europe: why so different? In: Gertler M, Rogoff K (eds) NBER macroeconomics annual, vol 20, pp 1–64
Axelrod R (1986) An evolutionary approach to norms. Am Polit Sci Rev 80(4):1095–1111
Braimah T (2014) Child marriage in Northern Nigeria: section 61 of part I of the 1999 constitution and the protection of children against child marriage. Afr Hum Rights Law J 14(2):474–488
Burgoon B, Baxandall P (2004) Three worlds of working time: the partisan and welfare politics of work hours in industrialized countries. Polit Soc 32(4):439–473
Diamond P (1971) A model of price adjustment. J Econ Theory 3:156–168
Esping-Andersen G (1990) The three worlds of welfare capitalism. Wiley, Hoboken, NJ
Huberman M, Minns C (2007) The times they are not changin': days and hours of work in old and new worlds, 1870-2000. Explor Econ Hist 44:538–567
Levine DI (2002) The new employment contract? W.E. Upjohn Institute for Employment Research, Kalamazoo, MI
Ichniowski C, Shaw K, Prennushi G (1997) The effects of human resource management practices on productivity: a study of steel finishing lines. Am Econ Rev 87(3):291–313
Milgrom P, Roberts J (1992) Economics, organization, and management. Prentice Hall, Englewood Cliffs, NJ
Ouchi W (1977) The relationship between organizational structure and organizational control. Adm Sci Q 22(1):95–113
Prescott EC (2004) Why do Americans work so much more than Europeans? (No. w10316). National Bureau of Economic Research
Sinclair U (1906) The jungle. Doubleday, New York
White J (1996) Dodging doom: how a creaky factory got off the hit list, won respect at last. Wall Street J 26:A1–A2
Zola E (1885) Germinal. G. Charpentier, Paris

Chapter 3
Scale

Abstract This chapter introduces the concept of scale as it is used in other disciplines, as an indicator of magnitude. It shows how economists can utilize this concept to add clarity, simplicity, and insight to their research. Applications to the incidental parameters problem, the Slutsky Equation, Mincer's wage equation, mortality dynamics, and more elucidate the power of scale analysis, both theoretically and empirically.

In economics, the concept of scale traditionally refers to the size of the plant and/or firm—it's often muddy which is which. When larger plants or firms have lower average costs, there are economies of scale; there are diseconomies of scale when it works the other way—though it's unclear whether such diseconomies occur in practice, so far as I can tell, or why they should.

In this sense of the term, scale refers to *size*. It would be just as good, though less impressive-sounding, to refer to economies of size. Better, actually, because in general use and scientific research the term scale refers to a concept that, though related, is different: *magnitude*. On what time scales does a persistent increase in oil prices generate new sources of supply: months, years, or decades? What is the scale of the damage inflicted on Hungary's corn crop by the Western Corn Rootworm: tens of thousands, hundreds of thousands, or millions of hectares? The concept focuses on orders of magnitude, which can be reasonably approximated, rather than exact values, which cannot.

In using the term the way that we do, economists distract ourselves from this more important meaning of the word. This is unfortunate, because the concept of scale as magnitude is valuable at every stage of an empirical study: the beginning, in which you denote the system and lay out your theoretical predictions; the middle, in which you describe and analyze the data; and the end, when you interpret your results and their implications. So let's acquaint ourselves with the concept and its uses.

D. Grant, *Methods of Economic Research*, Springer Texts in Business and Economics, https://doi.org/10.1007/978-3-030-01734-7_3

3.1 The Nature of Scale

Scale can be quantitative or qualitative. Here Joseph Pedlosky's *Geophysical Fluid Dynamics* (1979) sharpens its focus using a mathematical scalpel of scale:

> For the purpose of this text large-scale motions are those which are significantly influenced by the earth's rotation. Let L be a characteristic length scale of the motion, i.e., one that characterizes the horizontal spatial variations of the dynamical fields. This could be the distance between a pressure peak and a succeeding trough. Similarly let U be a horizontal velocity scale characteristic of the motion. [In a previous example] L would be on the order of 1,000 km, while U would be on the order of 20 meters/second.
>
> The time it takes a fluid element moving with speed U to traverse the distance L is L/U. If that period of time is much less than the period of rotation of the earth, Ω, the fluid can scarcely sense the earth's rotation over the time scale of the motion. Rotation, then, is unimportant whenever $L/U << \Omega$. For the L and U given above, we can expect the earth's rotation to be important.

This rotational effect works in opposite directions in the two hemispheres, and a common belief has toilets in the U.S. and Australia flushing in opposite ways for this reason. This analysis debunks that notion: there L is closer to 0.001 km—and even that's a mighty big toilet.

And here Michael Lewis' *Moneyball* (2003) leaves numbers behind altogether:

> Voros McCracken once saw someone say that no matter how much research was done, no one would be able to distinguish pitching from defense. That is, no one would ever come up with good fielding statistics or, therefore, good pitching statistics. If you don't know how to credit the fielder for what happens after a ball gets put into play, you also don't know how to debit the pitcher. You would never be able to say with real certainty how good any given pitcher was.
>
> When Voros read that, "I thought, 'That's a stupid attitude. Can't you do *something*?'" So he divided the stats a pitcher had that the defense behind him could affect (hits and earned runs) from the stats a pitcher did all by himself (walks, strikeouts, and home runs). He then ranked all the pitchers in the big leagues by this second category. For the 1999 season, his list was topped by these five: Randy Johnson, Kevin Brown, Pedro Martinez, Greg Maddux, and Mike Mussina. "I looked at that list," said Voros, "and said, 'Damn, that looks like the five best pitchers in baseball.'" He then asked the question: if looking at just walks, strikeouts, and homers identified the five best pitchers in baseball, how important could all the other stuff be?

What these excerpts both have, despite their differences, this next one doesn't (Cameron and Trivedi 2005, p. 781). Can you see what's missing?

> Neyman and Scott considered inference when some parameters are common to all observations but there are additionally an infinity of parameters, each of which depends on only a finite number of observations. The common parameters are of intrinsic interest, whereas the latter parameters are called incidental parameters.
>
> Consider the nonlinear model $E(y) = g(\alpha_s + \beta x_{st})$, where $s = 1 \ldots N$ indexes cross sectional units and $t = 1 \ldots T$ time. Here β is a common parameter, but $\alpha_1 \ldots \alpha_N$ are incidental parameters if the panel is short, as then each α_s depends on fixed T observations and there are infinitely many α_s since $N \to \infty$. The incidental parameters are inconsistently estimated as $N \to \infty$, since only T observations are used to estimate each parameter. In general the com-

> mon parameters are also inconsistently estimated, even though they are finite in number and are estimated using $NT \to \infty$ observations.

The difference comes out in relief: this has none of the sense of magnitude that pervades the first two excerpts. Pedlosky delineates when you need and need not account for the Coriolis effect in modeling air or ocean circulation, by comparing its time scale to that of the other forces influencing the circulation of the fluid. McCracken resolves an impasse about evaluating pitchers by showing that the statistics under their full control are of primary importance, while those that depend on the defense are secondary, conditional on the first set of statistics, and can be ignored.

In the last excerpt, however, Cameron and Trivedi tell us only that bias remains in infinitely large samples. We do not learn when this bias is likely to be large or small, or anything about its magnitude in finite samples. So how are we to know when this problem is serious, and when it can be safely disregarded? So far as I can tell, this has been a real issue in this literature, with some studies contorting themselves to avoid the incidental parameters problem and others ignoring it without apology. A mere example won't resolve the issue, unless it broadly represents the cases that occur in practice. You need a general characterization of when the incidental parameters problem fails to be relatively small. You need a scale analysis.

3.2 Scale Analysis in Economics

Economists do not ignore scale completely. We use it whenever we distinguish the long run from the short run, argue that a coefficient estimate is reasonable using "circumstantial evidence," or simplify a mathematical expression based on relative magnitudes: "when $z \gg x$ then the above expression reduces to the following…" Still, the concept is far from fully integrated into our research paradigm, and more papers prompt questions about magnitudes than answer them. All that highfalutin stuff in the opening chapter is true enough, and worth being said, but you don't need it to realize that our profession has not reached full maturity. You only need to note our inattention to the concept of scale.

Scale analysis involves approximating orders of magnitude, using elementary logic and essential information about the topic under study, in order to shape your research and interpret its conclusions. A GDP of $13,447,568,413 could be considered on the order of $10 billion. Or, let there be two variables, p and q, measured in the same units. If we say that p is an order of magnitude larger than q, this implies that p might be, say, five or ten or twenty times as large as q, but not merely twice as large. Perhaps the effect of p on some dependent variable y is large, while the effect of q on y is at least ten times smaller. We could then say that the effect of p is first order, while that of q is second order at best. In each case we are talking about magnitudes: we are imprecisely precise.

To illustrate how scale analysis can be used in economics, let's bring out that old warhorse, the Slutsky equation. This equation, which breaks down the demand response to price changes into income and substitution effects, can be written this way:

$$\epsilon_p\Big|_{u=\bar{u}} = \epsilon_p - w\epsilon_I \tag{3.1}$$

The Hicksian (constant utility) elasticity equals the Marshallian own-price elasticity minus the income elasticity multiplied by the good's budget share, w. The leftmost term in Eq. (3.1) denotes the substitution effect and the rightmost the income effect, with the measured elasticity in the middle.

Scale analysis compares the magnitudes of the two terms on the right. Usually, it isn't a fair fight. Consider the following:

$$\left|\epsilon_p\right| \gg w\left|\epsilon_I\right| \Leftrightarrow w \ll \frac{\left|\epsilon_p\right|}{\left|\epsilon_I\right|} \tag{3.2}$$

At the product demand level, a Marshallian price elasticity of -0.2 is quite small, while an income elasticity of ± 2 is quite large. Substituting these "extreme" values into Eq. (3.2) implies that the income effect is at least an order of magnitude smaller than the price elasticity whenever $w \ll 0.1$. (Substituting less "extreme" values only reinforces this conclusion.) Then product demand almost entirely reflects a substitution effect, and the difference between Hicksian and Marshallian demand is not meaningful.[1] Since $w \ll 0.1$ for almost everything, this statement applies quite broadly, the rare exceptions being goods with very large income elasticities or goods with large budget shares, such as housing or food (as a single good). That is scale analysis at work.

In my informal survey, this point can be found in graduate and undergraduate microeconomics textbooks not quite half of the time. You know what can always be found? Giffen Goods, though the conditions required for the Marshallian elasticity to be positive are even harder to satisfy. Meanwhile, numerical examples and homework questions routinely use implausible numbers that suggest sizeable income effects are widespread. It is as if we are being subconsciously instructed to ignore the concept of scale at the very inception of our economics training. Emphasizing the distinction between Hicksian and Marshallian elasticities and the oddity of Giffen Goods might seem smart, but if we lose the concept of scale in the process, we are the worse off for the bargain.

Scale analysis also applies to regressions whose dependent variable is an average of individual outcomes taken across a cross-sectional unit, such as a state or country, often at various points in time. (One such dependent variable would be the unemployment rate, which is calculated from individual survey data.) These regressions are typically specified as follows:

$$\bar{y}_{s,t} = \alpha + \beta X_{s,t} + \epsilon_{s,t} \tag{3.3}$$

[1] This conclusion is stated in relative terms, but a similar conclusion also holds in absolute terms, assuming that a difference in demand elasticities of 0.1 matters little.

with any fixed effects subsumed into *X*. We generally anticipate heteroskedasticity, because larger, more-sampled cross-sectional units have less sampling error in averaging *y*. Both the efficiency of the coefficient estimates and the accuracy of the standard errors are affected.

This issue is usually addressed by using weighted least squares (WLS), weighting by population or something similar. Because cross-sectional units often differ greatly in population, as do the countries in the world or the states or provinces in Brazil, India, Indonesia, etc., some cross sectional units receive vastly more weight than others.

We do this because we are sensitive to econometric problems, and here is one staring us in the face. So we act, without even considering the relative magnitudes of the underlying sources of error. We lose all sense of scale.

Unless *X* includes absolutely everything influencing *y*, this sampling error is coupled to specification error resulting from omitted variables or other unmeasured influences that are common to all agents within a cross-sectional unit. Expressing this specification error as a random effect, we can rewrite Eq. 3.3 as follows:

$$\bar{y}_{s,t} = \alpha + \beta X_{s,t} + \bar{v}_{s,t} + \xi_{s,t} \tag{3.4}$$

where $\bar{v}$ is the sampling error derived from averaging, which is heteroskedastic, and ξ is the state*year random effect, which is not. The overall degree of heteroskedasticity in this equation depends on the relative variances of these two error terms. It will be substantial only when the first dominates the second. Then population-weighted least squares will yield nearly-efficient estimates and accurate standard errors. Otherwise, it can easily make things *worse*.

Thus, if you must choose between this estimator and ordinary least squares (OLS), you should compare the relative variances of ξ and $\bar{v}$. When sample sizes are small and the set of controls is exhaustive, $\text{var}(\xi) < \text{var}(\bar{v})$ and WLS is better. In the reverse, with large sample sizes or limited controls, $\text{var}(\xi) > \text{var}(\bar{v})$ and OLS is preferred. These relative magnitudes are usually not hard to ballpark. Yet, despite our technical sophistication, this is still often not done.[2]

As we have seen so far, the ends to which scale analysis can be applied are quite varied. It cannot be reduced to a checklist. You have to use it when and how the circumstances permit, and remain vigilant for opportunities to do so.

Nonetheless, these cases all employ scale analysis for the same general purpose: to distinguish the highly relevant from the less relevant. Then you can keep what is

[2] Sampling variance is sometimes documented by the surveyor, or is easily estimated from the micro data. If the standard deviation of the N_j individual observations in state*year cell j is σ_j, then the sampling variance in that unit is $\sigma^2_j/(N_j\text{-}1)$; averaging these terms across all state*year cells yields the "average sampling variance." As sampling error is independent of the random effect, this value and var(ξ) sum to the variance of the residual. Solon et al. (2015) mention some recent studies that reached erroneous conclusions because of the inappropriate use of WLS and present more general ways to address this issue.

large and disregard what is small. Simplicity, simplicity, simplicity. The Coriolis Effect can be disregarded for the flushing of toilets, but not for ocean circulation. Heteroskedasticity arising from sampling error can be disregarded when this error is dominated by a random effect, but not otherwise. And so on.

3.3 Heuristics

Our use of scale so far has been reductionist: the mathematical equivalent of "don't sweat the small stuff." The theoretical analog of this principle involves stripping elaborate mathematical models of their frippery, and boiling things down to their core. This can yield more tractable functional forms and essential intuitions that transcend the mathematical specifics. Such intuitions can be expressed as heuristics: serviceable rules of thumb whose crudeness is outweighed by their insight and handiness. Scale can help develop these heuristics by identifying messy, higher order stuff that can be excised, or by showing when we can safely use simpler, limiting cases that are more easily characterized. Let's look at an example of each.

Among many species, humans included, mortality rates grow exponentially with age until leveling off in senescence. To explain this pattern, Gavrilov and Gavrilova (2001) develop a complicated expression for the age-mortality relationship from theory, and then simplify it in two broad strokes:

> The failure rate, μ, of the organism as a function of age, x, is given by:
>
> $$\mu = k\lambda mce^{-\lambda}\sum_{i=1}^{n}\frac{\lambda^{i-1}\left(1-e^{-kx}\right)^{i-1}}{(i-1)!\left(1-\left(1-e^{-kx}\right)^{i}\right)} \tag{3.5}$$
>
> where $1/k$ is the mean failure time of each component of the organism. In the early-life period, $x << 1/k$ and, therefore, $1 - e^{-kx} \approx kx$, so the mortality kinetics follow the exponential Gompertzian law: $\mu \approx R\exp(\alpha x)$, where R and α are functions of c, m, k, and λ. In the late-life period, $x >> 1/k$ and, therefore, $1 - e^{-kx} \approx 1$, so the failure rate levels off and the mortality plateau is observed: $\mu \approx mk$. The failure rate of the organism initially grows exponentially with age, then eventually decelerates and approaches an upper limit equal to mk. The model explains the exponential increase in mortality with age and the subsequent leveling off.

Here, scale analysis amounts to finding two asymptotes and their range of relevance: an exponential for the young, and a constant for the old. Actual mortality begins in the neighborhood of the first asymptote, and ends in the neighborhood of the second. Figure 3.1 illustrates exactly this pattern for French mortality, which holds surprisingly well after the accident-prone teenage years. The heuristics clarify the intuition and simplify the technicalities; matching their functional forms to the data builds confidence in the model. Biology, it turns out, is loaded with heuristics like these, both theoretical and empirical.

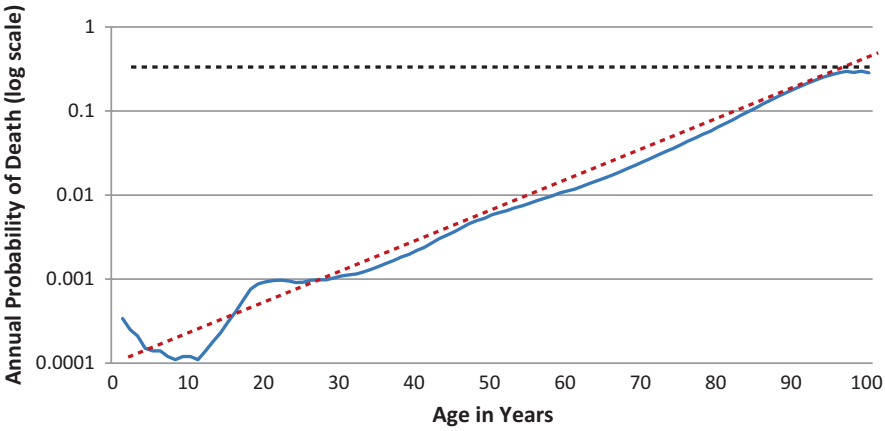


Fig. 3.1 Male mortality risk, metropolitan France, 2002–2004 (data from Insee Vital Statistics)

Another example, closer to home, arises in Mincer's (1974) development of the relationship between schooling and earnings. Scale shows up here before your coffee gets cold:

> Investments in people are time consuming.

Good opener! Deceptively simple. Worth coming back to later.

> Investments in people are time consuming. Each additional period of schooling postpones the time of the individual's receipt of earnings and reduces the span of his working life, if he retires at a fixed age. The deferral of earnings and the possible reduction of earning life are costly. These time costs plus direct money outlays make up the total cost of investment.
>
> An empirically convenient assumption is that all investment costs are time costs. This assumption is more realistic in such forms of human capital investments as on-the-job training, but less so in others, such as schooling. This assumption is not essential. Detailed information on direct costs can be incorporated into the model to yield a more precise empirical analysis. We forego precision, in order to gain in the simplicity of exposition and analysis.

This is an opportunity missed. Mincer justifies treating all investment costs as time costs based on convenience alone. A direct argument based on scale would be superior, for then we don't have to extol simplicity in abstract, just show that its cost is small.

This argument confirms that, in Mincer's time, simplicity sacrifices little. Mincer's (1975) data, from 1959, indicate that a fresh U.S. high school graduate earned about $3500 per year in contemporaneous dollars. The Digest of Education Statistics doesn't go back quite that far, but 5 years later, in 1964, average tuition at 4-year public institutions was $298. Direct costs are an order of magnitude smaller than time costs, and can be ignored without seriously impairing the argument.

OK, what's next?

> When earning life is long, it matters little whether each additional year of schooling reduces earning life by exactly one year, or whether more educated people retire at correspondingly later ages. What matters is the deferral of earnings: the cost of currently postponing earnings by one year is much more significant than the present cost of reducing earnings by one

> year, four or five decades hence. An infinite earning life can, of course, be viewed as a special case of the equal-span assumption. This formulation has greater tractability and flexibility in empirical interpretation.

Another opportunity missed. The cost of postponing earnings now is indeed far larger than the present value of reducing earnings decades later. To complete the scale analysis, which is there for the plucking, all you would need is this:

> With constant annual earnings and an annual discount rate of 5%, the present value of an infinite earnings stream is only 10% higher than that of an earnings stream forty-five years long, a typical working life in the U.S. This difference is small enough that the former can be assumed without much harm.

This is clearer and more defensible than the "special case" argument for an infinite earning life.

For these simplifications, you get something valuable in return—a simple heuristic that is not only easy to derive and to estimate, but which also admits a structural interpretation:

$$\ln E_{\mathrm{i}} = \alpha + rS_{\mathrm{i}} + \varepsilon_i \tag{3.6}$$

where E is earnings, S is years of schooling, and r is the discount rate or real interest rate; i indexes individuals, here and henceforth. The simple functional form and clear interpretation of the slope coefficient are significant compensation for the loss of realism produced by ignoring college costs and assuming infinite working life. Suitably augmented with terms for work experience, etc., this equation has held up well.

Scale analysis does not just help develop heuristics. It also indicates when they do and do not apply. In today's United States, for example, Mincer's first simplifying assumption no longer holds, as college costs have grown far more than high school graduates' earnings have. The interpretation of the schooling coefficient should be tempered accordingly.[3] The accuracy of the second simplifying assumption would be similarly debased wherever lifespans are sufficiently short.

3.4 Non-reductive Uses of Scale

While these reductive uses of scale are important, the concept can be useful in other ways, too, that give you a better handle on just what is going on in your study. Three are worth mentioning now (and revisiting later).

First, scale analysis can distinguish theoretical or empirical findings that are at heart low-order, and thus heuristic in nature, from more subtle, higher-order findings. Sometimes, we encounter elaborate equations whose novelty and robustness are hard to assess. Often, such an equation can be thought of as a robust, first-order, heuristic trunk festooned with higher-order branches, in which the novelty resides. To determine how much these branches influence the equation's implications, bring

[3] I don't know if this has happened in practice, however, and I'm scared to look and see.

out the equation's low-order properties using scale analysis, and compare them to the properties of the equation itself. For example, Fig. 3.1 suggests that Eq. (3.5) is mostly trunk, since actual mortality hugs the asymptotes so closely.

Second, to truly understand the nature of the key variables in your analysis, identify the spatial and/or temporal scales over which they move. This was essential to our opening excerpt: only by understanding the lengths and periods of the waves of interest could Pedlosky distinguish large-scale motion from the flushing of a toilet. Similarly, for the Child Rights Act from Chap. 2, it might be better to think of the Act as being adopted regionally, over a time scale of a decade, than by individual states over the course of a year.

Finally, scale can help you understand the magnitudes implied by your estimates. McCloskey has taught us to attend to economic significance as well as statistical significance, and we generally recognize when our estimates take the wrong sign, or are unusually large or small. But don't stop there. Probe to see if all essential magnitudes implied by your estimates are reasonable, even those that aren't immediately obvious.

My first accepted paper was motivated by such probing (Grant 1999). The context was the "recycling problem," in which a monopolist produces a durable commodity that is sold to consumers and eventually recycled by secondary producers, who then compete with the monopolist. As increasing amounts of material ends up in the hands of secondary producers, the problem eventually reaches a steady state.

Two papers had already analyzed the archetypal recycling problem: the U.S. aluminum market before World War II. In both, the estimated price coefficient for secondary supply was positive and statistically significant.[4] This was reassuring at first glance, but a deeper look revealed trouble: the first paper's estimates implied that less than 1% of the available stock of aluminum was recycled each year, while in the second paper it was 100%. The first number implies that the typical aluminum-bearing product lasted over 100 years before being recycled—an order of magnitude too large. The second implies it lasted just 1 year—an order of magnitude too small. This alone shows that something was seriously wrong with both papers, as this time frame is an integral aspect of the recycling problem. My more general model estimated the interval to be about 15 years—not only the right order of magnitude, but seemingly in the right ballpark. The major aluminum-bearing products of the time, such as cookware and automobile engines, had lifespans about that long. That was very reassuring.

3.5 Conclusion

In the end, scale, like systems, isn't a mathematical tool, but a conceptual tool. It doesn't diminish the value of precise, unbiased estimates, or make them harder to achieve. So reconcile yourself to a world in which a broad sense of magnitude is not the enemy of good research, but its consort. Scale offers guidance in making modeling choices, both theoretical and econometric, and assistance in gauging the theoretical

[4] One only because of a major data error (Swan, 1980)

and quantitative implications of your analysis. None of this is averse to obtaining good estimates—and it can help you avoid obtaining bad ones.

The main limitation of scale analysis is that it is sometimes difficult to conduct. In contrast to many physical quantities that readily admit of measurement, behavioral or social phenomena can be hard to quantify, as can the magnitudes of the "forces" underlying them. But let's not make the perfect the enemy of the good. Though obtaining this knowledge often takes time, it is time well spent, for scale analysis is, in fact, a very cost-effective way to enhance the credibility of your findings. That is why so many fields employ it to begin with.

Food for Thought

1. "In reasonably competitive markets, a second-order change in costs yields a first-order change in profit." Treat costs in this statement as average accounting costs, which is the most likely way that a business professional would use the term in this context. Show this statement is true, always working with multiples of ten, and using a first order approximation for the typical business's profit margin as a fraction of sales.

2. My grandfather is a profit maximizing farmer. One spring, after he has ordered his seed, insecticide, and fertilizer for the season, I give him $10, to be spent on seed, insecticide, or fertilizer: just one of the three. Which one should my grandfather spend this $10 on: seed, insecticide, or fertilizer?

   (a) The correct answer is: it doesn't matter. Explain why, employing the concept of scale.
   (b) Assume a unit of seed, insecticide, and fertilizer each costs $1. Then write out a Taylor series expansion for the marginal product of ten units of each, to the second order. The answer in (a) assumes what about the relative magnitudes of the terms in the Taylor series expansion?
   (c) Would the answer change if I gave my grandfather $1, instead of $10? What about $1000? Explain.

3. The text refers to the unequal populations of the states or provinces within Brazil, India, and Indonesia, the nature of the heteroskedasticity that this gives rise to, and the problems that result by "overcorrecting" using population weights. Consider three analyses of the same phenomenon, estimated using Eq. (3.3): one using data from Brazil, another data from India, and another data from Indonesia.

   (a) Intuitively, in which country should the "overcorrection" problem be largest? Smallest? You will have to look up state populations for each in order to answer.
   (b) Can you articulate the heuristic(s) you used in answering part (a)? What statistics weighed most heavily in determining your answer?

Table 3.1 Relative automobile crash risks by BAC (from Blomberg et al. 2009)

Blood alcohol concentration (g/dl)	Relative crash risk: final adjusted estimate	Blood alcohol concentration (g/dl)	Relative crash risk: final adjusted estimate
0.00	1.0	0.13	12.1
0.01	1.0	0.14	16.4
0.02	1.0	0.15	22.1
0.03	1.1	0.16	29.5
0.04	1.2	0.17	39.1
0.05	1.4	0.18	51.0
0.06	1.6	0.19	65.3
0.07	2.1	0.20	81.8
0.08	2.7	0.21	99.8
0.09	3.5	0.22	117.7
0.10	4.8	0.23	134.3
0.11	6.4	0.24	146.9
0.12	8.9	0.25+	153.7

Note: Reprinted from the *Journal of Safety Research*, 40(4), pp. 285–292, 2009, with permission from Elsevier

4. Table 3.1 lists relative automobile crash risks as a function of blood alcohol concentration. The relationship in this table can be described with a very simple and natural heuristic, expressed not in terms of blood alcohol concentration, but in "standard drinks." Articulate this heuristic.

5. Table 3.1 confirms a truism about drinking drivers: though rare, they are far more likely to crash than sober drivers are. Using this fact, we can relate the incidence of traffic crashes to the extent of drinking and driving in a simple, clean way.

   Define s and d as the number of miles driven by sober and drinking drivers, r as the average per-mile crash risk of sober drivers, and k as the average crash risk of drinking drivers relative to sober drivers. None of these are observed, only the total number of miles driven and the number of accidents involving sober and drinking drivers.

   (a) Using scale analysis and a Taylor series approximation, derive a linear equation relating crashes per mile to r and the fraction of crashes involving drinking drivers.
   (b) Does this relationship hold when k varies across time, or does it need to be generalized? Does it hold in first differences? Does it hold in Turkey, where very few road accidents involve alcohol? What about the country of Blottonia, where almost all do?
   (c) In practice, no national repository records all the traffic accidents in the U.S. How, then, might we put this identity into practice with the data that is actually available?

References

Blomberg R, Peck R, Moskowitz H, Burns M, Fiorentino D (2009) The Long Beach/Fort Lauderdale relative risk study. J Saf Res 40(4):285–292

Cameron AC, Trivedi PK (2005) Microeconometrics: methods and applications. Cambridge University Press, Cambridge

Gavrilov LA, Gavrilova NS (2001) The reliability theory of aging and longevity. J Theor Biol 213(4):527–545

Grant D (1999) Recycling and market power: a more general model and re-evaluation of the evidence. Int J Ind Organ 17(1):59–80

Lewis M (2003) Moneyball: the art of winning an unfair game. WW Norton & Company, New York

Mincer J (1974) Schooling, experience, and earnings. University of Michigan Press, Ann Arbor, MI

Mincer J (1975) Education, experience, and the distribution of earnings and employment: an overview. In: Juster FT (ed) Education, income, and human behavior. National Bureau of Economic Research, Boston, pp 71–94

Pedlosky J (1979) Geophysical fluid dynamics. Springer, New York

Solon G, Haider S, Wooldridge J (2015) What are we weighting for? J Hum Resour 50(2):301–316

Swan PL (1980) Alcoa: the influence of recycling on monopoly power. J Polit Econ 88(1):76–99

Part II
Ways of Seeing

Chapter 4
Vernacular Knowledge

Abstract This chapter argues that an empirical analysis of an economic phenomenon should be rooted in a thorough understanding of that phenomenon's social, institutional, and technical context. It discusses what it means for a study to be so rooted, and how to go about acquiring the appropriate level of knowledge required. These ideas come to life in applications to whaling, sports betting, participation in union certification elections, and more.

Systems and scale are the yin and yang of your study preliminaries. The first is expansive, and serves a broadening function, identifying factors or relationships that need to be considered. The second is reductive, and serves a narrowing function, identifying things that are small in magnitude and thus can be cast aside. But both spring from the same root: an intimate familiarity with the phenomenon you are studying. This familiarity comes partly through formal measurement, the subject of the next chapter, and partly through informal, qualitative understanding, the subject of this one. Following Swann (2006), let's call this understanding "vernacular knowledge."

Vernacular knowledge is earthy, not abstract. It consists of the world of the agents you are modeling: what they observe and worry about, what they ignore, and what is beyond their control that they may or may not understand. This includes technical detail, such as production methods; institutional detail about market organization, functioning, and regulation; and social detail such as agents' circumstances, motivations, and social structures.

While vernacular knowledge helps you understand scale and systems, its fundamental purpose is deeper than that. To see what it is, let's turn to the work of a master.

D. Grant, *Methods of Economic Research*, Springer Texts in Business and Economics, https://doi.org/10.1007/978-3-030-01734-7_4

4.1 The Role of Vernacular Knowledge in Economic Research

In an early chapter of his classic, *Moby Dick* (1851), Herman Melville sends his narrator, Ishmael, to Nantucket to find work on a whaling ship. The account of his search amounts to a vast, cleverly disguised disquisition on the economics of the industry.

> Seated on the transom was a most uncommon and surprising figure: Captain Bildad, who along with Captain Peleg was one of the largest owners of the vessel; the other shares being held by a crowd of old annuitants; widows, fatherless children, and chancery wards; each owning about the value of a timber head, or a foot of plank, or a nail or two in the ship. People in Nantucket invest their money in whaling vessels, the same way that you do yours in approved state stocks bringing in good interest.

This ownership is, naturally, separated from control:

> "Sir, I want to see what whaling is. I want to see the world."
>
> "Want to see what whaling is, eh? Have ye clapped eye on Captain Ahab?"
>
> "Who is Captain Ahab, sir?"
>
> "Aye, aye, I thought so. Captain Ahab is the Captain of this ship."
>
> "I am mistaken then. I thought I was speaking to the Captain himself."
>
> "Thou art speaking to Captain Peleg. It belongs to me and Captain Bildad to see the ship fitted out for the voyage, and supplied with all her needs, including crew. We are part owners and agents. But if thou wantest to know what whaling is, I can put ye in a way of finding out before ye bind yourself to it. Clap eye on Captain Ahab, young man, and thou wilt find that he has only one leg."

This separation allows what Leeson (2007) calls captain predation:

> Now, Bildad, I am sorry to say, was, in his sea-going days, a bitter, hard task-master. They told me in Nantucket, that when he sailed the old Categut whaleman, his crew, upon arriving home, were mostly all carried ashore to the hospital, sore exhausted and worn out. He never swore at his men, they said; but somehow he got an inordinate quantity of cruel, unmitigated hard work out of them. When Bildad was a chief-mate, to have his drab-coloured eye intently looking at you, made you feel completely nervous, till you would clutch something—a hammer or a marlin-spike, and go to work like mad, at something or other, never mind what. Indolence and idleness perished before him.

Oh. But at least there is the incomparable ocean view?

> "Now then, thou wantest to go a-whaling in order to see the world? Well then, just step forward there, and take a peep over the weather-bow, and then back to me and tell me what ye see there."
>
> Going forward, I perceived that the ship, swinging to her anchor with the tide, was now obliquely pointing towards the open ocean. The prospect was unlimited, but exceedingly monotonous and forbidding.
>
> "Well, what's the report?" said Captain Peleg when I came back; "what did ye see?"

> "Not much," I replied—"nothing but water."
>
> "Well, what does thou think then of seeing the world? Do ye wish to go round Cape Horn to see any more of it? Can't ye see the world where you stand?"

And, slowly, we get it. These guys are stuck with each other for a long time in remote, austere conditions, ruled with absolute authority by an agent—the captain—with minimal investment in the vessel under his command. I wonder if anything will go wrong?

These quotes are fiction, true, but only in nominal terms. Melville lived the whaling life before he wrote of it; *The Perfect Storm* (1997) contains the same material in nonfiction form. In essence, these passages contain the type of institutional description that can be found in many economics articles, with some window dressing thrown in.

But to focus on this similarity is to miss the point. Melville is not providing us with general background information, or merely acquainting us with the life of a sailor. He is developing the foundation upon which his epic is built. The themes of his tale require that it be firmly grounded in reality—otherwise the book is worthless. Thus Melville must stealthily show that Ahab's behavior is not fantastic, in the original sense of the word, but merely an extension of conditions that already exist in this industry: harsh in the unrelenting vigor of nature and in the treatment of the men who face it. This is what we discover in these quotes and their companions. Only then are we ready to join Ishmael on his long journey.

Your empirical study should be rooted in reality this same way. Vernacular knowledge is the soil in which your study is planted. The richer the soil, the stronger your study will grow. Cognitive scientist Daniel Willingham (2009, p. 67) emphasizes that "we understand new things in the context of things we already know." Context matters, even for phenomena that themselves do not depend on context, and even more for social phenomena that do. The ultimate purpose of vernacular knowledge is to root your study in the real-world context of the phenomenon you are analyzing.

4.2 Vernacular Knowledge as Context

Applied microeconomists do develop and communicate new knowledge in context. We use three types, in fact: the literature, to place our work within existing research; the facts of the situation, what I have called vernacular knowledge; and the "technical context"—the theory and models around which the analysis is framed.

But we rely on this last type the most, almost as a crutch, giving short shrift to the other two types. Theory and models are thoroughly developed in most papers, and dwelled on in seminars. Meanwhile, literature reviews are more cursory, as discussed in a later chapter, as is the presentation of vernacular knowledge. We often do just enough to justify conducting our study, to support our model and empirical methods. This approach provides us with a common language with which to communicate our research and a common basis on which to evaluate it. But it does not maximize the correctness or the credibility of our results. Our focus on technical context prevents us from fully appreciating the importance of the other two types.

So does our training, which, for better or worse, does not catalogue the various behaviors, organizational forms, types of motivation, formal and informal institutions, social networks, and types of interactions to be found in economic life.[1] This is somewhat unique among the social and natural sciences. In the classroom, the field, and the lab, biologists learn a pantheon of behaviors exhibited by cells, plants, and animals, and their relation to the habitats in which they reside, as chemists do for chemicals, in isolation and in combination with each other, and psychologists and anthropologists

Table 4.1 Measures of cost: Conrad's coffee shop (from Mankiw 2012)

Quantity of Coffee (cups per hour)	Total Cost	Fixed Cost	Variable Cost	Average Fixed Cost	Average Variable Cost	Average Total Cost	Marginal Cost
0	$ 3.00	$3.00	$ 0.00	—	—	—	
							$0.30
1	3.30	3.00	0.30	$3.00	$0.30	$3.30	
							0.50
2	3.80	3.00	0.80	1.50	0.40	1.90	
							0.70
3	4.50	3.00	1.50	1.00	0.50	1.50	
							0.90
4	5.40	3.00	2.40	0.75	0.60	1.35	
							1.10
5	6.50	3.00	3.50	0.60	0.70	1.30	
							1.30
6	7.80	3.00	4.80	0.50	0.80	1.30	
							1.50
7	9.30	3.00	6.30	0.43	0.90	1.33	
							1.70
8	11.00	3.00	8.00	0.38	1.00	1.38	
							1.90
9	12.90	3.00	9.90	0.33	1.10	1.43	
							2.10
10	15.00	3.00	12.00	0.30	1.20	1.50	

Note: from *Principles of Economics*, 6th edition, by Mankiw, N. Gregory (2012). Reproduced with the permission of South-Western Cengage Learning; permission conveyed through Copyright Clearance Center, Inc.

[1] Furthermore, much of the knowledge that we do acquire in this training is episodic, based on an isolated circumstance that may not be representative of the phenomenon in question. Such studies include the primary example of efficiency wages, Henry Ford's doubling of pay in 1914 (Raff and Summers, 1987); the primary example of the effect of effort incentives, Lazear's (2000) analysis of an auto glass company; the primary example of information on consumer choice, Los Angeles' required posting of restaurant hygiene "report cards" (Jin and Leslie 2003); the primary example of predatory pricing, Standard Oil around 1900 (McGee 1958); and the primary example of the health effects of improved air quality, Chay and Greenstone's (2003) analysis of the 1981–1982 recession.

do for people. Without this kind of catalogue, it's harder for economists to give this type of context the attention it deserves.

Not that that many economists appreciate its importance to begin with. Even at the principles level, many texts utilize unrealistic values for costs and report no real-world price or income elasticities of demand. Table 4.1, from Mankiw (2012, p. 266, and subsequent editions), is representative. Even the setup is problematic: in a chapter that lays the groundwork for the output choice of perfectly competitive firms,[2] Conrad's Coffee Shop serves its product on demand in an imperfectly competitive market. Then the numbers: Why are fixed costs so low, and who pays them by the hour, anyway? Why does diminishing returns set in so quickly? And how hard is it, really, to make ten cups of coffee in an hour? Your response that these numbers are only intended to be illustrative is my point exactly. Absurdities like these not only impede the learning process—which, as we've just said, relies on context—but also stealthily inculcate the irrelevance of context into our students, like Melville in reverse. Phooey on you, Conrad, and your stupid coffee shop. I hope you go out of business.

In economics, the relevance of real-world context stems from the fact that most economic phenomena are deeply embedded in social and institutional structures. The spirited Mr. Tawney (1926, p. 221) puts it this way:

> Few tricks of the unsophisticated intellect are more curious than the naive psychology of the businessman who ascribes his achievements to his own unaided efforts, in bland unconsciousness of a social order without whose continuous support and vigilant protection he would be as a lamb bleating in the desert.

In Europe, modern capitalist economies evolved from a medieval world wholly dominated by such structures, just as the capitalist whalers in *Moby Dick* followed traditional whaling communities, such as the Inuit, that operated on communitarian principles. The most visible manifestation of this transition in Tawney's own England was the (often forced) "enclosure" of common agricultural land into private use. This land, which had been held mostly for prestige, was profitably put to sheep farming, while the previous tenants went to work for subsistence pay in the slowly-industrializing cities.

> The aim of the great landowner was no longer to hold at his call an army of retainers, but to exploit his estates as a judicious investment. The prosperous merchant, once content to win a position of dignity and power in fraternity or town, now flung himself into the task of carving his way to solitary preeminence. To the immemorial poverty of peasant and craftsman, pitting their pygmy forces against an implacable nature, was added the haunting insecurity of a growing proletariat, detached from their narrow niche in village or borough, the sport of social forces which they could neither understand, nor arrest, nor control (Tawney, pp. 117, 118).

In America, a similar, later transition was spurred by the Industrial Revolution (O'Donnell, 2015, pp. 4–6):

[2] While diminishing returns applies universally, of course, it is used most fundamentally, in this book as in others, to explain the output choices of competitive firms.

> Pre-industrial order centered on the artisan. The economy and society was dominated by time-honored traditions, obligations, and codes that constituted what E.P. Thompson termed "the moral economy." The culture of early 19th century artisans celebrated their independent lifestyle and flexible work routine rather than the headlong pursuit of wealth. When artisans spoke of virtue—the placing of the common good above private interests—they emphasized a sensibility that dominated the master-journeyman-apprentice relationship, which restrained the unbridled pursuit of self-interest over the well-being of the trade.
>
> This centuries-old world rapidly disintegrated in the decades following 1815. Some master craftsmen, enticed by a growing ethos of individualism and profit, began to alter the traditional norms of craft production. They took in more apprentices but taught them only a portion of the craft. They substituted market-rate cash payments for customary obligations like food, board, and education. They eliminated the casual rhythm of the workday in favor of longer days regulated by a clock, fewer breaks, and no alcohol.

In these earlier times the scope of economic analysis as we know it today would have been heavily circumscribed. Modern economics is possible due to economic development and the social changes it has wrought—what Peter Drucker (1939) would call the inception of Economic Man. But this development has not eliminated social and institutional structures, only replaced them with modern equivalents, which radiate like a fractal out from the local community toward the wider world.

Our medieval cousins wouldn't recognize many of today's social norms, but their importance is undiminished. A sense of fairness pervades employment relationships, affecting wage and employment flexibility. Extrinsic incentives that run counter to these norms can backfire, by diminishing intrinsic motivation. Many contracts are highly incomplete, with the unwritten interstices filled by social convention and expectations of "fair dealing." Firms' sourcing, production methods, and employment terms are influenced by their customers' social norms.

Social networks have become less dense, but more varied and wide-ranging. Networking for job opportunities is a commonplace, but this is only the beginning. Firms are networked with suppliers (Scherner et al. 2014), consumers (this is but one aspect of Chinese *guanxi*—see Wong and Leung 2001), and each other (Japanese *keiretsu*). There are networks of workers (Acheson 1988), government officials (Marcum et al. 2012), and workers and government officials together (Ge 2014). Such networks shape preferences, gather and distribute information, influence collective decision making, and expedite action compared to the more ponderous methods of large bureaucracies, either civic or corporate.

Finally, the social control of traditional leaders has been replaced, to some extent, by that of administrative, civil, and criminal law. Yet, as any cop on the street could tell you, this law's creation and enforcement are deeply embedded in social and institutional structures, so much so that anthropologists conduct fieldwork within government organizations. Economists and political scientists have instructively compared formal and informal methods of addressing market failures, especially common property resource problems. Neither side gets a knockout blow.

We need not turn our research methods inside out to accommodate these facts. But it is callow to peremptorily dismiss their effects simply as "regression error." This is appropriate only when these effects are small or unsystematic, and usually they are neither. Then your study must be shaped accordingly.

4.3 Vernacular Knowledge in Action

Just how should you go about doing this? The uses of vernacular knowledge are so varied that there is no simple guide. Many of these uses are best illustrated in future chapters, where we discuss the limits of your data, the testing of hypotheses, and the legitimacy of your empirical findings. Vernacular knowledge leaves no part of the research process untouched.

Still, as the footing on which your study is founded, its central role lies in model-building. A theoretical model is informed by economic theory, but not founded on it. An econometric model is informed by econometric theory, but not founded on it. Theoretical and econometric models are not castles in the air. They are *built on something*, and that something is a good working understanding of the phenomenon of interest. This comes, in large part, from vernacular knowledge.

In this role vernacular knowledge serves three purposes, which appear in order in the applications that follow. It can establish the accuracy—or inaccuracy—of causal claims or model assumptions. It can indicate the relevance or irrelevance of economic and social forces, things that must be considered or that can be ignored. And it can gauge the sensitivity of the model, and its findings, to the institutional, social, and technical environment.

4.3.1 Point Shaving in College Basketball

In our profession, the field most connected to vernacular knowledge is sports economics. It is just dripping with context. These folks *love* sports. As a result, it didn't take long to uncover the flaw in Wolfers' (2006) attempt to show that point shaving in college basketball is widespread.

Wolfers lays out how this point shaving would work:

> The University of Pennsylvania played Harvard on March 5, 2005, and was widely expected to win. The spread offered by bookmakers was 14.5 points, meaning that a bet on Penn would pay only if the margin of victory was at least 15 points, which is a comfortable margin.
>
> This example is ripe for corruption. The joint surplus of the Penn players and a gambler betting on Harvard occurs if Penn wins the game but fails to cover the spread. The contract required to induce this outcome simply involves the gambler offering a payment to the player that is contingent on Penn failing to cover the spread. Given the player's rough indifference to the size of the winning margin, even small bribes may suffice, and these, in turn, yield large profits for the gambler who has bet accordingly. There is an easy way for the

gambler to commit to paying this outcome-contingent bribe: simply give the player the ticket from a $1,000 bet on his opponent not covering the spread.

Wolfers argues that this should be most common among strong favorites, like Penn in his example, that can still win the game comfortably without covering the spread. Sure enough, the data show exactly this: games between evenly matched opponents cover the spread about half the time, but when the opponents are not evenly matched, an unusually high proportion of strong favorites fail to cover the spread. The estimated frequency of attempted point shaving in these games is 6%.

Oh my. I am not sure which is more shocking, the charge that players try to fix 6% of uncompetitive games, or the idea that a player would attempt to cash a ticket on a game he played in.

A telling feature of Wolfers' article is the absence of any mention of the expected costs of point shaving, a felony at the state and federal level, which would permanently end the professional careers of anyone involved. This is present in Borghesi (2008), who sets things straight. He shows that the expected costs to professional athletes far exceed the potential benefits: "it is implausible that conspirators could avoid detection when betting the enormous amounts of money that would be required to compensate professional athletes for shaving." Thus "in the past fifty years, there has not been a single documented case of a player fixing games in any of the four major American professional sports." So, when Borghesi shows that uncompetitive contests in pro basketball and pro football fail to cover the spread at rates similar to those for college basketball, it undermines the legitimacy of Wolfers' test.

Borghesi then provides a more tenable explanation for these findings, which is based on vernacular knowledge obtained from papers and reports in finance, gambling, and sports economics and from quotes by sports book managers. Bookmakers typically "balance their books," setting the spread so that roughly equivalent amounts of money are bet on either side, and making their profit from the vigorish (the commission for taking the bet). But bettors are inclined to over-favor teams that are strong favorites to begin with. In these games bookies will "shade the line," raising the spread above a balanced book and putting some of their own money at risk. This profit-increasing strategy reduces the fraction of games that cover the spread, but only in those contests that are not expected to be close—just as Wolfers and Borghesi find.

Wolfers' paper was built on the assumption that bookies always balance their books. That faulty foundation led to implausible findings. Either could have been smoked out using vernacular knowledge.

4.3.2 The Incentive Effects of Grades

When classes are graded on an A/B/C/D/F system, students who lie on the border between two letter grades have an incentive to "bump themselves up" with a strong performance on their final exam. I expected it would be easy to confirm, empirically, that students do this, and even imagined that I could see it in my own classes, as I entered students' exam scores into my spreadsheet at the end of the semester.

When my coauthor and I estimated how much this happens, however, we got a surprise: there was no effect at all. Not in my own classes, nor in others' (Grant and Green 2013).

I was unprepared for this. So were seminar audiences. This finding directly contradicted a basic economic tenet in one practical area where economists have great experience—the classroom.[3] We all viewed the situation in econometric terms. Something must be misspecified! An important variable must be omitted! The tests must lack statistical power! So my coauthor and I added statistical tests, purged even minor assumptions from the specification, and conducted additional robustness checks. Nothing. At this point we had a total of 20 nonparametric tests across four data sets, and the distribution of p-values across these tests was uniform—a strong sign of support for the null hypothesis. Still, in seminars, the hunt for econometric problems went on.

Looking for these problems is natural and appropriate, and it made our paper stronger. In the end, however, the issue wasn't econometrics, but a lack of vernacular knowledge. Context matters. Psychologists find that one set of tasks responds well to incentives, and another set does not. Studying for exams falls in that second set. I was unaware of this, as were my seminar audiences.

In the end, this was remedied by rummaging through the literature on educational psychology. This should have happened in the beginning: it wasn't too much to ask, and was highly relevant to the topic. After doing so, we knew to frame the question not as an obvious test of the pertinence of economic theory, but as an inquiry into the effectiveness of a particular type of motivator in circumstances where it may or may not be effective.

That's not the only vernacular knowledge I acquired too late. After the paper was in press, I told my students about the project, and asked them if they focused their exam-week effort on the classes whose grades were near the borderline. It was a reasonable question. After all, students doing this would do so deliberately, and would be aware of their intent. Fortunately, their response was a relief. They looked at me as if I was crazy to even suggest such a thing.

4.3.3 Turnout in Union Certification Elections

Some models turn out to be tall, elegant towers that crumble and fall with the slightest shift in the ground below. Vernacular knowledge can help you build sturdier models and can delineate the circumstances in which a given model should and should not apply. Thus, in the grades example, vernacular knowledge says not to extrapolate our null finding for exam performance to, say, extra credit assignments, because these lie in the set of tasks that do respond well to incentives. Our findings aren't that general.

[3] Of course, we included a scale analysis showing that the expected labor market benefits of higher grades were worth the study time required to achieve them. But this did not make people any more receptive.

I was vexed by this issue when testing a "rational voter" model of election turnout, which posits that more voters will turn out in close elections, because they might end up casting the deciding vote. My initial foray, using a sample of congressional elections, was successful, as we will see later. Then I went double or nothing. With a coauthor, I refined the model and tested it again, this time on a set of union certification elections, which determine whether the union in question has the right to represent the workers in that "bargaining unit" (Grant and Toma 2008).

I was very optimistic. Econometrically, this sample was far superior: it heightened the effect that was expected to obtain, contained many more observations, and expanded the range of tests that could be conducted. Yet, empirically, the basic turnout patterns in the data turned out to be almost comically out of sync with the predictions of theory. The model failed catastrophically. And when it did, I was defenseless. I had no answer for the questions that naturally arose. Why did the model apply in one set of elections and not the other? Why did it fail? Under what circumstances should you expect it to succeed? If our model doesn't explain turnout in these elections, what does?

The paper *contained* vernacular knowledge: a dutiful discussion of the mechanics of union certification elections and their outcomes. But it was not *rooted* in vernacular knowledge. Just what happens, in the workplace, in the interval between the filing of the election petition and the election itself? What pressures—from the employer, the union, from co-workers—operate on the voting decision, and how do they work? Does the easy visibility of turnout, in an election held at work, matter? Does the winning margin or the extent of turnout affect the union's post-election effectiveness? We didn't know. We had no sense of how this situation unfolds, from the perspective of a worker in that bargaining unit. As a result, we never understood why our model failed, what should replace it, or what its range of application should be. To this day the paper feels incomplete.

4.4 Acquiring Vernacular Knowledge

With such a wide range of potential uses, how can we anticipate when and where vernacular knowledge will turn out to be useful? Without the aforementioned catalogue of behavioral and institutional variation, how can we be sure to acquire the right pieces of vernacular knowledge, instead of useless bits that get left on the cutting room floor?

You can't, of course. The acquisition of vernacular knowledge is subject to diminishing returns, like anything else, but you cannot be overly utilitarian in your search. Still, there are guiding principles that will light your way.

The first is that the amount you need depends on the characteristics of your study. The more you ask of your model—in causal understanding, in generalizability, in precision—the more vernacular knowledge you need. Thus a model of turnout in congressional elections requires less vernacular knowledge than a more general turnout model does.

The second principle is to let your lodestar be intimacy: to see things the way the agents you are modeling do. Caroll Spinney puts it this way: "You have to have a good sense of the deep motivations of the character." Imagine you sat down with participants in the phenomenon you are studying and laid out your model, data, and results in laymen's terms. Their critiques—what they think you got right, got wrong, and forgot to consider—would be based on vernacular knowledge.

We cannot obtain this intimacy without recognizing that vernacular knowledge is local, shaped by the various circumstances in which people find themselves. Sometimes this is easy to see. We know production methods respond to an important aspect of the economic environment, input prices, just as inputs' marginal products depend on other environmental factors, from the weather to transport networks. We are thus well-primed to appreciate differences in the production methods of, say, Chinese, Egyptian, and American rice farmers. But elsewhere neither theory nor intuition is such a ready guide. Why are some federal agencies more effective than others, even when their tasks are similar? What explains the variety of organizational forms of indigenous peoples across the Americas? There are probably explanations for these phenomena, but we probably don't know them.[4] Then all we can do is be aware of the variety of circumstances that pertain, to the extent it impinges on our research question.

This is hard to anticipate in advance, which motivates the final principle: it is OK for the acquisition of vernacular knowledge to be piecemeal and unsystematic. Sometimes, it simply involves broadening your knowledge beyond the narrowness of your experience: reading about various times, places, and people; chatting with workers in different occupations; going to diverse sessions at conferences. Sometimes, the terms are specific to a particular phenomenon or data set: communicating with individuals actively involved in this phenomenon, digging through obscure documents, or scouring the literature in neighboring fields, where vernacular knowledge plays a larger role. Either way, the amount you acquire will inevitably exceed the amount that you need, which in turn will exceed the amount you convey to your readers.

4.5 Conclusion

When you boil it down, scale, systems, and vernacular knowledge simply help you develop a good working understanding of the phenomenon of interest. Vernacular knowledge is the raw material, a hodge-podge of context; scale and systems prune that material, organize it, to reveal the structure underneath.

This is unobjectionable, as a matter of logic, yet still it feels unfamiliar and strange. I think this stems from our profession's focus on the local and the small. We theorize in marginal terms: how small changes in one variable influence another. We

[4] There are! See, for example, Mashaw and Harfst (1990) and Cornell and Kalt (1995), respectively.

use regression discontinuity (RD) designs, which identify the coefficient of interest from small changes in the "running variable." We estimate how one variable affects another, holding constant everything else. We do these things for good reasons, which I needn't bother to elaborate. But all your thinking need not be this way.

Sometimes we need to think big: by identifying the interlocking mechanisms governing the motion of your dependent and independent variables, with systems; by distinguishing the large from the small, with scale analysis; by relating your model to agents' everyday lives, with vernacular knowledge. This helps us shape our study to best fit the circumstances, avoid multiple sources of error, and recognize the limitations of our analysis. Part of good craftsmanship is to temper the small and local with the big and grand.

Food for Thought

Answering these problems may necessitate a modest amount of groundwork. This is intentional, since this is how most vernacular knowledge is acquired.

1. Another provocative sports economics paper, Romer (2006), concluded that professional football teams punt far too often on fourth down, instead of going for another first down by running or passing. Using a dynamic programming model, Romer concludes that, at midfield, one should go for it on any fourth down with five yards or less to go (to earn another first down), and that it should do so even on its ten yard line if it needs only three yards to go.

   Skepticism should meet these findings, because teams' actual choices concur with them only 10% of the time. The weak link is that, because "going for it" on fourth down is so rare, Romer analyzes third down plays instead, arguing without formal evidence that "one would not expect either side (offense or defense) to behave very differently on the two downs." Evaluate this claim informally using vernacular knowledge. Do football aficionados agree that the fourth down plays offenses usually run, and the defenses that are used against them, resemble those used on third downs?

2. The best known paper on the topic of job networking involves the "strength of weak ties," based on Granovetter's (1973) finding that people find more jobs through acquaintances that they don't know all that well, rather than from friends that they do. Treating the job hunter's friends and acquaintances as a network, sketch the idea of information flow that underlies Granovetter's conclusions (which have been recently questioned). Compare this information flow with that underlying Hayek's "The Use of Knowledge in Society" (1945).

3. Jensen and Murphy (1990) argued that the compensation of chief executive officers (CEOs) was not sufficiently performance-based, harming firms' stock values as a result. The following decade saw dramatic growth in use of a form of executive compensation that is particularly sensitive to the value of the firm: stock options. Yet Jensen (2001) explicitly disavowed conventional stock options, arguing they actually encouraged poorer CEO decision-making.

(a) The growing use of stock options in the 1990s, and its reversal in the following decade, had little to do with incentivizing sound managerial decision-making. What did it have to do with instead? The answer to this puzzle involves vernacular knowledge.
(b) The use of performance-based incentives is supposed to remedy the principal-agent problem and align the objectives of the CEO with those of stockholders. Yet much about the reward structure of CEOs suggests that the principal-agent problem applies to board of directors as well. Give at least two significant examples, which can involve, but are not limited to, the facts reported by Jensen and Murphy and the treatment of stock options.

4. The acquisition of a large body of vernacular knowledge has been key to the work of several Nobel Laureates. Illustrate with regard to the work Coase and Ostrom.

5. The effect of malpractice liability on the physician services market is a big subject of study in health economics. Consider a federal law that decreases the number of malpractice suits against physicians, either by making it harder to sue or by reducing the payout in the event the suit is successful. This should reduce the premiums that physicians pay for malpractice insurance.

   (a) In the U.S., what kinds of laws have been used to accomplish this goal? As a matter of scale, are malpractice premiums a practice cost of sufficient magnitude to merit attention?
   (b) Using a standard price setting graph, illustrate the short run effects of such a law on the price and quantity of physician services. The price and quantity effects on this graph depend on whether malpractice insurance is a fixed cost or a variable cost. Which is it?
   (c) Using a standard price setting graph, illustrate the long run effects of such a law on the price of physician services. Approximately what period length would correspond to the long run in this context? Why?
   (d) The price effect on the graph you just drew depends on the degree to which increased profits can draw new doctors into the profession. How flexible is entry into U.S. physician services over the long run period that you specified above?

6. A well-known study of job displacement, Jacobson et al. (1993), finds that high-tenure, prime age, continuously-employed Pennsylvanian workers who lost their jobs between 1980 and 1986 received a persistent wage penalty in excess of 20%.

   (a) Which aspects of this study would affect the generality of its conclusions? Comment on how this study's findings might depend on the institutional, demographic, and macroeconomic aspects of the data it uses. Overall, would these aspects tend to push the wage effects of displacement up or down?
   (b) How do you think the extent of wage loss differs by industry? Compare your guesses to the authors' empirical findings.

References

Acheson J (1988) The lobster gangs of Maine. University Press of New England, Lebanon, NH
Borghesi R (2008) Widespread corruption in sports gambling: fact or fiction? South Econ J 74(4):1063–1069
Chay KY, Greenstone M (2003) The impact of air pollution on infant mortality: evidence from geographic variation in pollution shocks induced by a recession. Q J Econ 118(3):1121–1167
Cornell S, Kalt JP (1995) Where does economic development really come from? Constitutional rule among the contemporary Sioux and Apache. Econ Inq 33(3):402–426
Drucker P (1939) The end of economic man. Transaction Books, Piscataway, NJ
Ge Y (2014) Do Chinese unions have "real" effects on employee compensation? Contemp Econ Policy 32(1):187–202
Granovetter M (1973) The strength of weak ties. Am J Sociol 78(6):1360–1380
Grant D, Toma M (2008) Elemental tests of the traditional rational voting model. Public Choice 137(1):173–195
Grant D, Green WB (2013) Grades as incentives. Empir Econ 44(3):1563–1592
Hayek F (1945) The use of knowledge in society. Am Econ Rev 35(4):519–530
Jacobson LS, LaLonde RJ, Sullivan DG (1993) Earnings losses of displaced workers. Am Econ Rev 83(4):685–709
Jensen M (2001) How stock options can reward managers for destroying value and what to do about it (No. 480401). Social Science Research Network
Jensen M, Murphy K (1990) Performance pay and top-management incentives. J Polit Econ 98(2):225–264
Jin GZ, Leslie P (2003) The effect of information on product quality: evidence from restaurant hygiene grade cards. Q J Econ 118(2):409–451
Junger S (1997) The perfect storm. W. W. Norton, New York
Lazear EP (2000) Performance pay and productivity. Am Econ Rev 90(5):1346–1361
Leeson PT (2007) An-arrgh-chy: the law and economics of pirate organization. J Polit Econ 115(6):1049–1094
McGee JS (1958) Predatory price cutting: the standard oil (NJ) case. J Law Econ 1:137–169
Mankiw NG (2012) Principles of economics. Cengage Learning, Boston
Mashaw JL, Harfst DL (1990) The struggle for auto safety. Harvard University Press, Cambridge, MA
Marcum CS, Bevc CA, Butts CT (2012) Mechanisms of control in emergent interorganizational networks. Policy Stud J 40(3):516–546
Melville H (1851) Moby Dick. Harper & Brothers, New York
O'Donnell E (2015) Henry George and the crisis of inequality. Columbia University Press, New York
Raff DM, Summers LH (1987) Did Henry Ford pay efficiency wages? J Labor Econ 5(4, Part 2):S57–S86
Romer D (2006) Do firms maximize? Evidence from professional football. J Polit Econ 114(2):340–365
Scherner J, Streb J, Tilly S (2014) Supplier networks in the German aircraft industry during World War II and their long-term effects on West Germany's automobile industry during the 'Wirtschaftswunder'. Bus Hist 56(6):996–1020
Swann GP (2006) Putting econometrics in its place: a new direction in applied economics. Edward Elgar Publishing, Cheltenham
Tawney RH (1926) Religion and the rise of capitalism: a historical study. Harcourt, Brace & Company, New York
Willingham DT (2009) Why don't students like school?: a cognitive scientist answers questions about how the mind works and what it means for the classroom. Wiley, Hoboken, NJ
Wolfers J (2006) Point shaving: corruption in NCAA basketball. Am Econ Rev 96(2):279–283
Wong YH, Leung TK (2001) Guanxi: relationship marketing in a Chinese context. International Business Press, Horsham

Chapter 5
Data

Abstract This chapter portrays the economic analysis of data as a sophisticated way of "seeing like a state," a perspective which highlights the inherent limitations of most economic data. It articulates the data qualities the researcher should be familiar with, and the econometric consequences of failing to understand these qualities. These ideas come to life in applications to patents, informal markets in Peru, school accountability ratings, drug dealing, medical coding, and the employment effects of the minimum wage.

5.1 How to Think About the Economic Analysis of Data

Malcolm Gladwell's *Blink* (2005) opens with a vignette about an ancient Greek statue, an exceptionally well-preserved example of what is called a "kouros." The museum it was offered to moved carefully before buying it. They ran all the necessary tests, took all the necessary measurements, even had a bit of the core removed and analyzed to confirm the kouros's age and provenance. The only problem: it was a fake. Later evidence would establish this more conclusively, but the early signs were there too. Experts brought in to observe the statue had a visceral negative reaction, thought it didn't "look right." Turned out they were on the mark. The moral of the story? Trust your gut.

Unless you are a pilot. Then the mantra is different (Marshall 2016):

> When flight instrumentation first became viable, aviation educators began teaching pilots instruments-only flight. The gist was simple: only pay attention to your instruments. Many veteran pilots insisted that in a crisis situation the key was to go by intuition instead. But numerous experiments and tragic experience showed that this was not true. With poor visibility, pilots' perceptions were consistently wrong. Even today, much of a pilot's training turns on the difficult process of learning to disregard what your senses tell you is happening and following the instrument panel that tells you what actually is happening.

The moral of the story? Ignore your gut. Trust Your Instruments.

These vignettes contrast two ways of viewing the world. One—informal, unstructured, qualitative—is the province of vernacular knowledge. The other—formal, structured, quantitative—is the realm of data. As portrayed above, these two conflict: they make opposing claims, only one of which can be correct.

D. Grant, *Methods of Economic Research*, Springer Texts in Business and Economics, https://doi.org/10.1007/978-3-030-01734-7_5

This conflict is not foreign to academia. Its misbegotten kouros is basketball's "hot hand"—mid-game streaks during which players' shot-making ability is thought to rise. Long a belief among professional basketball players, it was convincingly debunked 30 years ago (Gilovich et al. 1985), in a study since repeatedly replicated: testament of people's ability to draw unwarranted causal inferences from patterns that occur by random chance.

Until it wasn't. The hot hand is real (Miller and Sanjurjo 2015), previous analyses flawed, some so far off they failed to detect even substantial hot-hands. Who would have guessed? Perhaps the professionals, who could perceive these patterns after years on the court. We should have listened to them after all.

But there are Trust Your Instruments moments to match. One of the best known is psychologist Paul Meehl's (1954) comparison of two methods of predicting behavior from observation: objective, formalized algorithms and clinicians' subjective judgment. Meehl's well-substantiated finding, a smashing victory for formalism, extends to a related literature in human resource management, which shows that structured interviews (with scripted questions) outperform unstructured interviews (without scripted questions) as predictors of job performance.

We could go on with stuff like this all day. The material is there, all right. But why would we do so? It would cause more harm than good.

In the first place, it is not constructive to frame the relationship between formal measurement and informal observation as us-vs.-them. This tension is a central feature of some social sciences, such as sociology. Perhaps, there, it is useful, with the qualitative and quantitative camps each serving to check the other. It would not be so in our profession. One side would win in a rout. As if we need to reinforce the self-indulgent illusion that data is the objective, hard-nosed, scientific alternative to subjective, informal, wishy-washy observation.

Furthermore, framing things this way sets up a false choice. The two perspectives are not mutually exclusive. For example, Daniel Kahneman's *Thinking Fast and Slow* (2011) stands the structured-unstructured interview dichotomy on its head, describing an analogous situation in which subjective assessments are valuable when, and only when, they are combined with objective performance indicators. Similarly, the book you now hold is replete with instances of vernacular knowledge embellishing empirical analysis, and vice versa.

We need a different way to think about what we are doing when we analyze data. This formulation should accept the primacy of data analysis in economics, yet impel us to an understanding of its limits. It should portray data and vernacular knowledge as the complements that they are, while placing the latter into the role of supporting the former. That formulation is to use the data that we have, mindful that, in doing so, we are "seeing like a state" (Scott 1998):

> Officials of the modern state are, of necessity, one or more steps removed from the society they are charged with governing. They assess the life of their society by a series of typifications that are always some distance from the full reality these abstractions are meant to capture. Thus the foresters' charts and tables do not quite capture the forest in its full diversity. Thus the cadastral survey and the title deed are a rough, often misleading representation of actual, existing rights to land use. The economic plan, survey map, record of ownership, forest management plan, classification of ethnicity, passbook, arrest record, and

> map of political boundaries acquire their force from the fact that these synoptic data are the points of departure for reality as state officials apprehend and shape it.

The economic analysis of data is merely a sophisticated way of seeing like a state.

5.1.1 Measurement and Legitimacy

This formulation seems rather political, but the inescapable reality is that measurement confers legitimacy on the people and activities that are quantified. Peruvian economist Hernando DeSoto knows this well, having spent his career detailing the costs of "informality"—illegitimacy in the eyes of the state—in housing, trade, and transport. In *The Other Path* (1989), DeSoto legitimizes these activities through the simple act of counting.

> In Lima, informal markets begin when vendors who are already operating on the streets seek to end the insecurity of doing so and begin to build their own markets, without complying with the legal provisions governing legally developed lots or land that has been overrun by squatters. The number of such markets was unclear when we began our research. They were simply ignored by official researchers, and there was no benchmark from which to make an estimate. Exhaustive field work showed that there were 274 informal markets in the capital city, as compared with 57 built by the state.
>
> Our researchers then assessed these markets, funded principally by street vendors who desired to move off the streets, and estimated their current value at $40.9 million. More detailed research showed that there were 38,897 people working at 29,693 stalls in these markets. Altogether, there are 439,000 people dependent on informal trade carried out on the street and in markets.

The excessive numerical precision seems odd, as does the sample design: there is no sampling. But that is because we are looking at things the wrong way. While unnecessary from the perspective of social science, both features are ideal for conferring legitimacy—for seeing like a state.

The conflux of measurement with legitimacy tempts us to overlook Scott's Hayek-ish warning that data often obscures as much as it illuminates. It crudely represents a much richer process. Our challenge is to see past the crudeness, and not limit our understanding of what is going on to that which is measured.

5.1.2 Patents

Take the study of patents, measures of innovative activity that, by definition, are seen by the state. It is the empiricist's classic conundrum. This data is obviously imperfect (Griliches 1990):

> One recognizes, of course, a whole host of problems associated with using patent data. Not all inventions can be patented, not all inventions are patented, and the inventions that are patented differ greatly in the magnitude of inventive output associated with them.

Yet what is better?

> When Schmookler began his work, data on research and development (R&D) expenditures or research-related employment barely existed. Even today, with data much more plentiful, the available detail in published R&D statistics is still quite limited. Thus, showing that patent statistics are a good indicator of inputs into inventive activity is a useful accomplishment on its own merit.

It is the perfect invitation to overly align one's thinking with the things you can measure—that is, to see like a state:

> Schmookler started out thinking he could use patent statistics as an index of inventive output and as an explanation of the growth in the aggregate efficiency of the U.S. economy. Unfortunately, the relationship did not work—there was little correlation between total factor productivity and the number of patents. Schmookler did not give up. In response, he redefined patent statistics as an index of inventive "activity," an input rather than an output. He moved, essentially, in the direction of what patents can measure rather than what we would want them to measure. His interpretation of inventive activity became quite narrow.

Recent work has shown how narrow, by looking outside the realm of patents. A few studies examine industry-specific output measures, such as sewing machine speeds or crop yields, to measure technical progress directly. Other data allows inter-sectoral comparisons (Moser 2005, 2016):

> A new, internationally comparable data source permits an empirical investigation of the effects of patent law on innovation. Constructed from the catalogues of 19th century World Fairs, it includes innovations that were not patented as well as those that were, and innovations from countries both with and without patent laws.
>
> These data show that most innovations were not patented at all, even those high-quality inventions that received prizes, even in countries with low patent fees. The absence of patent laws in a few countries, such as Switzerland, did not preclude high levels of innovation. Regressions yield no evidence that patent laws increased levels of innovative activity, but strong evidence that they influenced the distribution of innovative activity across industries, such that inventions in countries without patent laws were concentrated in industries where secrecy was effective relative to patents.

Here we experience what it truly means to understand that your data sees like a state. You must appreciate the "flaws" in your data, which could possibly be fixed, along with the "imperfections," which can't. Moser's results suggest a problem in the patent data that really can't be fixed. And Griliches didn't have those results to draw on, anyway. He simply recognized the data's inherent limitations as a measure of innovative activity. Imperfections like these occur even in good data, data that isn't exactly flawed. This is just what happens next, in the economics of education, where studies gorge on mountains of data that sometimes tell us less than it seems.

5.1.3 School Accountability

The U.S. school accountability movement that picked up steam in the 1990s began to pay dividends to researchers during the following decade, with the release of administrative data, collected and recorded by the state. This data contained

standardized test scores for thousands (Tennessee's Project STAR) or millions (North Carolina) of students, grouped by class, with teacher identifiers and basic student demographic information. Students could be tracked across time, sometimes even to later life outcomes, such as college attendance or earnings (the Texas Schools Project).

Before economists could access this data, unfortunately, education professors leapt on it and analyzed it so thoroughly there was hardly any meat left on the bones. Just kidding right there. This is exhibit #1 for the worlds that open up with strong analytical skills. Economists' possession of these skills lets them go where most education professors fear to tread.

A central objective of this literature is to estimate how much the teacher effectuates student learning. To do so, researchers migrated to the concept of "value added," the teacher's estimated effect on her students' current test scores, controlling for their previous year's score on the same test (reading, science, etc.). The coefficient on the prior score runs about 0.7, significantly less than one, testament to the vagaries of intellectual growth and its measurement. These and other statistical issues make the determination of value added a fairly complicated exercise; nonetheless, numerous studies find large variation in value added across teachers (Kane 2014).

The excitement surrounding these findings has been palpable, and indeed they are a significant achievement. Yet in such exuberance we can forget that, by focusing on what is measured, we are seeing like a state (Kane and Staiger 2002):

> In most states, the improvements on the test used for accountability purposes are far larger than improvements on the National Assessment of Educational Progress (NAEP) scores, which are not used for these purposes (Linn and Dunbar 1990). In Kentucky's fourth grade math test, for example, the former were four times the latter, in comparable units. When Koretz et al. (1996) asked Kentucky teachers to account for improvements like these, more than half said "increased familiarity" with Kentucky's test and "work with practice tests and preparation materials" had been important, while only 16% said the same for "broad improvements in knowledge and skills."

So the scores can be gamed, substantially, through test prep. On top of this are the skills these tests miss in the first place. Jackson (2016), for example, creates an index of "non-cognitive" skills from attendance, grades, etc., and compares it to the cognitive skills represented in test scores. Both are equally important in predicting ninth graders' future success, such as whether they enroll in college. But teachers' skill at improving the cognitive index, i.e., value-added, was virtually uncorrelated with their skill at improving the non-cognitive index. Chamberlain (2013) takes this a step further, linking teachers to the earnings of their students in their late 20s. He concludes that only about one-fifth of the effect of teachers on their students' early-career earnings is mediated through test score improvements, that is, through value-added. The literature has emphasized these test scores in vast disproportion to their importance. In doing so, it too sees like a state.

5.2 Validity

Recognizing how economic data sees like a state is a cautionary perspective, which serves to temper the conclusions of our analysis. The other side of this coin is an affirmative perspective that asks what the data does see, rather than what it doesn't. The term for this is validity: does the data measure the constructs we want it to measure?

Economists encounter this issue less frequently—one reason it is so easy to see like a state. Does GDP measure aggregate production? Does the Herfindahl Index measure concentration? Absolutely (though imperfectly). But this cannot be taken for granted elsewhere, in other social sciences or experimental economics, where the concepts of interest are less obviously measurable, in theory or in practice. Validity must be established, rather than assumed, by comparisons with measurements known to be valid, appropriate consistency checks, calibrations on groups known to have (or not have) the qualities being measured, etc. (see Litwin 1995).

In *Gang Leader for a Day* (2008), sociologist Sudhir Venkatesh learns this the hard way. A new Ph.D. student at the University of Chicago, Venkatesh sets out one Saturday afternoon to interview young men in some hard-up projects a couple of miles away.

> As I started to climb the stairs, the smell of urine was overpowering. On some floors the stairwells were dark; on others there was a muted glow. I walked up four flights, maybe five, and then I came upon a landing where a group of young men, high-school age, were shooting dice for money.

I get it, I get it! The darkness is a metaphor for how little he knows as a Ph.D. student! But—nice touch, saying "four flights, maybe five" instead of "four or five flights" in order to heighten the ambiguity.

> "Man, what the f— are you doing here?" one of them shouted. I tried to make out their faces, but in the fading light I could barely see a thing. The young men rushed up to me, within inches of my face. I told them the numbers of the apartments I was looking for. They told me that no one lived in the building.
>
> Two of the young men started to search my bag. They pulled out my questionnaire sheets, pen and paper, a few sociology books, my keys. Someone else patted me down. A guy with a too-big hat, who had taken my clipboard, looked over the papers and then handed everything back to me. He told me to go ahead and ask a question.
>
> By now I was sweating besides the cold. I leaned backward to try to get some light to fall on the questionnaire. The first question was one I had adapted from several other similar surveys; it that targeted young people's self-perceptions. "How does it feel to be black and poor?" I read. Then I gave the multiple-choice answers: "Very bad, somewhat bad, neither bad nor good, somewhat good, very good."
>
> The guy with the too-big hat began to laugh, which prompted the others to start giggling. "F— you," he said. "You got to be f—ing kidding me."

Lost in the humor of the story is the soundness of the Likert-scale method, which is supported by research. But even sound methods can't rescue what is, here, a preposterous question. Even within the context of the study, it is unclear what, if anything, it truly measures. Its internal validity is suspect.

Because most economic data has internal validity, this concept doesn't help us that often. We get more traction with the concept of external validity. Just as we want our theories to generalize, to apply to more situations than one, so too do we want this of our data, so that our study's findings extend beyond its borders.[1]

This point is central to Card and Krueger's (1994) well-known study of the minimum wage. Using the quasi-experimental approach popularized by Campbell and Ross (1968), they compare employment changes in fast-food restaurants in New Jersey, which raised the minimum wage, with those in Eastern Pennsylvania, which didn't, using survey data they collected themselves. Their finding was of a null effect at best—employment seemingly increased on the New Jersey side.

Econometrically, the authors demonstrate that this finding is quite robust, insensitive to various data handling and specification quibbles. What they don't demonstrate is the validity of their data. The collective criticism of this study raised three major concerns on this front (see, especially, Brown's, Hamermesh's, and Welch's comments in Ehrenberg 1995).

First is the suitability of the control group—a question of internal validity. I had always envisioned this study straddling the Delaware River, which separates the two states. But, in fact, metropolitan New York City dominates the New Jersey sample, while its Pennsylvania counterpart contained more…bucolic counties such as Bucks, Chester, and Luzerne (Card and Krueger 2000). Comparing these two is suspect.

Second is the focus on fast-food—a question of external validity. While this industry definitely employs many low-wage workers, it may not be representative of all industries, and its demand could shift as the minimum wage increase affects the income distribution.[2] Thus, it is quite troubling when Welch points out that changes in aggregate low-skill employment in these areas deviate substantially from the employment changes in Card and Krueger's restaurants.

Third is the study's nine-month time frame, which might be too short to gauge the full effects of an increase in the minimum wage. After all, this increase should not only raise marginal costs and prices, but also lower profits, driving firms out of the market. Nine months is sufficient to detect the first effect, but not the second (see Baker et al. 1999).

Better data may not have rescued this study, given its other problems (see Chap. 8's food for thought). Still, under the circumstances, supporting the data's internal and external validity, or trying to, should have been considered mandatory. Note that, for the first two concerns at least, simple description would have sufficed.

[1] This issue is especially important in experimental economics, for example, whether the findings of an experiment using college students as subjects apply to other populations.

[2] This appears to be an issue in recent studies of Seattle's increase in the minimum wage.

5.3 Getting to Know Your Data

Knowing what your data does and does not see, coupled with a good, quotidian understanding of its properties, can inform your research from start to finish.

It informs your econometric model by indicating how much you can ask of the data. Better data supports more complex estimators and identification strategies, and allows the estimation of weaker, more subtle relationships.

It informs your final conclusions by helping you see the big picture, how much light your data sheds on various parts of the system. We have seen, for example, that value-added measures of teacher effectiveness are not unsound, just a relatively small part of a much larger picture. The estimates of such a study need not be tempered, but the conclusions must be, nevertheless.

It also helps you troubleshoot your empirical estimates—a vital part of any economic analysis. For example, Moser found that patents had large effects on the type of innovation, but not the overall level of innovation. Thus cross-country or intersectoral analyses that proxy innovation with patents would have substantial omitted variables bias, even if patents correlate reasonably well with R&D expenditures, as they do, and even if patents enhance firm value, which they do.

For these purposes one should attend to three qualities of your data: precision, accuracy, and span.

5.3.1 Data Precision

Because his purpose was to legitimize a forgotten people, DeSoto could not afford for his measurements to be imprecise. Thus the number of informal markets was reported exactly: there were 274.

This level of precision is more than most researchers need. We often encounter variables that are imprecise. They measure something that varies continuously using multiples of one hundred, or a sequence of numerical intervals, or a binary variable indicating whether its value is relatively large or small. Even if the measurements themselves are not erroneous, this lack of precision can be consequential. We can overlook this when we too hastily accept the data as it is presented to us.

This is an under-appreciated issue in health economics, where administrative data abounds in fields of plenty, machine-readable and ready to go. Such ready access cloaks problems with precision. Much of this data contains codes: ICD codes to identify diagnoses associated with medical events, DRG codes to classify hospital admissions, CPT codes to identify the inpatient or outpatient procedures that were performed.[3] One cannot overstate the ubiquity of such codes in health care, in

[3] ICD: International Classification of Diseases, now in its 10th iteration. DRG: Diagnosis-Related Group. CPT: Current Procedural Terminology. Those are the codes sprinkled all over your bill after an office visit with the doctor. Present here, too, is such codes' ability to legitimize behavior, as with the 1980 addition of Post-Traumatic Stress Disorder to the Diagnostic and Statistical Manual of Mental Disorders, or the 1974 removal of homosexuality from that manual.

documenting and justifying the care that was delivered, and in getting reimbursed for providing it.

Economists use these codes when comparing the effectiveness of alternate forms of treatment, estimating the variation in physicians' practice styles, or imputing physicians' therapeutic skill (analogous to teacher value added). For these purposes one must "risk adjust" for the underlying health of the patient when they enter the medical system and for the acuity of the condition that they are being treated for. Diagnosis codes help you do that, since they record the conditions or complications exhibited when the patient seeks treatment.

But these codes often have little sense of severity, as designed or as they are actually recorded. Sometimes this is fine: a breech birth either is or isn't, and a binary code suffices. Sometimes it is not, as for hypertension, which can vary substantially in severity. Furthermore, this variation could be systematic across areas or providers (as could variation in how a given condition is coded). You are looking at the right things, but you are seeing them—like a state—in crude, blunt terms. For studies that try to tease out small effects conditional on many controls, such as two-way fixed effects, this could represent a significant problem.

5.3.2 Data Accuracy

This term does not refer to the avoidance of errors in *handling* the data—transforming variables, dealing with missing data, etc. These are more common than you would guess, at all levels of the profession. But there is little to be said about them beyond the obvious.

More can be said about the accuracy of the data itself. Most data is somewhat erroneous, due to sampling error, recording error, or subjectivity in recording. Sometimes this error is quantified, as with sampling error in the unemployment rates calculated from the Current Population Survey. Often it is not. Yet estimates can depend quite sensitively on data accuracy, because many estimators reduce the *effective* signal that identifies the coefficients of interest, and inflate the noise, as discussed at length below.

5.3.3 Data Span

Few data sets are fully comprehensive. The variables within them generally represent only a subset of everything that we'd like to know. Understanding your data includes understanding how large that subset is.

To fix thoughts, consider a basic regression with a single independent variable of primary interest, which may be correlated with any of several controls, some measured in your data, some possibly not. These controls are likely to be interrelated, so one can

distill from them, in linear algebra terms, a "control space": a set of underlying conceptual factors that influence the dependent variable and may also correlate with the independent variable of interest. For example, in a standard wage discrimination regression, a gender or race dummy might be the key independent variable, while the "control space" might have four dimensions: skill, cost of living and location amenities (such as the area's climate), job characteristics (such as safety or pleasantness), and labor market competitiveness. These, in turn, would be proxied by observable variables such as work experience, region dummies, union coverage, etc.

For each factor, you would like to know three things: (1) how large it is, that is, its variation, (2) how strongly it correlates with the dependent variable and the independent variable of interest, and (3) how well it is represented by the control variables that you actually measure. That is, you want to understand the control space and the degree to which your data spans it.

In our wage discrimination example, there is cause for concern across the board. Each element of the control space varies substantially, affects wages, and relates to race and gender. Each also is somewhat poorly proxied with observable measures. In the generic "wage equation," skill is represented by education, job tenure, and "potential" experience; location amenities and the cost of living by state fixed effects and city size dummies; job characteristics by industry and occupation dummies; and labor market competitiveness by dummies for union coverage and public sector employment. But potential experience is a far cry from actual experience,[4] while the dummies used to represent the remaining factors are textbook examples of data imprecision, especially for broad, "two-digit" industry and occupational dummies. Several papers have fruitfully explored these points (e.g., Light and Ureta 1995; Neal and Johnson 1996), and find that estimates of wage discrimination are quite responsive to them.

5.4 Consequences of Data Problems

When data has insufficient precision, accuracy, or span, the problems that arise are generally pretty typical—nothing you haven't seen before. But the extent of these problems depends on how the data is used. Consider a binary variable, B, utilized in a simple regression, which is mis-coded or imputed incorrectly 2% of the time. If B is the dependent variable, the consequences are probably minor—a slight embellishment of the error term in the regression, and thus of the standard errors as well. If B is a control variable instead, the ramifications will be greater, but probably still small, especially if other controls are prevalent. These other controls will partially compensate for error in B, and the estimate on the independent variable of interest

[4] Potential experience is defined as the worker's age minus the age at which they "should have graduated," given the degree that they hold. This doesn't account for delays in graduation, working while attending school, periods of unemployment or part-time employment, or the rate at which skills are acquired on the job.

should still be fine. The repercussions grow if *B* is itself that key independent variable, for errors-in-variables bias will follow.

As we know, this bias is worse when there is a lower ratio of signal to noise—when coding error more heavily obscures the true variation in *B*, or when more of this variation is accounted for by the controls, in effect "washing out" the signal. Techniques that parse the data more finely, or that try to draw "deeper" conclusions from it, tend to reduce the ratio of signal to noise, and so are more vulnerable to data problems. Thus simply adding controls to a regression exacerbates the effect of mis-coding our key independent variable *B*. It will be exacerbated further in a differences-in-differences model that includes space and time fixed effects, which identifies the coefficient through changes within cross-sectional units over time. This effectively increases the variance of the error while reducing that of the signal, inflating errors-in-variables bias dramatically. If the typical individual truly changes *B* 4% of the time, then the data will imply that *B* changes, in fact, 8% of the time: 4% true change, 2% false first period value to true second period value, and 2% true first period value to false second period value.[5] Half of the measured changes in *B* will be error, and the bias in the coefficient will be extreme.

This is pernicious enough, but it's not the only way fancier estimators are more vulnerable to data problems. There is also the issue of transparency. In complex estimators, the impact of data problems is harder to see and appreciate. These days, for example, differences-in-differences is old-school: the rage is to estimate treatment effects using inverse probability weighting. Let *B* now be a binary treatment indicator. Regressing this on a set of observed variables yields the predicted probability of treatment: a propensity score. Then, in estimating the effect of *B* on the dependent variable, observations are weighted by the degree of "surprise" in being treated. The weight is largest when you are treated when the propensity score is low, or aren't treated when the propensity score is high. When the propensity score is 0.95, for example, an untreated observation will receive $(0.95 - 0)/(1 - 0.95) = 19$ times more weight than a treated one. That's a big difference.

The logic for this is perfectly understandable. More information is imparted when a person is "unexpectedly" treated or not treated, so these observations should receive more analytical weight. But you know what is the most unexpected treatment—or lack thereof—of all? Mis-coding. The bias is also towards zero, and could be quite large when treatment is fairly predictable. But how large exactly, in theory or in practice? We don't really know. At this point, the issue has not received sufficient attention. I have watched two well-known econometricians discuss this technique at length, giving little heed to the possibility that the data could be mis-coded. On the other hand, the effect of coding error (and its close cousin, imputation) in simpler, more tractable differences-in-differences models has been more thoroughly studied (e.g., Card 1996).

[5] Not exactly, if you want to be technical, but close enough.

The best way to anticipate problems like these is not by walking through the statistical assumptions of your model or a checklist of potential biases. That is often impractical, especially with sophisticated methods that are not intuitively transparent. It starts in the wrong place anyhow. There is no excuse for not understanding your data, its strengths and weaknesses, what it records and what it overlooks. Only with this understanding are you properly prepared to design your econometric model and to think through the problems these weaknesses can cause in estimation.

Food for Thought

1. The introduction to this chapter gave two examples in which intuition trumped formal measurement, and two counterexamples that went the other way. This discussion side-stepped a book, excerpted in an earlier chapter, whose entire theme was the tension between subjective intuition and objective measurement. What book? What did it conclude?
2. Table 5.1 is one of many similar entries in the classic book, *The Philadelphia Negro: A Social Study* (1899). (I'm not sure why the totals in the table are occasionally off a bit.)
   (a) Using this table, detail the ways in which this book's author uses measurement to legitimize an overlooked people, as did DeSoto.
   (b) Who authored this book? Was it important to this author to legitimize these people?
3. In a recent cross-sectional analysis of partisan primaries in Texas (Grant 2017), I used the following county-level controls: the fractions of the population that are Anglo, black, and Hispanic; the percentage of adults with at least a high school diploma and with a college degree; the percentage of housing that is owner-occupied; median age; per capita income; the unemployment rate; mean annual rainfall; and the logarithms of the value of agricultural production, population, area,

Table 5.1 "Colored Methodist Episcopal Churches in Philadelphia, 1897" (title of the original)

Church	Members	Pastor salary	Value of church	Present debt	Current expenses	Benevolent collections
Bainbridge St.	354	$1312	$20,000	$601	$1274	$326
Frankford	70	720	1500	146	155	87
Germantown	165	828	4000	400	270	177
Haven	72	440	3400	…	277	25
Waterloo St.	31	221	800	50	25	37
Zoar	508	1270	20,000	2171	257	583
Total	1202	$4791	$49,700	$3368	$2255	$1235

the number of registered voters, the number of voters in the election being analyzed, and the number of votes received by that party's 2012 Presidential nominee.

(a) Sketch out the control space spanned by these variables. Which elements of the control space should vary the most across Texas counties? Which should vary the least?
(b) The key independent variable in this paper is the order in which each candidate is listed on the ballot. By law, this must be determined at random within each county. What, then, is the primary benefit of including controls in the analysis?
(c) If your answer to part (b) was "by accounting for more variation in the dependent variable, you get more precise estimates," then you are seeing (or thinking) like a state. Scan the paper's introduction, and then answer this question again.

4. The North American Industrial Classification System (NAICS) extends, in some cases, to six digits: a pretty fine level of detail. This gives analysts the option of controlling for industry using broader, higher-level one- or two-digit dummies, or finer, lower-level three-or-more-digit dummies. Review the NAICS online at Statistics Canada's website, and then answer the following questions.

(a) In deciding the level of detail to use in controlling for industry, one must trade off data precision against degrees of freedom and data accuracy. Expound on this tradeoff. What levels of detail are used most commonly in papers in labor economics? Industrial organization?
(b) Consider how this tradeoff might differ for industry dummies used in the following contexts: (1) to control for working conditions in a wage analysis, (2) to control for market power in a pricing analysis, and (3) to account for industry-specific systemic risk in estimating stock returns using the capital asset pricing model.

5. Sometimes the easiest way to recognize the issues that can arise with formal measurement is to compare multiple measurements of the same general thing.

(a) In addition to the formal unemployment rate, the Bureau of Labor Statistics (BLS) reports five other measures of labor underutilization. Speculate on what these other measures might be, and then check your work on the BLS website. Does the formal unemployment rate see like a state more or less than these other measures?
(b) Three sources measure the number of homicides in the U.S.: the Uniform Crime Reports, the Supplemental Homicide Reports, and the Fatal Injury Reports. Identify the main differences in the way these data are compiled, and compare the strengths and weaknesses of each measure.
(c) Consumer inflation is measured using the PCE Deflator and the CPI. Identify three important differences in how the two measures are created. Which measure do you prefer?

(d) The text analyzed a binary variable that was measured in error 2% of the time. Do the homicide measures in part (b) differ by more than 2%, or less? What about the inflation measures in part (c)? Are estimates using alternative measures of homicides or inflation likely to differ by more than 2%, or less?

References

Baker M, Benjamin D, Stanger S (1999) The highs and lows of the minimum wage effect: a time-series cross-section study of the Canadian law. J Labor Econ 17(2):318–350

Campbell DT, Ross HL (1968) The Connecticut crackdown on speeding: time-series data in quasi-experimental analysis. Law Soc Rev 3:33–53

Card D (1996) The effect of unions on the structure of wages: a longitudinal analysis. Econometrica 64(4):957–979

Card D, Krueger A (1994) Minimum wages and employment: a case study of the fast food industry in New Jersey and Pennsylvania. Am Econ Rev 84(4):772–793

Card D, Krueger A (2000) Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania: reply. Am Econ Rev 90(5):1397–1420

Chamberlain GE (2013) Predictive effects of teachers and schools on test scores, college attendance, and earnings. Proc Natl Acad Sci 110(43):17176–17182

DeSoto H (1989) The other path: the economic answer to terrorism. Basic Books, New York

DuBois WEB (1899) The Philadelphia negro: a social study. The University of Pennsylvania Press, Philadelphia

Ehrenberg R (ed) (1995) Review symposium: myth and measurement: the new economics of the minimum wage. Ind Labor Relat Rev 48(4):827–849

Gilovich T, Vallone R, Tversky A (1985) The hot hand in basketball: on the misperception of random sequences. Cogn Psychol 17(3):295–314

Gladwell M (2005) Blink: the power of thinking without thinking. Little, Brown, and Company, Boston

Grant D (2017) The ballot order effect is huge: evidence from Texas. Public Choice 172(3):421–442

Griliches Z (1990) Patent statistics as economic indicators: a survey (No. w3301). National Bureau of Economic Research

Jackson CK (2016) What do test scores miss? the importance of teacher effects on non-test score outcomes (No. w22226). National Bureau of Economic Research

Kahneman D (2011) Thinking, fast and slow. Farrar, Straus and Giroux, New York

Kane TJ (2014) Do value-added estimates identify causal effects of teachers and schools? The Brown Center Chalkboard, Brookings Institution

Kane TJ, Staiger D (2002) The promise and pitfalls of using imprecise school accountability measures. J Econ Perspect 16(4):91–114

Koretz D, Barron S, Mitchell K, Stecher B (1996) Perceived effects of the Kentucky instructional results information system. Rand Corporation, Santa Monica, CA

Light A, Ureta M (1995) Early-career work experience and gender wage differentials. J Labor Econ 13(1):121–154

Linn R, Dunbar S (1990) The nation's report card goes home: good news and bad about trends in achievement. Phi Delta Kappan 72(2):127–133

Litwin M (1995) How to measure survey reliability and validity. Sage, New York

Marshall J (2016) Trump is no mystery. There've been no surprises. https://talkingpointsmemo.com/edblog/trump-is-no-mystery-there-ve-been-no-surprises. Accessed 6 May 2016

Meehl PE (1954) Clinical versus statistical prediction: a theoretical analysis and a review of the evidence. University-of-Minnesota-Press, Minneapolis
Miller JB, Sanjurjo A (2015) Surprised by the gambler's and hot hand fallacies? A truth in the law of small numbers (No. 2627354). Social Science Research Network
Moser P (2005) How do patent laws influence innovation? Evidence from nineteenth-century world's fairs. Am Econ Rev 95(4):1214–1236
Moser P (2016) Patents and innovation in economic history (No. w21964). National Bureau of Economic Research
Neal DA, Johnson WR (1996) The role of premarket factors in black-white wage differences. J Polit Econ 104(5):869–895
Scott JC (1998) Seeing like a state: how certain schemes to improve the human condition have failed. Yale University Press, New Haven
Venkatesh SA (2008) Gang leader for a day: a young sociologist crosses the line. Allen Lane, London

Part III
Ways of Doing

Chapter 6
Theory and Models

Abstract This chapter distinguishes a theory from a model and lays out the threefold objectives of economic models. It demonstrates how to negotiate the tradeoffs involved in meeting these objectives, and specifies three instruments that can be used for this purpose. These ideas come to life in applications to attorney compensation, insurance pricing and valuation, the behavior of government bureaucracies, the size of cities, and more.

You get out some paper and a Dr. Pepper, write down a few assumptions, work out their logical implications and voilà: you have a theory specifying how changes in x should affect y. To see if this theory is supported by the evidence, you write down a basic regression model such as...

Let me stop you right there. Let's back up a second.

If you were to follow the scientific method strictly, you wouldn't model anything. You'd test the theory with an experiment. You would hold constant as much of the experimental environment as possible, and set the values of x across experimental units randomly, so that they are unrelated to any uncontrolled influences. The $x-y$ relationship would then be estimated and tested statistically in ways that we know well.

The most notable aspect of this theory-testing procedure is this: you need not care one whit about y. How it is determined in reality, that stuff in the previous chapters, is all largely superfluous. In this universe, all you really need from y is the ability to measure it. The experiment takes care of the rest.

This strikes us as odd because, outside of experimental economics, that universe is largely alien to us. Our econometric gymnastics often derive from concerns about the broader process by which y is determined. Our discussion of scale, systems, and vernacular knowledge was similarly motivated. These concerns arise because our hypotheses are tested not by experiments, but by models. In this other world, we care about y very much.

So don't think of a model merely as a serviceable bridge between theory and estimation, a mild, convenient tweak on the traditional scientific method. It is a radical departure from the scientific method (Manzi 2012), with implications for empirical

D. Grant, *Methods of Economic Research*, Springer Texts in Business and Economics, https://doi.org/10.1007/978-3-030-01734-7_6

analysis that are equally large. Most importantly, using a model instead of an experiment forces a shift in emphasis away from the hypothesis of interest toward explaining the behavior of the dependent variable. This is something entirely different.

6.1 The Nature of a Model

What is a model? The word is asked to serve too many masters, to describe everything from basic reduced-form regressions to elaborate mathematical postulates that are impervious to testing and far removed from reality (see, e.g., Boland 2014 or Morgan 2012). If the Bedouins can have more than 100 words for "camel,"[1] surely our language can clarify exactly what a "model" is and what it does. Otherwise we have an escape: without clarifying what our "model" is trying to achieve, we don't have to decide whether it has achieved it.

There was no such ambiguity when I worked for the Navy in Mississippi. We analyzed the oceans with the vintage of computer equipment used in "Pong." The model–I mean The Model, which is how we spoke of it–was a stack of Fortran code an inch thick. You submitted it to HAL in the morning and got your results back in the afternoon, page upon dot-matrixed page. There was never any question what The Model was intended to do: represent reality, first and foremost. Its use to test hypotheses or provide insights was subordinate to that overriding objective and dependent on it being achieved.

Accordingly, our definition of a model is based not on form, but purpose:

Definition of a Model: *A model is an attempt to describe observed outcomes that is intended to be taken seriously*.[2]

The description is both of the process generating the outcomes and the results of that process, the values of the dependent variable. This definition accords with the most common use of the term among professionals generally (see Gilboa 2015); it encompasses econometric models and theoretical models, while differentiating the latter from mathematical assemblies with other purposes, such as elaborating a theory or illustrating a theoretical result.

This definition, catholic about methods, finicky about results, both corners us and lets us roam free. It rules out anything so stylized that it cannot be seriously intended to describe actual outcomes, such as the fanciful economy in which Debreu proves the First and Second Welfare Theorems. It lets in anything that helps explain the behavior of the dependent variable, including theory, econometrics, and vernacular knowledge, in whatever measure the modeler deems appropriate. Thus, models may be quantitative or qualitative, theoretical or atheoretical.

---

[1] "Depending on whether the animal in question is male or female, old or young, healthy or ill, nasty or agreeable, smooth or bumpy to ride, and so forth" (Barash 1994, p. 20).

[2] As opposed to an attempt to describe observed outcomes by someone who takes themselves seriously, which is something completely different.

Distinguishing a theory from a model forces you to say what you mean and mean what you say, both in the world of abstract thought and in that of empirical reality. That shouldn't be too much to ask of scientists such as ourselves. The litmus test is uncompromising. If you say, here is my model, and I say, great, let's take it out for a spin, apply it to some data, and you object: I didn't mean for my model to be taken like that. It was meant as more of a metaphor, an elaborate example (Gilboa et al. 2014), an elegant working out of how certain factors could generate a certain outcome. Please don't test my model! I didn't intend it to be a description of outcomes that was meant to be taken seriously! Then I need only say, for crying out loud, calm down. And stop calling it a model.

In this way of thinking, theory and models have similar aims but different emphases. Both would like to elucidate cause and effect and account for observed outcomes. For theory, the latter purpose is subordinate to the former. For models, the situation is reversed. When there are significant tradeoffs between lucidity and accuracy, insight and description, tractability and utility, theory and models will diverge. They will converge when these tradeoffs are negligible. Both instances are worth illustrating.

At first glance, Baik and Kim's (2007) game-theoretic analysis of attorneys' compensation in civil cases looks like any number of theoretical papers, with lemmas, streams of equations, and insightful, plentifully-labeled diagrams. But what they do with their findings makes this more than just a theory:

> After solving for the subgame-perfect Nash equilibrium, we derive the condition under which delegation to lawyers brings both litigants higher expected payoffs than no delegation. For this to occur, the hourly fee for the defendant's lawyer should be about two to three times the hourly wages of the litigants. We then find that the equilibrium contingent-fee fraction for the plaintiff's lawyer is about one-third. Finally, in equilibrium, the defendant's lawyer works harder than the plaintiff's lawyer, and has a higher chance of winning as well. All of these findings are quite descriptive of the real world.

Baik and Kim's animating assumptions differ from those usually used to explain these compensation arrangements. Which set is superior, I don't know. But in providing a description of actual outcomes that is intended to be taken seriously, Baik and Kim have located themselves squarely on the field of battle: the facts. Their analysis is a theory—and a model, too.

Now, instead, imagine that you wish to explain an employer's short-run demand for workers in an industry that features competitive input and output markets. So you posit that the firm maximizes profits:

$$\pi = \mathrm{pQ}\left(\mathrm{L},\underline{K}\right) - \mathrm{wL} - \mathrm{r}\underline{K} \tag{6.1}$$

where all variables have the usual meanings, and K is underscored because it is fixed in the short run. The optimum amount of labor, L^*, satisfies the common first order condition:

$$p\partial Q\left(\mathrm{L}^*,\underline{K}\right)/\partial L = \mathrm{w} \tag{6.2}$$

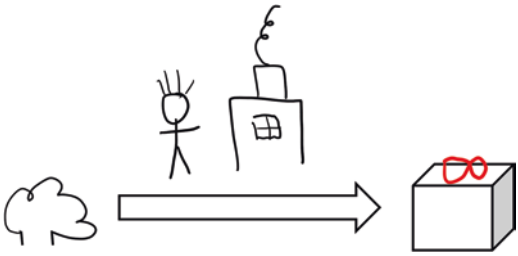


Fig. 6.1 Professor Grant's model of a firm

In other words, the employer adds labor until the rate at which it creates additional output, multiplied by the selling price, equals its wage. If you trace this out for various wages, you get a labor demand curve. This is the basic derived demand for inputs, which ties L^* to profit maximization via its marginal productivity, and which we'd be foolish not to know.

This theory is rarely claimed to be a model, however, and for good reason: as a description of observed outcomes, it is flat-out crazy. If you told a factory supervisor, a restaurant owner, or a sales manager that they should add labor until the additional output it generates, multiplied by the selling price, equals its wage, they would give you a quizzical look and then ask about raw materials or intermediate goods, which are left entirely out of the equation. The envelope theorem won't save you on this one: in these applications and many others, raw materials or intermediate goods enter in fixed proportion, the labor input must be changed in discrete amounts, or both. Either invalidates the envelope theorem.

These practical defects are remedied in the simple model with which I introduce labor demand to my students, presented in Fig. 6.1. It depicts a value added process in which raw materials or intermediate goods (the nondescript symbol) are transformed into output (the box) via the application of capital and labor. Calling the cost of the intermediate good c and the discrete change in labor ΔL, I point out that the value added by this extra labor is $(p-c)\Delta Q$, which is worth it as long as this exceeds the extra labor cost $w\Delta L$. Thus labor is added until $(p-c)\Delta Q/\Delta L$ no longer exceeds w.

This is clumsy and obtuse—the concept of profit maximization is never explicitly invoked–and so is clearly inferior as a theory, indeed hardly deserves that name. But it is far more realistic, both in fact and design. It is, despite its informality and its flaws, a model: an attempt to describe observed outcomes that is intended to be taken seriously. As such, it serves its designated purpose well.[3] The theorist Hal Varian (1989) rhapsodizes that

> A well-constructed economic model has an aesthetic appeal well-captured by the following lines from Wordsworth:
>
> Mighty is the charm
> Of these abstractions to a mind beset
> With images, and haunted by herself
> And specially delightful unto me

[3] For working MBA students, especially, the dissonance between reality and the pure derived demand theory of labor is genuinely confounding. My model solves that problem.

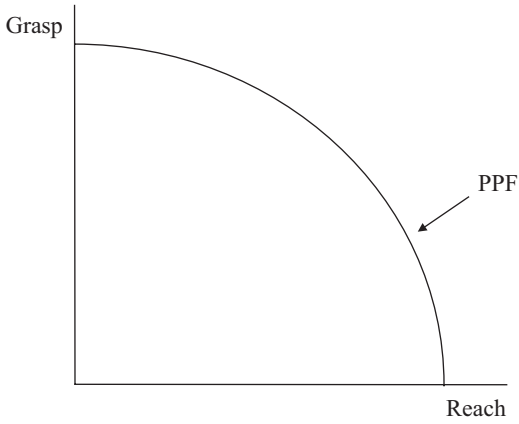


Fig. 6.2 The reach-grasp problem

> Was that clear synthesis built up aloft
> So gracefully.

Well, I just traded in all that for a diagram that turns an overgrown piece of broccoli into a Christmas present—because I'd rather be right than elegant.

6.2 The Central Conundrum of Economic Modeling

Theory's causal emphasis elevates it above the messy business of implementation: incorporating other influences on the dependent variable that lie outside the theory's scope and determining when in practice its assumptions do and do not apply. This stuff is the modeler's problem. The modeler must aspire to achieve three objectives: to draw a clean line from cause to effect, to describe observed outcomes with fidelity, and to do so across a panoply of situations, places, and times, a property I call "versatility."[4]

Do not undervalue this last quality. If you've never seen one of your precious models, which you supported so carefully in your data, sputter and flail elsewhere, well, you haven't been living hard enough. I have experienced it not just in studying election turnout, but a few times. It is humbling. One theory of wage dynamics, artfully shaped (by others) into a simple regression equation, I supported on five data sets spanning over 30 years (Grant 2003). These findings would surely live forever! On newer data, apparently not (Molloy et al. 2014).

[4] Versatility can be achieved in two ways. The model could apply directly to a broad range of circumstances: *generality*. Or the model can be easily adapted to circumstances other than those that it predicates: *generalizability* (see the discussion of Lucas' model, below).

The central conundrum of modeling is that these tripartite goals–causality, fidelity, and versatility–are often at odds with each other, a phenomenon I call the "reach/grasp problem" (Fig. 6.2). There are tradeoffs involved, and a good model dexterously manages these tradeoffs. You have three tools with which to do so: the specificity with which your model describes process and outcome, or "exactitude"; the degree to which it is detached from concrete facts and figures, or "abstraction"; and the fullness of the causal chain it embodies, or "causal depth." Each tool can be deployed to varying degrees, a fact lost upon "corner solution" modelers who always give it all or nothing. Much of the creativity in modeling lies in these intermediate options, as we will now see.

6.2.1 Exactitude

To describe observed outcomes with fidelity is to make predictions about them that are suitably accurate and precise. There is a tradeoff between these two qualities: more precise predictions are inherently more inaccurate, more likely to be wrong. This tradeoff is diminished in economics, however, because we rarely offer much precision in the first place.

Two Economists Pretend to Be Physicians

Economist 1: Doctor, have you taken the patient's temperature?
Economist 2: I have, doctor. It is 100 °F, plus or minus 10%.
Economist 1: Excellent. Proceed with the operation.

As a result we emphasize accuracy: the truth of your claims, however precise they may be. Baik and Kim, for example, were clearly satisfied with predictions that were only correct to the first order, and no one would fault them for this.

Nonetheless, adjusting a model's exactitude—the numerical or functional specificity with which it describes process and outcome–can help us make the most of this tradeoff, while improving the model's versatility as well. A model composed of large, square blocks should be more sturdy than one tenderly constructed from small, smooth stones. Crude predictions formed from crude assumptions generally apply more broadly and fail less often, as we will now see.

Ultramarathoning. It turns out that contestants running a 100 mile race are highly motivated to finish within 24 h. The psychic benefit from doing so can be regarded as a "threshold incentive" that offers a reward only when observed performance exceeds some pre-specified value.

This incentive problem is intriguing, but once I started building a model of ultramarathoners, I had to care about them too (Grant 2016). It would be natural to treat their racing strategy as the solution to an optimal control problem, in which they continuously adjusted their pace in order to minimize their finish time. My data, which tracked runners' progress throughout the race, did not rule this out. But it sure would be complicated. Fortunately, the physiology of long distance running, particulars of the race that I studied, and descriptive statistics all supported a less exact, more tractable approach: to treat runners' racing strategies as a two-stage procedure, in which you run a manageable pace for first two-thirds of the race, and

then, if feasible, speed up in the last third in order to "break" 24 h. In this framework, threshold incentives and innocuous functional form assumptions generated a relationship between the first-stage and second-stage paces that could be taken directly to the data as a regression.

These functional form assumptions needn't be strictly accurate, however, and the discontinuous, nonlinear, non-monotonic relationship they generated was unintuitive and opaque. The solution was to become even less exact, via the use of heuristics. The predicted relationship could be adequately described with five heuristics, which themselves could be derived heuristically, without relying on specific functional forms, as seen in Fig. 6.3 and explained in the note to the figure.

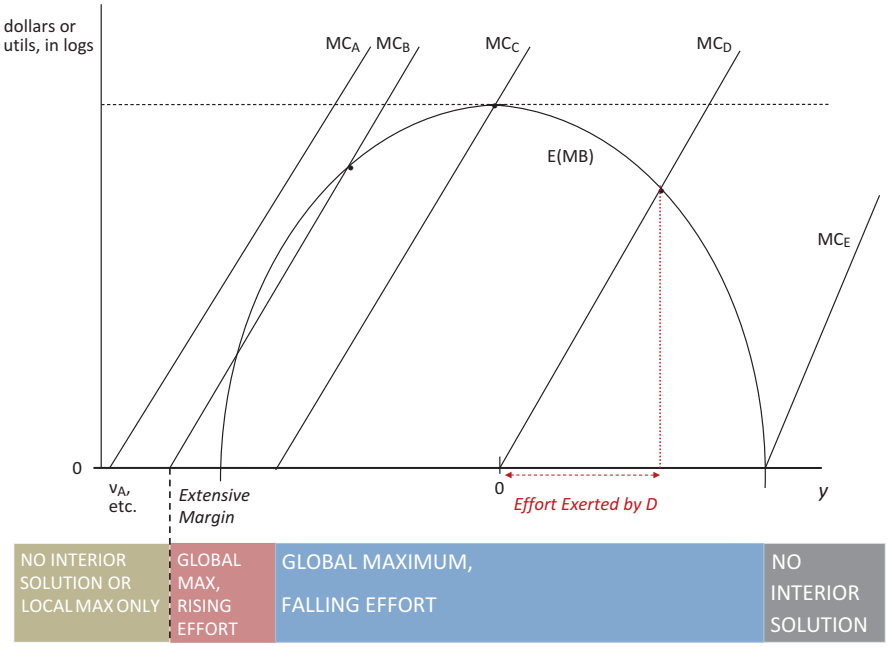


Fig. 6.3 "Derivation" of the hypothesized relationship between the first-stage pace, ν (in m/s, so larger is better), and the overall pace, y, for five different runners, A–E, with the 24 h threshold scaled to zero (from Grant 2016). Note: As runner D illustrates, threshold-induced effort can be thought of as the horizontal distance between the intercept, ν, of that runner's marginal cost curve and that curve's intersection with expected marginal benefits. Though not shown here, the model's structural parameters relate to the slope of marginal costs (MC), the width of expected marginal benefits (E(MB), which take this shape because of uncertainty in predicting one's finish time at the end of the first stage), and the height of the dashed horizontal line. The five heuristics referred to in the text all extend from this figure. Two concern those runners who give the most extra second-stage effort (runner C): (1) they are not quite on pace to break 24 h at the end of the first stage, and (2) have at least a 50% chance of succeeding (since their expected value of y is at least 0). Two concern the ν-y relation, which (3) is always positive (moving right in ν always increases y), and (4) takes a saw tooth shape (since the region of rising effort is smaller than the region of falling effort). Finally, (5) runners who are on track to pass the threshold (runner D) exert precautionary effort. Reprinted from the *Journal of Economic Behavior and Organization*, 130, pp. 180–197, 2016, with permission from Elsevier

These heuristic predictions were at a more appropriate level of precision. When the formal structural model was estimated, the parameter estimates were reasonable, but the functional form did not perfectly match the model's specification. The heuristics, which were easily satisfied, were another matter. Repeatedly blunting the edges of the theory sacrificed a bit of causality and precision in return for vastly increased accuracy, versatility, and interpretability. Not a bad deal, when you can get it.

The Value of Insurance. How much do people value the risk-sharing that insurance provides? This question is not easy to answer, especially for un-priced public insurance such as Unemployment Insurance, Medicare, or Social Security disability insurance.

That's what structural models are for, right? Cabral and Cullen (2016) aren't buying it:

> In the recent literature on private insurance, one common approach to welfare analysis is to specify a structural model of decision-making, estimate its parameters using premium and product attribute data, and from that investigate broader welfare questions of interest. The same approach could be applied to public insurance as well, utilizing data on the market for supplemental insurance. While this type of analysis allows for a broad range of counterfactuals, however, the results are typically quite sensitive to the analysis' assumptions regarding the form of utility, the distribution of risk, and individual heterogeneity.

These authors' solution is to bound the value of insurance rather than predict it exactly. This bound can be calculated using two elementary assumptions: the value of additional insurance declines in the amount of insurance people already have, and the people with the greatest need for risk sharing will buy insurance "first." In return:

> This approach allows us to remain largely agnostic as to the primitives underlying individuals' decisions, and to investigate a range of welfare questions without strong assumptions.

As in the previous example, by specifying the *process* less exactly, a measure of causality and precision was exchanged for much more accuracy and versatility. This trade was worth it.

6.2.2 Abstraction

Any theory or model abstracts, to some degree, from the particulars of the situation it analyzes. It leaves some details out. Whether this helps depends on whether these details are relevant. If they aren't, the model becomes more versatile and more easily interpreted. If they are, though, it works the other way: accuracy, causal depth, and versatility are lost.

What is the ideal degree of abstraction? Guidance in answering this question begins with the fact that theory is not the only input into a model. A model isn't just a dumbed-down theory. Its base, as we know, is vernacular knowledge, which is the opposite of abstract. This knowledge is especially valuable when outcomes depend sensitively on factors that are not easily anticipated, quantified, or articulated theoretically. My labor demand model, for example, was built on vernacular knowledge

about the nature of production: the frequency with which labor utilization is lumpy and intermediate goods are a fixed factor. Creating a model that was sufficiently versatile to accommodate this fact meant sacrificing causal depth, insight, and abstraction. So be it.

We have already seen a model that is fully abstract: Baik and Kim (2007), whose clean equations and graphs accounted for only the most essential elements of the problem. Now let's look at the other extreme: research that abjures abstraction completely.

Bureaucracy. James Q. Wilson's (1989) book of this name is not quite a model, despite its subtitle—*What Government Agencies Do and Why They Do It*. It's somewhat too broad and discursive to satisfy our definition.

But this is secondary to the central point, which is how Wilson navigates the tradeoffs between causality, fidelity, and versatility. These are confronted right at the start (p. xi):

> I wish this book could be set forth in a way that proved a simple, elegant, comprehensive theory of bureaucratic behavior. As a young and giddy scholar, I had hoped to create such a theory, and even tried my hand at a few versions. But I have come to doubt anything worth calling "organizational theory" will ever exist. Any such theories will be so abstract or general as to explain rather little. [Hint: he's talking about you, Niskanen and Tullock.] At some point, distinctions become more important than generalizations.

Because bureaucrats' circumstances are so varied, versatility is vital, and abstractness must be sacrificed to obtain it. In its place, we get a compendium of vernacular knowledge that illuminates the world of government employees, managers, and executives, the unique constraints they face and the multiple influences on their behavior, both economic and non-economic. To ensure we appreciate these influences' variety, this compendium is multifaceted: Wilson discusses prisons, schools, welfare offices, the Department of State, the Social Security Administration, the CIA, FTC, TVA, and Forest Service, OSHA, the old Civil Aeronautics Board, and NASA, among others—and that's in just the opening chapters.[5]

Here is Wilson imagining the forces arrayed against a manager impelled to speed things up at the driver's license bureau (p. 135, 136).

> Better service might require more money. Why would your political superiors give it to you? No legislator benefits from shorter lines. There may be fewer complaints, but these are episodic and ineffectual. By contrast, shorter lines at McDonald's means more customers and more money, for the restaurant and the manager who runs it.
>
> Improving service at the bureau might require replacing slow and surly workers with quick and pleasant ones. But you can neither hire nor fire them at will, and look enviously at the McDonald's manager who can. Alternatively, you may wish to mount an extensive training program to imbue a culture of service in your employees. But you would have a tough time convincing anyone this wasn't a wasteful expenditure on a frill project.

[5] CIA: Central Intelligence Agency. FTC: Federal Trade Commission. TVA: Tennessee Valley Authority. OSHA: Occupational Safety and Health Administration. NASA: National Aeronautics and Space Administration.

> If, somehow, your efforts succeed, clients will start coming to your office instead of the other, slower office down the road. As a result, the lines you succeeded in shortening will become longer again. To keep complaints down, even more must be spent on your office. But if it was hard to persuade people to do that before, it is impossible now. Why should taxpayers spend more on your office when the other one, fully staffed–naturally, no one was laid off when the clients disappeared–has no lines at all?
>
> Finally, you remember that your clients have no choice. Unlike McDonald's, you need not fear that they will take their business to Burger King or Wendy's. In the long run, all that matters is that there are not "too many" complaints to the legislature about service. Perhaps you should just relax.

This story is simple, the understanding it conveys neither abstract nor causally "deep." But it is accurate. Add dozens of other such stories, and this understanding becomes versatile, as the reader appreciates how bureaucrats' behavior flows from their variegated circumstances. The whole thing has a rationality the economist cannot help but love.

An Intermediate Case: Nonlinear Multiproduct Pricing. The theory of linear price setting is easily explicated, via the classic Lerner Index, and straightforwardly applied to data on one product or many. But this is not so for nonlinear pricing, especially when many interrelated products are involved. The pure, general theory is mind-bogglingly complex, difficult to simplify or take to data without heroic assumptions. It does not clearly justify even common pricing arrangements such as tying or bundling, despite obvious heuristic arguments for each.

Another way to put it is that the pure theory is too concrete for its own good. You get more with a less exact, more abstract model. This can be constructed by reengineering things in terms of general constructs instead of more familiar formal elements, such as demand and income elasticities. These abstract constructs, indirectly measured and informally justified, are called latent variables. In the set of firms to which my coauthor and I applied this technique, most price variation could be attributed to two such variables, "the general demand for the firm's products" and "the ability to price discriminate" (Depken and Grant 2011). Abstraction allowed us to answer questions the pure theory couldn't even ask.

In that paper latent variables played a particular role: representing heuristic concepts that defy formal development. They have other uses as well. They can be used as a data reduction technique, extracting a few major threads from a large number of variables (as we did conceptually in Chap. 5's discussion of data span). For example, psychologists don't quantify the five major personality traits–neuroticism, and I don't remember the others–by asking direct questions like "How neurotic are you?" Instead they ask a bunch of questions like "Do you always close the window when you shower?" or whatever, and distill from these the level of neuroticism and the other traits. Latent variables also can serve as a theoretical bridge to an observable end, as in binary choice models, which treat the final, binary outcome as being one if a latent variable, I^*, is positive, and zero otherwise. The independent variables don't predict the outcome–they predict I^*. Latent variables have the same function in each of these roles: to represent an underlying concept, something that is quintessentially abstract. Their popularity attests to their usefulness.

6.2.3 *Causal Depth*

Understanding causality is intrinsically valuable and practically useful. If we want to influence outcomes using policy, a causal model can recommend which levers to press, and can predict what will happen when other levers are pressed instead. Without this knowledge, we are apt to be lulled into assuming people are either overly static or overly amenable to our policy instruments.

But, as we know, there is a cost for everything–and modeling causality is no exception. One major cost is that structural models demand more to succeed, in data, econometrics, and vernacular knowledge. The other major cost is that causal models tend to be more ill-conditioned. They are more likely to fail, due to errant assumptions or changed circumstances, and they fall apart more spectacularly when they do. Imagine an outcome, y, the result of purposeful choice, governed by the flexible relation $E(y) = H(G(X,\beta))$, where $G(X)$ is a general, vector-valued function, β are parameters, and X are independent variables that could include institutional details that do not vary in the data. The function H represents untestable assertions that result from a priori theorizing, but we aspire to estimate G and β using some combination of parametric and nonparametric methods.

However, this can be done only if H is known and all X variables are observed and vary in our data. Rarely is this the case. So instead we estimate the model $E(y) = h(g(x,\beta_x))$, where h is our best guess of H, x is the subset of independent variables that are observed, and β_x is the coefficient vector associated with the subset of x variables that are non-degenerate. We cannot suppose that estimates of β_x and g will be correct, or that we can safely ignore the remaining elements of β. What follows is often more pernicious than Type I or Type II error. In-sample, the β_x and g estimates will try to compensate for any misspecification of h, making it harder to discover. But out of sample–different in time, in space, in institutional arrangements–two wrongs need not make a right. They could just pile on, one atop the other, or even reinforce each other, especially for policy analyses that contemplate a change in h. In real life what this looks like is this:

> In recent decades, a vast risk management system has evolved, combining the best insights of mathematicians and finance experts supported by major advances in computer technology. A Nobel Prize was awarded for the discovery of the pricing model that underpins much of the advance in derivatives markets. The whole intellectual edifice, however, collapsed in the summer of last year because the data inputted into the risk management models generally covered only the past two decades, a period of euphoria. Had these instead been fitted more appropriately to historic periods of stress, capital requirements would have been much higher and the financial world would be in far better shape today. (Statement of Federal Reserve Chair Alan Greenspan to the House Oversight and Government Affairs Committee, Oct. 23, 2008, quoted in Birks 2015, p. 43.)

These problems tend to be more severe when the model more fully articulates the causal process we believe to be at work, when it has a greater degree of "causal depth." Therefore, it may be preferable to decrease this down-side risk by using "softer," less sensitive, more neutral assumptions. This is what we do when we estimate the garden-variety reduced-form regression $Y = X\beta + \varepsilon$. This model forsakes

causal depth almost entirely, holding theory at arm's length, often to a sign prediction on a key coefficient. And it handicaps its versatility by contemplating only other, similarly-valued instances of X within a stable economic environment. It reaches for little, hoping to grasp it all.

At the other extreme is a model with full causal depth. This model is formed from the primordial ooze: basic principles, such as maximization; behavioral primitives, such as risk aversion; or technological fundamentals, such as returns to scale. We have seen two examples already: my ultramarathoning model and Baik and Kim (2007), which both have great causal depth, despite varying levels of abstractness and exactitude.

But you can also choose to have an intermediate level of causal depth, a partial explanation of events. We have also seen two examples of this: Cabral and Cullen, who actively advertise their ignorance of the "primitives underlying individuals' decisions," and my multiproduct pricing model, which relied on latent variables that were only "somewhat causal." Here is yet another intermediate strategy.

City Size. The right tail of the distribution of city size follows Zipf's Law: the number of cities with more than N inhabitants is inversely proportional to N. "Power laws" of this sort can be found throughout the natural sciences: a similar relation applies to the distribution of earthquake magnitudes, while the cube of an organism's metabolic rate is proportional to the fourth power of its size.

But that doesn't make them easy to explain. A model of Zipf's Law, in particular, can't just explain expected outcomes, appending an error term for convenience. It must match the actual distribution of all (right-tail) city sizes pretty closely. It is too much to form such a precise prediction with full causal depth. Gabaix (1999) pushes the ball forward by thoroughly explaining half of the problem. In classic theoretical fashion, he shows that Zipf's Law follows when cities' growth rates are randomly drawn from a common distribution, leaving open the question of how this growth process comes to be.[6]

This explanation is exact, interpretable, and versatile. When sub-areas follow Zipf's Law, it also holds in the aggregate. Inter-city variation in growth rates does not invalidate the law if it is "sufficiently structured," say as a combination of region-specific, industry-specific, and idiosyncratic shocks. And the infrequent appearance of new cities is not problematic. All this is sufficient compensation for limited causal depth.

The reach-grasp problem was portrayed as a production possibilities frontier for a reason: there many ways to negotiate the tradeoffs involved, as the varied examples in this section have shown. We see these options more clearly when we think of our ourselves not as theorists first and statisticians afterwards, but as modelers

[6] Unfortunately, Gabaix cites only one study fully confirming the growth dynamics that his model requires. His paper would have benefitted from a greater helping of vernacular knowledge, as the precise dynamics of city growth are more controversial than Gabaix assumes (see Duranton and Puga 2014). A better-structured paper would have treated these growth dynamics and Zipf's Law both as unsolved mysteries, amply documenting each, and then showing that solving the first mystery would also solve the second.

throughout. Modelers have more arrows in their quiver–tools we must deploy to negotiate these tradeoffs to maximum effect.

6.3 Setting Up Estimation and Testing

A model does not stand aside diffidently from the rest of your study. Its creation involves not just theory, but also all the preliminaries we have covered so far: a sense of scale and the system, of what is going on "on the ground" and in your data. Similarly, it has many uses once developed: estimating parameters, evaluating causal claims, and appraising policy. Altogether, your model is the fulcrum of your analysis, by which you take your understanding of the situation, informed by theory, vernacular knowledge, and data, and express from it the framework that is used for estimation, testing, interpretation, prediction, and policy experiments.

To do this properly, the model must be shaped with all of these uses in mind. This requires more than one chapter to cover fully—three, in fact, one for each incarnation of an economic model. For now, it is enough to discuss how theory and our study preliminaries can make your model better suited to all these uses.

6.3.1 Taking a Model Seriously

Our definition of a model seems to be so loose and ambiguous as to be worthless. But that has it backwards. Flip it around, make it a question, and now you have a weapon: "You take *that* seriously?" If the answer is "no," then there is a problem.

Models' flexible nature and varying exactitude permits great variety in their descriptions of the outcome. Thus, we have seen so far "sharp" predictions of functional form (Gabaix), more blunt heuristic predictions (ultramarathoning), approximations of parameter values (Baik and Kim), and crude sign predictions (the generic reduced form regression). Still, taking models seriously requires us to say what we mean and mean what we say. We must specify our model's predictions as precisely as circumstances permit. Only then have we fully exposed ourselves to the rigor of testing, which may ultimately prove our predictions wrong.

Most fundamentally, this imperative pertains to the implied magnitude of effect, as a matter of scale. If you put forward a model and say, "I have no idea whatsoever how large the effects should be," then you know I will respond: "Seriously?"

This issue is especially problematic for canonical "two types" theories, which presume people/goods to be either high cost or low cost, high quality or low quality, informed or uninformed, etc. Such theories, though mathematically tractable, rarely reveal the magnitude of the effects in question (or of the forces generating them). Without a continuity of types, the margin of interest is obscured; often the only genuine prediction is that of a coefficient sign. That's all you've got? For empirical purposes, at least, these papers would benefit by forsaking algebraic proofs for a

specific numerical example with realistic numbers and a more realistic distribution of types. Though far less elegant, this would at least indicate how large the effects of interest should be. If the estimates differ in magnitude, the theory is not supported.

To see how it should be done, let's look at Lucas' model of the welfare effects of consumption instability over the business cycle (Lucas 1987, Chap. 3; see also Lucas 2003). Consumers' utility in each period is constant relative risk aversion; they maximize the expected value of the discounted sum of these utilities over their lifetime (which is assumed to be infinite). Income uncertainty causes consumption to evolve over time according to a stochastic log-normal process with a deterministic trend and constant standard deviation.

This approach is extraordinarily tractable. The assumptions are simple, the specification of production and distribution can be set aside, and the solution is analytical. The general (each period) percentage increase in income that would be needed to compensate someone for this consumption uncertainty equals the coefficient of relative risk aversion, multiplied by the variance in consumption, multiplied by 50. Using estimates of this coefficient from the literature and the approximate standard deviation of annual detrended consumption in the U.S., Lucas shows that this percentage increase is at most 0.2%: one nice lunch per person per year, which is pretty small potatoes. Lucas proposes "taking these numbers seriously as giving the order-of-magnitude of the potential marginal social product of additional advances in business cycle theory" (p. 27).

There are good reasons to dispute this conclusion, some of which Lucas addresses, but this is not the fault of his scale analysis–it is a beneficial result. This is what models are *supposed* to do: give magnitudes to effects, to imply what the theory means quantitatively as well as qualitatively. Find fault with Lucas all you want—but give him credit for saying what he means, and meaning what he says.

Our new weapon pertains not just to a model's predictions, however, but also to its description of the process at work, which too must be taken seriously.[7] Again the concept of scale applies, here identifying the magnitudes of relevant forces along the margin of interest. Consider, for example, an environmental performance standard that specifies how much pollution firms can emit, which is satisfied via unobservable modifications of the production technology. The cost increase associated with these modifications is the "relevant force" that drives the ultimate changes in price and output. In a structural model, it would be bounded by the estimated shadow cost (Lagrange multiplier) associated with the pollution limit under the original technology. This could be compared to the price of tradeable emissions credits in places where they exist, such as Europe or California.

In addition, the imperative that a model's process must be taken seriously can serve as a practical guideline in model development. We can see its usefulness by returning

[7] Describing process may not be endemic to theory, but it is endemic to modeling. It is in the name–we model the process in order to predict the outcome. Thus, our focus on modeling elides a controversy that many of you are aware of, concerning the realism of assumptions. This controversy is not worth getting into right now–I'm not ready to mention You Know Who.

to Lucas (1987) and asking a simple question that is hiding in plain sight: how long should a period be? One quarter, the frequency with which consumption is measured in the data? One year, Lucas' implicit standard? A couple of years, the time frame over which the monetary authority can detect and stabilize movements in aggregate production? Half the length of the business cycle? Or something else entirely?

This is a puzzling question, isn't it? It is hard to know where to look for an answer. Even Lucas seems unsure just what to do. The key is to start with process. The model implicitly assumes within-period consumption is stable or can be shifted around without sacrifice. Over what time frame is this assumption reasonable? In my little town, where garage sales are held on the first weekend of the month because the correctional officers have just been paid, and where people take out payday loans to purchase school clothes for their children, even one quarter is too long. A month is more like it. A recession might mean you eat peanut butter for the last 7 days of that month, instead of the last two or three.

Now, there may be good reasons why a longer period is allowable. Perhaps the "cost" of within-period consumption shifting is small, on average, or dominated by that due to cross-period shifting. Perhaps high-frequency variation in consumption can be considered independent of low-frequency, business cycle variation, for theoretical or empirical reasons. Fine. Show this.[8] We don't have to take the prescribed process literally, just seriously. Neither Lucas nor I would be at all surprised if a longer period works out in this case. Elsewhere in macro, I'm not so sure.[9]

6.3.2 Causal Predictions

Models can have causal breadth as well as causal depth. Baik and Kim (2007) had full causal breadth because it explained everything in the system: contract types and pay rates for plaintiff's and defense attorneys, their relative effort levels and chances of winning. But other models may have less causal breadth or none at all—it too is a matter of degree.

Expanding a model's degree of causal breadth and causal depth can enhance testing in two valuable ways.

The first way, involving causal depth, is by outlining the pathways linking cause and effect. Sometimes multiple pathways are possible. Then, the failure of any one or two need not disqualify the theory: other (possibly untested) mechanisms are available. The best you can do, after presenting your main findings, is to append auxiliary

[8] If the answer is obvious, it should be easy to show. If it is not easy to show, perhaps the answer is not obvious.

[9] In my experience, macro models commonly lack empirical guidance concerning the appropriate period length, reverting to annual or quarterly periods by default. This is not justified when period length carries economic import, as in Lucas' model or a cash-in-advance monetary model that requires purchases to be financed with cash obtained in the previous period.

results that explore which mechanisms are most important. But elsewhere there is only one candidate pathway, or one that dominates. Now the number of potential hypotheses available for testing has tripled. If A leads to C because of B, you can test the reduced form relation between A and C, and the structural relations from A to B and B to C. Everything better come up sevens if your model is to succeed.

The second way, involving causal breadth, is to connect multiple predictions of the model and show they all spring from the same source.

> Theory can bring an underlying rationale that integrates otherwise diverse observations into a single whole. Thus plate tectonics explain the worldwide pattern of earthquake activity and volcanic activity, explain why continental geology differs from that of the ocean floor, show how mountains form, and account for geophysical data on the earth's interior (Smith 1982, p. 940).

Similarly, the theory that HRM practices X, Y, and Z are all complements generates seven predictions: X, Y, and Z should occur together; X should be more effective when Y is present, and vice versa; Y should be more effective when Z is present, and vice versa; and X should be more effective when Z is present, and vice versa. Were we to obtain these seven empirical results, we wouldn't view them as a jumble of unrelated findings, but as powerful evidence in support of this theory. They would form a tight conceptual unit.

The opposite of this occurs in Larson's (2015) recent look at how licensing affects teacher quality. To his credit, Larson estimates the effect not just at the mean, but across the entire quality distribution. Despite the putative belief that public school quality should be reasonably homogenous within states, it varies widely in practice. Does licensure reduce this problem or exacerbate it? A simple regression of the mean effect can't tell us. Quantile estimation can.

In significantly increasing the number of findings, however, Larson opens a Pandora's Box. Quality is measured different ways, as is the certification used for licensing. The results are broken down by teacher experience and student demographics, while unanticipated consequences–the change in the student-teacher ratio and the use of uncertified teachers–are also investigated. And all of this by quantile.

The results resemble spaghetti–findings going every which way. Some certification measures have no effect; others lower new teacher quality, especially in low-SES (socioeconomic status) schools, and do the converse for experienced teachers, especially in high-SES schools. Some certification measures lower new teachers' student test scores, and some raise them; other measures do the same things for experienced teachers. Similarly, the effects of certification on class size, the use of uncertified teachers, and teacher pay also vary by quantile, teacher experience, and the measure of certification, in seemingly haphazard ways.

How should we think about this disparate collection of results? Do they go together, or conflict with each other? Do they coalesce into a cohesive whole? If so, what is at the center of this whole? For this we need theory, of which the paper has little. Consequently, we are less convinced that these findings are correct, and more

uncertain about what is really going on. This paper amounts to less than the sum of its parts.

6.3.3 *Competing Theories*

In many provinces of applied micro, great attention is paid to parameter identification. But a theory can also be strongly identified, weakly identified, or unidentified, depending on how distinct its predictions are from those of its competitors. Models set up testing better when they heighten those distinctions, allowing more vigorous inter-theory competition.

Choi et al. (2002) face this problem in their analysis of underwriting cycles in property and liability insurance. Every 6–8 years, these markets tend to cycle from lower prices and looser underwriting standards to higher prices and higher standards. There is no shortage of explanations for this: these authors consider a total of six.

Identification of this many *theories* is bound to be a problem, so the authors adopt a systemic perspective. For each theory, they spell out how the insurer's "economic loss ratio" relates to three independent variables: interest rates, financial reserves (sort of), and the variance of claims. Even with this causal breadth, few theories are well identified in the short run. Three make identical sign predictions for all three independent variables, rendering them unidentified (Table 6.1). Another matches these predictions for two independent variables and makes no prediction for the third, leaving it weakly identified.

Table 6.1 Summary of alternative models' implications for insurers' economic loss ratio (from Choi et al. 2002)

Model	*Time frame*/independent variable					
	Short run			*Long run*		
	Interest rate	Insurer surplus	Variance of loss	Interest rate	Insurer surplus	Variance of loss
Actuarial	+	+	–	+	0	0
Capacity constraint	+	+	–	+	0	0
Economic	+	+	–	+	+	–
Financial	+	0	0	+	0	0
Financial quality	+	+	±	+	–	±
Option pricing	±	–	0	±	–	0

Note: Republished with permission of John Wiley and Sons, from Choi S, Hardigree D, Thistle PD (2002) The property/liability insurance cycle: a comparison of alternative models. South Econ J 68(3), 530–548. Permission conveyed through the Copyright Clearance Center, Inc

Here, the solution was to expand the predictions to encompass the long run as well as the short run. Now, as the table shows, four theories are strongly identified: their predictions differ in multiple ways, from each other and from those of the remaining two theories. Hypothesis testing can now separate the strong theories from the weak.

6.4 Conclusion

There is a time and a place for everything. It is appropriate for theorists to remain somewhat aloof from the subject of their study, and experimentalists too. These worlds isolate testing from theorizing. Modelers do not have this luxury. We can't remain aloof from social phenomena and describe them in a way that can be taken seriously. This is one reason we dwelled on the preliminaries in the previous chapters: in bringing us closer to the subject of our study, they prepare us for successful modeling and equip us to negotiate the modeling tradeoffs discussed here.

However, I haven't clarified how to incorporate one preliminary into model-building: our data. This, too, would never happen in the world of theory and experiment, but does in modeling—in its next, empirically-oriented incarnation.

Food for Thought

1. The Precautionary Principle holds that, in the presence of risk or uncertainty about the range of potential outcomes of an action or policy, greater weight is given to (avoiding) those that are worse. If an analyst adopts the Precautionary Principle, how does that shift the modeling tradeoffs discussed in this chapter?
2. The chapter introduced three modeling tools: exactitude, abstraction, and causal depth. Models use these tools in varying amounts. Discuss the degree to which each of the following models employs each tool, and how this combination contributes to the model's success.
   (a) Akerlof's lemons model.
   (b) The original Phillips Curve.
   (c) Donabedian's model of health care quality.
   (d) One of these three "models" is not a model as defined here. Which one? Why isn't it a model?
3. In the *2015 GNH Survey Report*, the Centre for Bhutan Studies argues that "Gross National Happiness [GNH] is more … important than Gross National Product [GNP]."
   (a) Trace the historical development of Gross National Happiness from a latent variable to a manifest variable.

(b) This survey report explicitly delineates the span of the data it collects. Compare this data span with that of the Current Population Survey in the U.S.
(c) As currently measured, which variable sees more like a state: GNH or GNP? Can a latent variable such as pre-measurement GNH help prevent you from seeing like a state?
(d) How would theories or models of GNH compare with those of GNP in terms of abstraction, causal breadth, and exactitude?

4. Many textbooks teach that, in long run equilibrium, monopolistically competitive firms operate with "excess capacity" (Bade and Parkin 2017) or "below efficient scale" (Acemoglu et al. 2015), so that average production costs exceed their minimum. As with the canonical derived demand theory of labor, however, this claim is often flat-out crazy as a description of observed outcomes. There is simply no way that Dove Soap or Cheez Whiz isn't produced at lowest average cost, and the idea that their production facilities have excess capacity is laughable.

(a) Reconcile the theory with the reality.
(b) When product demand is stable, which is more likely to occur in practice: monopolistically competitive firms that have excess capacity, or monopolistically competitive firms that operate below minimum efficient scale? Under what circumstances would these phenomena be most likely to occur? Are these circumstances satisfied by Dove Soap or Cheez Whiz?

5. Table 6.2 summarizes the 2017–2018 teacher salary schedule for the Hurst-Euless-Bedford Independent School District in the Dallas-Ft. Worth Metroplex. The Euless High School football team is famous for doing the haka before games, because many of their players are Tongan.

For schoolteachers, almost all "on the job training" qualifies as general human capital, which is portable across employers. Mincer's *Schooling, Experience, and Earnings* (1974) argues that this should generate concave earnings profiles, in which real wages increase at a decreasing rate with experience until topping out toward the end of one's career; this is empirically confirmed in the economy as a whole. Carefully review the table and answer the following questions.

(a) What is going on?

Table 6.2 Academic year salary for 2017–2018, teachers with a bachelor's degree, H-E-B school district

Years of teaching experience	Annual salary	Years of teaching experience	Annual salary
0	$55,000	25	$66,731
5	$57,102	30	$71,112
10	$58,378	35	$74,503
15	$59,562	40	$77,957
20	$62,194	45	$88,324

(b) Does this unusual earnings profile plausibly result from a lack of competition in this market? Is this question easier to answer factually or theoretically?

(c) Any model of public schoolteachers' earnings profile of must account for a major contextual factor, pensions. Gathering vernacular knowledge as appropriate, explain how Texas' teacher retirement system "bends" the shape of teachers' earnings profiles. Do you think Texas' system unique in this respect?

(d) Sketch out, in broad terms, an "illustrative model" of teachers' earnings profiles that takes into account how pensions affect the costs and benefits of working over the course of one's career.

References

Acemoglu D, Laibson D, List J (2015) Economics. Pearson, London
Bade R, Parkin M (2017) Foundations of economics. Pearson, London
Baik KH, Kim I (2007) Strategic decisions on lawyers' compensation in civil disputes. Econ Inq 45(4):854–863
Barash D (1994) Beloved enemies: our need for opponents. Prometheus Books, Amherst, NY
Birks S (2015) Rethinking economics: from analogies to the real world. Springer, New York
Boland LA (2014) The methodology of economic model building: methodology after Samuelson. Routledge, Abingdon
Cabral M, Cullen MR (2016) Estimating the value of public insurance using complementary private insurance. (No. w22583). National Bureau of Economic Research
Choi S, Hardigree D, Thistle PD (2002) The property/liability insurance cycle: a comparison of alternative models. South Econ J 68(3):530–548
Depken C II, Grant D (2011) Multiproduct pricing in major league baseball: a principal components analysis. Econ Inq 49(2):474–488
Duranton G, Puga D (2014) The growth of cities. In: Aghion P, Durlauf S (eds) Handbook of economic growth, vol 2. Elsevier, Amsterdam, pp 781–853
Gabaix X (1999) Zipf's law for cities: an explanation. Q J Econ 114(3):739–767
Gilboa I (2015) A world of models: review of Mary S. Morgan, the world in the model: how economists work and think. J Econ Methodol 22(2):235–240
Gilboa I, Postlewaite A, Samuelson L, Schmeidler D (2014) Economic models as analogies. Econ J 124(578):F513–F533
Grant D (2003) The effect of implicit contracts on the movement of wages over the business cycle: evidence from the National Longitudinal Surveys. Ind Labor Relat Rev 56(3):393–408
Grant D (2016) The essential economics of threshold-based incentives: theory, estimation, and evidence from the Western states 100. J Econ Behav Organ 130:180–197
Larson B (2015) Occupational licensing and quality: distributional and heterogeneous effects in the teaching profession. Stanford University, Stanford, CA. https://web.stanford.edu/~bjlarsen/Larsen%20(2015)%20Occupational%20licensing%20and%20quality.pdf
Lucas RE Jr (1987) Models of business cycles. Basil Blackwell, Oxford
Lucas RE Jr (2003) Macroeconomic priorities. Am Econ Rev 93(1):1–14
Manzi J (2012) Uncontrolled: the surprising payoff of trial-and-error for business, politics, and society. Basic Books, New York

Mincer J (1974) Schooling, experience, and earnings. University of Michigan Press, Ann Arbor, MI
Molloy R, Smith CL, Wozniak AK (2014) Declining migration within the US: the role of the labor market. National Bureau of Economic Research working paper
Morgan MS (2012) The world in the model: how economists work and think. Cambridge University Press, Cambridge
Smith VL (1982) Microeconomic systems as experimental science. Am Econ Rev 72(5):923–955
Varian H (1989) What use is economic theory? Manuscript. University of California, Berkeley
Wilson JQ (1989) Bureaucracy: what government agencies do and why they do it. Basic Books, New York

Chapter 7
Description

Abstract This chapter portrays data description as an integral prelude to economic modeling, a key aspect of shaping a model. It lays out three principles that exemplify effective description and presents three techniques for creating graphs and tables that adhere to these principles. These ideas come to life in applications to ultramarathoning, beer prices, and the incentive effects of letter grades.

One of the most-studied laws in health economics, .08 laws, sets a per se illegal blood alcohol concentration (BAC) of 0.08 dL/g. At or above that limit, no demonstration of driver impairment is needed to convict for drunk driving–the blood alcohol alone is enough.

Hansen (2015) wants to "understand whether punishments and sanctions are effective in reducing drunk driving," and uses .08 laws for this purpose. Using precise data on the blood alcohol levels of all tested drivers in the state of Washington and their subsequent drunk driving convictions, he structures a regression discontinuity analysis around the .08 threshold. He first must show that there is no manipulation of the running variable, BAC, so he plots its density (see Fig. 7.1). It resembles a bell curve shockingly closely, and "shows little evidence of endogenous sorting to one side [of this threshold]" (p. 1587). The analysis finds fewer future offenses for drivers just above .08, compared to those just below, from which Hansen justly concludes that convictions reduce future drunk driving.

This conclusion is noteworthy less for what it tells us about drunk driving, however, than for what it tells us about social science. While the evidence supports Hansen's conclusion, you see, it supports a less sanguine conclusion equally well: these laws do not work as intended. This conclusion, left unmentioned in the paper, comes directly from the density of BAC. The whole purpose of these laws is to reduce the amount people drink before driving. But, if so, BAC should gather, to some degree, below .08–exactly the manipulation Hansen rules out.

Altogether, this episode is an unusually pure distillation of a general principle best enunciated by dating analyst Christian Rudder (2014, p. 170):

> In social science, knowledge, like water, often takes the shape of its container.

D. Grant, *Methods of Economic Research*, Springer Texts in Business and Economics, https://doi.org/10.1007/978-3-030-01734-7_7

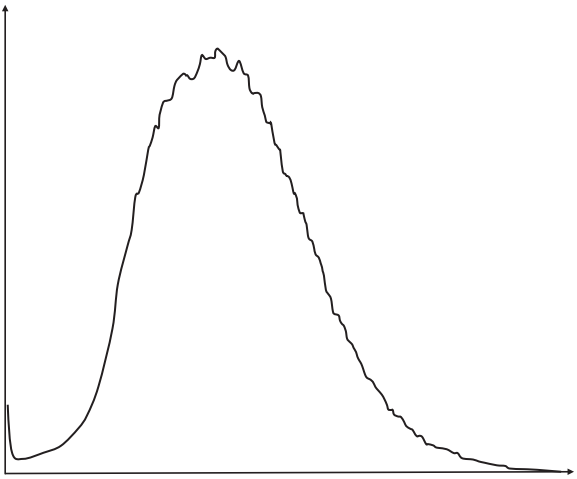


Fig. 7.1 The BAC distribution in Hansen (2015), crudely smoothed, with axis labels deliberately omitted. Can you tell where the 0.08 dL/g per se illegal BAC threshold is? Note: This figure is adapted from Hansen (2015), "Punishment and Deterrence: Evidence from Drunk Driving," *American Economic Review*, 105(4):1581–1617 (2015), Copyright American Economic Association, and reproduced with the permission of the *American Economic Review* (The BAC of 0.08 lies right above the letter "i" in the word "in" in the title to this figure)

Not always. Sometimes social science data spills its guts at the slightest provocation, like an octopus. But usually it needs an exoskeleton to give it structure. It requires a container. And it may not be an identity, but it's close enough: wrong container ≡ wrong conclusion. So before you pen yourself into the container you have designed, let the data flow freely, using description.

As most economists understand it, the purpose of description is to familiarize the reader with the data, say with a table of means or a graph of the dependent variable. This is fine, to start with. But if that's as far as it goes, then we miss description's larger purpose: to let the data speak for itself and, in doing so, shape your econometric model, the container from which you will draw the conclusions of your analysis and understand their implications. I don't care who else has used your priceless econometric method. I don't care how clever your identification strategy is. I do care about the suitability of the container for the task set before it.

This, then, is the second incarnation of a model: as the container that structures your empirical analysis. As in Chap. 6, our preliminaries–vernacular knowledge, economic theory, and a sense of scale and the system–all shape this container, and econometrics too. But it also must be infused with a good quantitative understanding of the phenomenon of interest: the best way to view it, the best way to measure it, the best way to analyze it.

This comes, in large measure, through description. In fact, you should think of the act of constructing descriptive statistics as the act of developing that understanding, which you then communicate to the reader upon completion. So go ahead, familiarize the reader with your data. But then there is more to be done. Only when it is complete will you have properly set up the estimation and testing that forms the centerpiece of your study.

7.1 Principles of Effective Description

Because descriptive statistics result from a good quantitative understanding of the phenomenon of interest, the relevant principles pertain to both the making and the telling–arriving at that understanding and relating it to the reader.

7.1.1 Self-Determination

Estimation usually employs an analytical framework that is imposed by the researcher. Then, given a set of results, we cannot readily distinguish the contribution of the ingredients (the data) from that of how they are cooked (the analytic structure). Description is different. It lets the data represent itself, without embellishment, only applying such structure as serves this purpose.

The issue here is rarely that we impose so much structure on descriptive statistics that we obfuscate their meaning. We just neglect to present many unstructured statistics to begin with, when we can leap forward to the structured ones right away. But this runs counter to the maxim that your study should be rooted in the context of the phenomenon you are analyzing. Chapter 4 discussed doing this using vernacular knowledge; now it is time to do the same thing with your data. You can't do this by imposing a bunch of structure and starting there.

7.1.2 Transparency and Redundancy

Chapter 6 argued that a model should "say what it means and mean what it says." The term for this is transparency: the full implications of the model should be plainly stated. Description, too, should promote transparency.

This principle commands us to furnish any descriptive evidence that bears witness to the legitimacy of the model's assumptions or the soundness of its estimates. To check a model's estimates, you can attempt "disproof by contradiction," and see whether they imply implausible or incorrect outcomes somewhere in the system. This may require additional information that description can provide. The most frustrating task in economics is reading a table of coefficient estimates without being told the mean and standard deviation of the dependent variable. Without this information, how can we know whether these estimates make sense?

If these consistency checks are satisfied, there is an element of redundancy: all the evidence implies the same thing. Though the term has a negative connotation, here redundancy is a virtue. Because we can be sure of so little in social science, redundancy helps affirm that we know what we think we know, and makes that knowledge more nuanced. Attorneys use the same technique in cross-examining

witnesses, repeatedly probing about the same thing, in different ways, to see if they get the same answer. The witness's credibility suffers if they don't.

7.1.3 Honoring Scale

Description should also honor scale, both in representing your data to the reader and in setting up the mechanics of your analysis. The concept applies in two ways.

The first way concerns variable values, which should be recorded in appropriate units. Often the best choice is such that a one unit change is meaningful–not too small, not too large. Sometimes this is established by convention, which is why infant mortality is reported per thousand births and air crash fatalities per million departures. And sometimes not: per capita income is often reported in dollars, when it should be in thousands (at least for wealthy economies). Do not simply default to custom or to the units in which the data are recorded. This would be to proceed in ignorance of a fundamental property of that variable.

The second way is less obvious but more important. When applicable, identify the appropriate spatial or temporal scale over which to observe key variables. Geologists rarely use the continent as the unit of observation, since the scale of geological formations is far smaller. Defining the temporal scale in hours would be equally unsuitable for evolutionary biologists, for the converse reason. Here, too, convention is not a reliable guide, as we will now see.

Industrial organization recognizes the importance of defining the market and the difficulties that can attend this exercise. A modern take on this issue involves the welfare effects of online retail via its expansion of the variety of products available to consumers. Custom, as established by several studies, has been to examine this on a national level. Does that suffice, or does it define the market too broadly? The answer depends on the degree to which tastes are local, and to which brick-and-mortar stores' offerings cater to them.

Quan (2015) investigates using simple description. With data from an large online shoe retailer, he calculates the revenue share of the top 1000 products at the city, state, regional, and national level. If local tastes are irrelevant, the product rankings should be similar at all levels, and shoes that sell well locally should also account for a large share of national sales. This is not what happens. The top thousand local products account for seven-eighths of local sales, but only one-eighth of national sales, a 7:1 difference, which Quan shows has huge effects on welfare imputations. The corresponding ratio at the state level is still lopsided, 3:1, but not at the regional level, where it is almost 1:1. The proper spatial unit for analysis is the state, at a minimum, and ideally the city.

7.1.4 Beer Prices and the Bundesliga

A particularly severe case of these principles' inaction arises in Empen and Hamilton's (2015) analysis of dispersion in beer prices associated with Bundesliga soccer games. German beer brands are regional, not national as in the U.S. So the authors identify the brands associated with each team and see if their (local) prices respond to home and away games. They do, in opposite directions, a finding the authors associate with a "tourist-natives" search model.

Empen and Hamilton provide a limited amount of description, moving fairly rapidly into estimation. Their key table of coefficient estimates is reprinted, as originally presented, in Table 7.1. There are two dependent variables, beer prices and

Table 7.1 Retail price adjustments for beer during Bundesliga game weeks (from Empen and Hamilton 2015)

	Model [I]		Model [II]	
	Beer prices	Promotion frequency	Beer prices	Promotion frequency
Constant	0.00222*** (−690.41)	−2.158 *** (−86.72)	0.00224*** (719.72)	−2.179*** (−89.76)
Time trend	0.00000034*** (−11.51)	0.000239 (1.14)	0.000000358*** (12.12)	0.000228 (1.09)
Christmas	−0.00000742 (−1.53)	0.117 (−3.7)	−0.0000133** (−2.76)	(3.98)
Easter	−0.00000433 (−0.69)	0.0258 (−0.6)	−0.00000071 (−0.11)	0.0207 (0.48)
Pentecost	−0.00000201 (−0.32)	0.266*** (7.27)	−0.0000111 (−1.77)	0.279 *** (7.68)
Father's day	−3.78E-07 (−0.06)	0.177*** (4.56)	−0.00000876 (−1.39)	0.189*** (4.9)
Temperature	−3.81E-07** (−2.65)	0.000202 (0.2)	−5.28E-07*** (−3.67)	0.000438 (0.43)
EM 2000	−0.00000316 (−0.56)	0.00393 (0.1)	−0.0000117* (−2.09)	0.0165 (0.41)
Oktoberfest	0.0000707*** (−5.84)	−0.0807 (−0.83)	0.0000809*** (6.74)	−0.0944 (−0.97)
Game-in-State	0.00000951*** (−5.33)	−0.00695 (−0.55)		
Visiting Beer			0.0000228*** (4.6)	0.0148 (0.44)
Home Beer			−0.0000649*** (−24.79)	0.101*** (5.17)
Adj. R^2	0.0367	0.0236	0.0388	0.024100
F (Chi^2)	705.31	999.47	679.57	1017.58

Note:*$p < 0.05$,**$p < 0.01$,***$p < 0.001$, t-statistics in parentheses. Republished with permission of John Wiley and Sons, from Empen J, Hamilton SF (2015) How do retailers price beer during periods of peak demand? Evidence from game weeks of the German Bundesliga. *South Econ J* 81(3), 679–696. © 2015 by the Southern Economic Association. Permission conveyed through the Copyright Clearance Center, Inc

promotion frequency, and two alternative specifications, models I and II. The controls are found in the top rows of the table, with the three key independent variables at the bottom: dummies for whether the brand's team has a game/away game/home game. In failing to communicate a good quantitative understanding of the phenomenon of interest, the paper violates all three principles of effective description. We see the consequences in this table.

- *Self-Determination*. Regressions like these are the first thing shown about the patterns in the data. But all controls are temporal except for temperature, which is largely seasonal, so these patterns should expose themselves graphically. (Doing this, using the techniques described below, is left as Food for Thought.) That is, a properly designed graph could show how prices change during game weeks, for brands at home and away, and compare this to other special occasions such as holidays.
- *Units and Scale*. Without this graph, we can only interpret the coefficient estimates in the table–which is really hard to do. There are lots of leading zeros and different numbers of total digits, even for similar variables (e.g., holiday dummies). The leading zeros indicate beer prices are measured in unnatural units, because their predicted change during game weeks is such a small number.
- *Transparency and Redundancy*. But as to what these units are, we never find out. The paper doesn't say. It's also unclear whether prices are in real or nominal terms.[1] These omissions, along with the absence of basic descriptive statistics and the graph suggested above, lack transparency and prevent us from conducting little redundancy checks that would enhance our confidence in the authors' findings.

This paper doesn't shape the container–it imposes it, and trusts that its findings are sound. Do you?

7.2 Techniques of Effective Description

Rarely can you fully achieve description's objectives with a mere table of means and graph of the trend in your dependent variable. Description, like modeling, is a creative enterprise, in both design and execution. Improvements in data accessibility and computing power have flooded the market with good examples, such as the revealing images regularly published by The Upshot in the *New York Times*.

[1] No quantitative interpretation of the coefficient estimates is offered in the text of the paper. After it discusses this table, we learn the average price of beer is 0.2207 Euros per 100 mL, with brand-specific price dispersion averaging around 0.1 Euros per 100 mL. But this can't be the unit used in Table 7.1—the coefficients are far too small. The units on temperature and the time trend are also unrecorded. The remaining variables are dummies, including the ambiguously-titled "promotion frequency."

Those are done by pros, but a few rules of thumb will take you a long way. Choose a medium that suits your objectives: tables when numbers are few, numerical detail is important, or connections don't easily fit in a two-dimensional plane; graphs otherwise. Structure these to focus on the quantities of primary interest, bringing out connections between them, and illustrating change along the important margins. Complement these design features with elements of execution that draw out key details, both large and small.

Three general techniques will help you do this effectively. To illustrate these, let's return to two papers we've seen before, and recall a central challenge faced by each. In the grades paper from Chap. 4, this was the unexpected nature of the finding that borderline students did not try harder on their final exam in order to earn a higher letter grade in the course. In the ultramarathoning paper from Chap. 6, this was the need to defend reducing a general optimal control problem (meting out exertion over a 100 mile race) into a more tractable "two-stage procedure." These challenges could not be met effectively with perfunctory description. They can be met using the techniques I will now show you.

7.2.1 Embed the Micropicture in the Macropicture

The "macropicture" concerns the major features of the landscape from which your data is drawn: variable trends, overall outcomes, etc. The "micropicture," in contrast, concerns the smaller-scale relationships that underlie the identification strategy used in estimation. Transparency requires that we become acquainted with

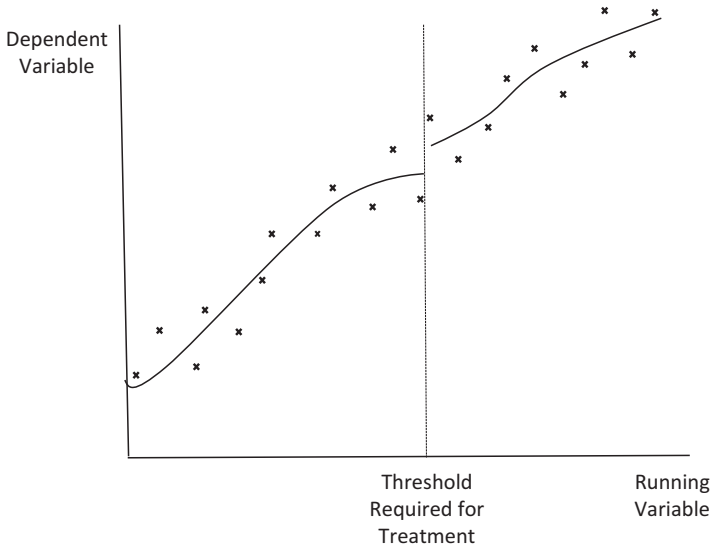


Fig. 7.2 Prototype regression discontinuity graph

both. Ideally this is done at the same time, in figures or tables that portray the latter within the context of the former, illustrating the system, giving a sense of scale, and setting up estimation all at once.

We have seen this kind of thing before. Regression discontinuity analyses, for example, often display a graph relating the outcome of interest to the running variable that determines whether the individual receives treatment (Fig. 7.2). These begin with a "bin scatter": chop up the domain of the running variable into bins, calculate the average values of x and y within each, and make a scatterplot of those. Overlaid on top of this are two trend lines, one short of the threshold required for treatment, one beyond it. These trend lines present the macropicture. Their difference at the discontinuity, which informally estimates the effect of treatment, presents the micropicture.

A more demanding example involves the ultramarathoning scenario. There, I needed to show how runners' times evolved over the course of the race, using the times recorded at the "splits" that occurred roughly every 10 miles. The conventional approach would use a sequence of histograms or kernel density plots of the time distributions at each split. But this would not illustrate the macropicture so much as a series of micropictures.

A better approach is to portray runners' split times as z-scores, standard deviations away from the split-specific mean. Plotting these, in quantiles, over the duration of the race gives you Fig. 7.3, which shows how these times evolve in a single, succinct graph.[2]

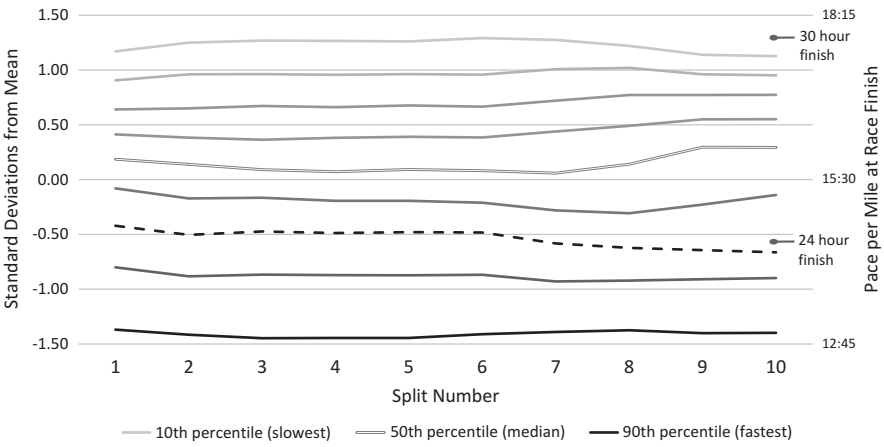


Fig. 7.3 Normalized times by decile, Western States 100 ultramarathon, for the nine intermediate race splits and the finish (associated with Grant 2016)

[2] It turns out that the z-scores in the latter splits were somewhat sensitive to the method used to handle censored observations–runners forced off the course for going too slow. As a result Fig. 7.3 was omitted from the published paper, replaced with a sequence of histograms. The censoring issue is irrelevant for the purposes of this chapter–and the approach used in Fig. 7.3 could be applied to the vast majority of races, which do not censor to any material degree.

Though it no longer commands center stage, the information in the histograms isn't lost, just repositioned, vertically, across the quantiles at each split. These manifest the somewhat-normal, left-skewed distribution of times in early splits, and their bimodal distribution later on. But now you can also follow the evolution of the race with a rightward glance of your eye. The quantiles' flatness during the first six splits, and their movement thereafter, affirms the "two-stage strategy" used to run the race; the downward thrust of the dashed quantile shows the extra effort exerted by runners who barely "break" 24 h, and compares its effect with the spread of times overall. All this places the phenomenon of interest and the racing strategy used to achieve it within a broader portrayal of the race as a whole. It embeds the micropicture in the macropicture.

7.2.2 Unlock Kinetic Potential

Early in our training we learn comparative statics. What an oxymoron! These "statics" are inherently about change: how one variable responds to activity in another. This movement may not occur in the temporal dimension, but across space or individuals or market characteristics, but it is movement nevertheless. Econometric models rely on this movement to identify the parameters of interest. Applied microeconomics is, with rare exceptions, inherently dynamic.

These dynamics are locked inside your data. Description should let them out, and realize what I call the data's "kinetic potential": the opportunity to illustrate movement, in the dependent variable or the system as a whole, with respect to time, space, or key independent variables.

Figure 7.3 was designed with exactly this purpose in mind. Here the dynamics of interest occurred temporally, over the course of the race, with runners just breaking 24 h behaving differently from infra-marginal runners far above or below that mark. The graph guides the reader through these dynamics, comparing these two groups in the process.

In the grades paper, we unlocked kinetic potential using a table. If borderline students study hard for their final exam, their final course averages should bunch slightly above the ten-point cutoffs used to distinguish A's from B's, etc. A histogram would display this bunching, if it occurs, but not the data's essential dynamics. These were conveyed with a carefully structured cross-tabulation that used the distribution of final averages as the backdrop.

This is shown in Table 7.2. Most bunching should show up in the lowest two points of any grade range (70.00–71.99, etc.), with "unbunching" in the highest two points (68.00–69.99, etc.). Thus, we simply sorted all pre-exam and post-exam course averages into three bins: the lowest two points of the grade range, the highest two points, and the six points in between. The last row and column presents the macropicture: the totals in each bin. In the center lies the micropicture: the transitions between these bins. These practically jump out at the reader. Dynamics lie literally at the heart of the table. (Soon we will see what they mean.)

Table 7.2 Transition matrix of student course averages, just before and just after taking the final exam (from Grant and Green 2013)

Post-final → / Pre-final ↓	Lower two points of range	Middle six points of range	Upper two points of range	Row totals
Lower two points of range	187 (0.27)	361 (0.52)	*150 (0.21)*	698 (0.203)
Middle six points of range	**346 (0.17)**	1380 (0.67)	**349 (0.17)**	2073 (0.604)
Upper two points of range	*154 (0.23)*	322 (0.49)	185 (0.28)	661 (0.192)
Column totals	687 (0.200)	2063 (0.601)	684 (0.199)	3434 (1.000)

Note: Reprinted by permission from Springer Nature, Grant D, Green WB. Grades as incentives. *Empir Econ*. © 2013

Always be on the lookout for ways to wheedle movement into a graph or table. A basic scatterplot allows one comparison. But add more variables to the plot by varying the markers' size, shape, and shading, and you multiply the number of comparisons dramatically. Listing the observations in a table alphabetically, say by state, isn't meaningful. Ordering them according to the values of a key variable, instead, springs the table into action, comparing that variable to every other. When techniques like these can't be used, and a given table or graph is essentially static, adding a second version (or more)–before and after, here or there, etc.–invites the reader to compare them. Where there are differences, we know something has changed. However you realize the data's kinetic potential, the reader's response will always be the same: their eyeballs flit around, seeing motion, drawing connections, developing a good quantitative understanding of the phenomenon of interest.

7.2.3 Maximize Information Transfer

The first two techniques concerned the design of your graph or table. This third emphasizes its construction, which helps the design realize its potential.

The bottom line is simple. Tables and figures transfer information from the researcher to the reader. Every aspect of their construction should facilitate this transfer. They should be a smooth, unswerving conduit of information, easy to navigate, without side-tracks, distractions, or impediments.

For tables, this requires attention to labeling, accoutrements, and content. Labels should be clear, without jargon or gibberish, and should identify the units of measurement. Thus, instead of "MTAX_HH," use "Household Marginal Tax Rate (percent)." Group related items together, and identify them as such. Thus, when listing controls, classify them by dimension: say, geographic characteristics, family demographics, economic variables, etc. This is not only more comprehensible, since the reader can take in the variables in "chunks," but also more informative, facilitating comparisons within and across dimensions and illustrating the span of the data.

Accoutrements such as gridlines, bolding, color, etc., should be used when, and only when, they convey structure or accentuate a key comparison. Thus, don't use gridlines to separate rows or columns–the rows or columns themselves do that–but to group related rows or columns together. (Unfortunately, this publisher won't always let me follow this guideline in the tables in this book.)

As to the numbers reported in the table itself, use natural units, as discussed above, and include only the degree of decimal precision that is economically meaningful and supported by the precision of the data and statistic reported. This harks back to the significant digits taught in high school chemistry. There, the accuracy of measurements of, say, weight is usually known; the number of digits recorded in that measurement should match its accuracy. Calculations using those figures are similarly constrained. Thus, the number 5.800 is a thing, if the weight is known that precisely, while a recorded weight of 4.72 may be preferred to 4.7183, if the instrument of measurement does not support that many significant digits. In the same way, it is silly to report the mean price of a ton of steel to the hundredth of a cent when its standard error is five dollars.

All of these dicta are exemplified in Table 7.2, in its organization, strategic use of gridlines, highlighting of key comparisons, clarity of labeling, and precision of numerical reporting. Table 7.1, not so much.

Similarly, graphs should have clear, restrained labeling, meaningful highlighting, and simple, clean presentation of key information. Figure 7.3 has all of this. Shading clearly orders the quantiles, while the dashed quantile for runners within "shootin' distance" of 24 h is easily compared to the others. Horizontal gridlines, useful for gauging the stability of times in the first stage and their movement thereafter, are retained but de-emphasized, so they don't compete with the quantiles. Vertical gridlines, which would be extraneous, are replaced with a simple, unobtrusive line demarcating the sixth-split shift from the first stage to the second stage of the race. Labels and legends are sparse but sufficient and easy to read, while the 24 h and 30 h finish times are marked unostentatiously, right down to the use of bulbs instead of arrows, which are more distracting.

For both Fig. 7.3 and Table 7.2, the key design feature is the identification of a common denominator in which to express the dynamics of interest: z-scores for the ultramarathon, the units digit of the course average for grades. Then you can identify a group that should respond to the incentive in question, and its counterpart, and structure a kinetic table or graph that compares these groups' behavior and outcomes. Technical features draw out and support these design features. Everything about Fig. 7.3 sustains the horizontal sweep that depicts the dynamics of the ultramarathon. Everything about Table 7.2 makes the bin transitions grab you at first sight.

7.3 Shaping the Container

Estimation and description are kin. The former descends from the latter via the imposition of assumptions. You can often think of econometric estimates as descriptive statistics that obtain after assumptions are imposed.

For example, we know that a pure difference in mean outcomes consistently estimates the effect of a binary treatment when that treatment is randomly assigned. No assumptions need be imposed at all. When this does not obtain, but all relevant controls are observed, and each treated observation has an untreated confederate with identical values of all controls, then a "pairwise difference in means" suffices. This estimator calculates a descriptive statistic, after pairing treated and untreated observations. When the data don't quite permit this, but each treated observation has untreated "near-matches" with similar values of the controls, nonparametric methods can "reconstruct" an equivalent untreated option through smoothing. You can then apply pairwise differencing to these. The estimator calculates a descriptive statistic, imposing the assumptions of matching and smoothness. If that isn't possible, but the probability of treatment can be accurately predicted by a logit model, then we have a propensity score matching estimator–still a difference in means, with even more elaborate assumptions imposed. And so on.

Thinking about econometric estimates this way carries two implications for description.

7.3.1 Continuity

After you familiarize the reader with your data, you needn't—and shouldn't—lurch to estimation. This is neither transparent nor redundant. Moving gradually, in a sequence of small steps, is both. You show how each element of the analysis contributes to its final results, and check the fidelity of the transition between adjacent steps, since only a little changes between them. Both boost credibility, if things are on track, or identify problems, if they aren't.

The principle we have just articulated, called continuity, is not limited to description. But it applies here with particular force. How better to infuse the container with a good quantitative understanding of the phenomenon of interest than by systematically segueing from unstructured to structured analysis? In this way, description can both preview formal estimation and reinforce its findings.

Our research on grades' incentive effects met resistance because our null finding was so unexpected. You can't respond to this resistance by ignoring it and letting the estimation results just drop out of the sky. To move more sequentially, Table 7.2 highlights three pieces of descriptive evidence concerning grades' incentive effects, all pointing in the same direction.

First are the fractions of final averages falling in each bin, found in the column totals at the bottom of the table. To a shockingly high degree, these match the fraction of the ten-point range occupied by each bin: 1/5, 3/5, 1/5. Next is the bin transition that would give rise to any bunching that occurs, from the top two points of one grade range to the bottom two points of another, found in the southwest corner of the transition matrix. Bunching implies asymmetry between this transition and its opposing counterpart in the northeast corner, but the transition probabilities are almost identical in each direction. Last are transitions from the middle six-point

range to the two-point bins on each end, for which the same thing is true. Nothing here suggests that borderline students try to claw their way to a higher letter grade with a strong performance on their final exam.

Thus, in the process of illuminating the dynamics driving the paper's econometric findings, this table also previews these findings. It checks them as well. Here, there was no reason to expect the formal econometrics to countermand the informal descriptive statistics–no control, for example, expected to affect dependent and independent variable alike. Therefore, had the two sets of results differed, I wouldn't have believed the more formal one–I would have been suspicious of both. Fortunately, they didn't differ. The formal analysis rendered the descriptive statistics redundant. As we now know, that's a feature, not a bug.[3]

7.3.2 Depicting and Defending the Assumptions

Transparency requires that the operative assumptions of our empirical model be not just stated but shown, when feasible, so we can gauge their plausibility. The standard of evidence–the conclusiveness given to mere description–depends on the situation. Sometimes the rule of reason applies, or we merely require the preponderance of the evidence–and sometimes we need more. Then description can be supplemented with formal testing.

In Grant (2016), I first justified ultramarathoners' two-stage racing strategy circumstantially, using the physiology of long distance running and the sudden change in race conditions between the two stages of the race. By itself, however, this isn't enough to act on. We need the quantitative reinforcement provided by Fig. 7.3. In this way, vernacular knowledge and description together shape the container used to analyze the data. (Similarly, Quan's descriptive statistics justified breaking down the unit of analysis more finely than previous studies had done.)

Though we haven't yet discussed it, the most integral assumption of any analysis is that its key independent variable has the "experimental content" claimed for it. As Chap. 8 will emphasize, this should be shown, not just stated, whenever possible, especially when this variable has a key temporal or spatial dimension over which it moves.

7.4 Conclusion

If we treat description as a perfunctory prelude to estimation and testing, then the whole exercise appears irredeemably quotidian, below our pay grade as highly-educated professionals. The problem with this perspective is that credibility is multifaceted. It depends on the whole assemblage that is your research, which also

[3] Elsewhere, of course, the two sets of results might be expected to differ in predictable ways. Suspicion follows if this does not come to pass; enhanced credibility if it does.

includes data, vernacular knowledge, and theory. Description contributes to this assemblage and connects its elements together.

I have two easy chairs at home that are 50 years old if they are a day. They have endured. Why? In part, because the pieces the chairs are made from–legs, arms, seat, back–are carefully and securely connected. This is a well-known hallmark of good craftsmanship in the construction of furniture. So it is in empirical microeconomics as well.

Don't worry. Your data will be tested all right. But so will you. How well does the model capture the process generating your data? Do your results make sense? Are their implications reasonable? What has been overlooked or left unexamined? By strengthening your analytical assemblage, good description prepares you for this test, so that the, uh, easy chair of your research is cushioned against the sallies of questioners and referees.

There is a final reason for stressing description, which you may have already surmised: the ideas developed here pertain to the rest of the assemblage as well. The same tabular techniques apply to reporting coefficient estimates, as Table 7.1 shows (by counterexample). The implications of your findings might best be displayed in a graph that embeds the micropicture–the effect of treatment–within the macropicture, or in a dynamic, kinetic table. And previous studies on your topic can be presented using a table or figure that describes the landscape of the literature, rather than that of your data. We will see tables or figures used below for these purposes, each of which manifests the principles outlined here.

Food for Thought

1. In that most basic of chemical constructs, the periodic table, what is the micropicture? The macropicture? How is the former embedded in the latter?
2. The method Quan (2015) used to examine spatial scale was relatively crude. Given a large set of products and localities, each product's local market share, and each market's size, sketch out both graphical and tabular ways of documenting the spatial scale of product preferences in this data. You may want to incorporate standard measures of "unequalness," such as the Gini coefficient, in the methods you propose.
3. The grades and ultramarathoning papers both examine the motivational effects of an achievement threshold, yet use different forms of description to depict their underlying dynamics. Design a table in the spirit of Table 7.2 that captures the dynamics of the ultramarathon discussed in this chapter. How would the essential content of the table differ from that of the graph? Why is a table inferior to the figure for displaying the dynamics of the ultramarathon, and vice versa for grades?

4. In a sizable literature on the effect of education on health, many authors take pains to instrument for education, expecting bias otherwise, as people with healthy habits should also get more schooling. But OLS and two-stage least squares estimates of the effect tend to be quite similar. Is this reassuring or not? How could descriptive statistics provide information relevant to deciding, one way or the other?

5. Consider the paired difference in means example discussed in the text, for the case of a single, discrete control variable that is associated with both the dependent variable and the treatment. Design a graph that segues from pure description to estimation.

6. This question applies the techniques developed in this chapter to Empen and Hamilton (2015).

   (a) Improve Table 7.1's organization, labeling, and content, making up any information needed to complete this task.
   (b) Design a graph to illustrate the price dynamics underlying Empen and Hamilton's findings. This graph should have more than one line, present both the macropicture and the micropicture, and be highly kinetic. It may help to know that Bundesliga "fixtures" typically alternate each team's home and away games. My favorite, of the options I generated, illustrated the mean of price changes, not that of prices per se.

7. A graph that exemplifies the principle of maximizing information transfer is the International Diabetes Center's Continuous Glucose Monitor AGP report (see agpreport.org). Identify at least five features of this graph that would make Edward Tufte proud.

References

Empen J, Hamilton SF (2015) How do retailers price beer during periods of peak demand? Evidence from game weeks of the German Bundesliga. South Econ J 81(3):679–696

Grant D (2016) The essential economics of threshold-based incentives: theory, estimation, and evidence from the Western States 100. J Econ Behav Organ 130:180–197

Grant D, Green WB (2013) Grades as incentives. Empir Econ 44(3):1563–1592

Hansen B (2015) Punishment and deterrence: evidence from drunk driving. Am Econ Rev 105(4):1581–1617

Quan TW (2015) Demand heterogeneity: implications for welfare estimates and policy. Doctoral dissertation. University of Minnesota, Minneapolis, MN

Rudder C (2014) Dataclysm: who we are (when we think no one's looking). Random House Canada, Toronto

Chapter 8
Econometric Modeling

Abstract This chapter treats an econometric model as an experiment, or set of experiments, that are embedded within an economic model. It shows how to unpack the "experimental content" of an econometric analysis, and demonstrates how this concept and the modeling principles from the preceding chapters contribute to the development of econometric models. These ideas come to life in applications to orchestra auditions, aluminum recycling, the returns to schooling, economic development in The Gambia, fracking, and more.

The "double-slit" experiment is simple by the standards of modern physics. You shine a light at a metal plate that has two parallel slits cut in it, close together. The light that passes through the slits shines on the wall behind the plate–but not in two lines. Instead, you get a sequence of light and dark bands, the classic interference pattern that comes from superimposing two waves of similar phases and frequencies. Light behaves like a wave! When this experiment was first performed, in the early 1800s, this was a bombshell discovery.

The double-slit experiment boils down to this: if you look at things a certain way, this is what you see. The same principle holds in econometric modeling. We draw causal inferences by looking at things a certain way. Data envelopment analysis, instrumental variables (IVs), multinomial logit, basic OLS—these all are our own versions of the double-slit experiment, in which we look at the data a certain way and draw causal conclusions as a result.

The last chapter introduced the concept of model as container. What do we put in this container? An experiment, or set of experiments: the third and final incarnation of an economic model. Unlike the pure experiments posited by the scientific method, however, most economic experiments are impure, governed by factors beyond our control, and imperfectly aligned with the basic regression specifications we are most familiar with. Because of this, it is vital to understand the "experimental content" of your econometric analysis: the nature of the experiments lurking within your equations.

As before, theory–in this case econometric theory–only takes you so far. We learn a lot of econometrics and generally know the assumptions of various models,

D. Grant, *Methods of Economic Research*, Springer Texts in Business and Economics, https://doi.org/10.1007/978-3-030-01734-7_8

how they can be violated, how to detect these violations, and how to address them. Ideally, applying this knowledge should be all we need to get our models right. But experience shows otherwise. A good mechanical understanding of econometrics is no substitute for understanding the experimental content of your analysis.

Furthermore, over-reliance on econometrics alone is dangerous. It encourages you to think about everything–the situation you are analyzing, the analysis itself, its potential problems and its implications–in terms of your econometric model. The model becomes the lens through which we view the situation. Then we are asking for trouble. The easiest way to do this is also the most tempting, which is to hardly look at things at all. This is what happens when we impose the container that is traditional under the circumstances, that best handles our main econometric concern, or when we nonchalantly assume our analysis has the experimental content that we have asserted for it.

In my experience, most problems with econometric models are not, in the end, econometric problems. Instead, they result from the ignorance or disregard of relevant social and institutional detail, the process generating the outcome, and the limits of the data. These are problems of perception. This is why we have spent so much time on ways of seeing: in the abstract, with theory; in reality, with vernacular knowledge; like a state, with your data. You cannot design a sound empirical analysis without understanding these things first. This is the essence of good craftsmanship. Thus, to fully appreciate this third incarnation of a model, we must flesh out this concept of experimental content.

8.1 The Experimental Content of Econometric Analyses

Two researchers wish to discover whether nourishing a certain plant with a little iodine improves its growth. They each establish two plots, some distance apart. The first researcher adds iodine to one plot but not the other, and then compares average plant height across the two plots. The second researcher adds iodine to randomly selected seedlings within each plot, instead, and then compares their average height to that of plants that did not receive iodine. For both researchers, the same number of seedlings receive iodine, and go without.

Statistically, each scenario has the same number of degrees of freedom. But that doesn't make them equivalent. Clearly they are not. In the first case, differences in growing conditions across the two plots can affect experimental outcomes, a problem that is absent in the second case. It would be better to say that the first case is a single experiment, adding iodine to one plot but not the other, observing the effects across all the plants in each plot. The second case consists of multiple, small, plant-level experiments. It has more "experimental content."

No biologist would conflate these two scenarios. They design their own experiments, and are trained in how to so do. Most applied microeconomists are not. We use observational data, and don't design anything. We can't control the experimental content of our analysis, and look enviously at the biologist who can. Sure, we

have an armamentarium of econometrics at our disposal. But we may not know how to command it, or even that we need to, until we first understand the experimental content of our analysis.

We see this in the history of analyses that meld state-level policy changes with individual-level outcomes. Two common problems occur in these analyses, elementary error structure issues comprehensible with the econometric knowledge of the 1950s. When regressions are conducted at the state level, with the dependent variable an average of individual outcomes, there is the weighting issue discussed in Chap. 3. For individual-level regressions, instead, there is "clustering," which makes OLS inefficient and its standard errors too small.

Yet both of these problems were almost wholly disregarded until a few decades ago, with Moulton (1990) and Dickens (1990), and then still often neglected until more recently, with the advent of the "clustering" option in Stata and the working paper version of Solon et al. (2015). We've waited half a century for practice to catch up to common sense.

The reason this is common sense is that no econometrics whatsoever is needed to recognize these problems' significance. You only need understand experimental content. Ignoring these problems is like being in the position of the first researcher, while pretending you are in the position of the second. (The analogy is perfect: states are plots, people are seedlings.)

Consider an analysis of this type, involving a policy adopted in West Bengal and Kashmir, among other states. West Bengal is seven times as populous as Kashmir, but to give it seven times the weight, in a state-level analysis, is like saying that it conducts seven natural experiments to Kashmir's one, which is obviously wrong. It is one natural experiment in each state, in which the mean outcomes are "recorded" to different degrees of accuracy. West Bengal should get more weight than Kashmir, but not seven times more. Similarly, in an individual-level analysis, unclustered standard errors imply that each person in each state is subjected to their own individual-level experiment, independent of everyone else. This is also obviously wrong, overly strong, yielding standard errors that are too small.

Most microeconomists are aware of these issues by now. But this is low-hanging fruit. Deciphering an analysis' experimental content is often difficult. Then shaping the container requires a more thorough understanding of the forces underlying your natural experiment, which in turn requires description and vernacular knowledge, as we will now see.

8.1.1 The Experimental Unit

Experimental content is a function of your research design and your data. It gauges, more in general terms than specifics, how many independent experiments your study consists of, the nature of those experiments, and their "strength." This, in turn, can be understood by examining the behavior of your key independent variable, X, and your error term, ε, at the level of the "experimental unit": the unit over which your natural experiment takes place.

This is easy to figure out in the iodine example. For the first researcher, the experimental unit was the plot; for the second, it was the plant. In practice, however, it need not be so simple. The experimental unit can differ from the unit of observation in the data and the units utilized by a theoretical model. This is one reason it's so important to attend to units in the first place.

Jaimovich (2013) explores economic development in The Gambia, using a household survey of rural villages. Remoteness, poverty, and social structure limit the number of households with economic links outside the village–most transactions take the form of intra-village "gift exchange" of labor, land, etc. Is such gift exchange less common when inter-village economic links are present? If so, is the relationship causal?

These questions span economic development, labor economics, and social network theory, and permit estimation approaches related to each. Accordingly, there are three possible experimental units. The first is the village, of which the sample contains only 60. Villages in close proximity to other villages could have more external connections. The second is the household, of which the sample contains about 2000. Households with more facility for making external connections could reduce the number of internal exchanges in response. The third is the "dyad," which indicates whether there is an economic connection between any two households within the same village. These connections could depend on household-level "homophily." If gift exchange is more likely between dissimilar households with different endowments, then households might go outside the village only when suitable internal partners cannot be found. There are over 100,000 dyads.

Across these three viewpoints are orders-of-magnitude differences in the number of experiments and in the implied mechanism driving the mix of internal and external economic connections. Deciding which to use requires an understanding of whether and how these mechanisms operate in practice, which in turn requires vernacular knowledge. You cannot finesse the issue by trying every type of analysis–village level, household level, dyad level–and seeing if the results are the same. Because the mechanisms differ, there is no reason that this should come to pass. It would be odd, in fact, if it did.

In that example, uncertainty about the experimental unit was generated by ambiguity in the identification strategy. No such ambiguity exists in Angrist and Krueger's (1991) analysis of the returns to schooling. If you leave school once you reach the "dropout age," your birthday determines how much schooling you get.[1] If that's all (of consequence) it determines, and birthdays are as good as randomly set, then you can compare birthday-driven changes in mean schooling with birthday-driven changes in mean earnings, and discern the causal effect of (forced) schooling on pay.

The U.S. Census data used by the authors reports births in 3 month intervals, so the natural experimental unit is the birth quarter: 1930:1, 1930:2, etc. Identification relies on the seasonality of schooling and earnings against the broader trends in both over the sample period. This–and everything else needed to set up the paper–naturally lends itself to description.

[1] Except in New Zealand, where everyone starts school on their fifth birthday, and thus has the same number of years of required schooling.

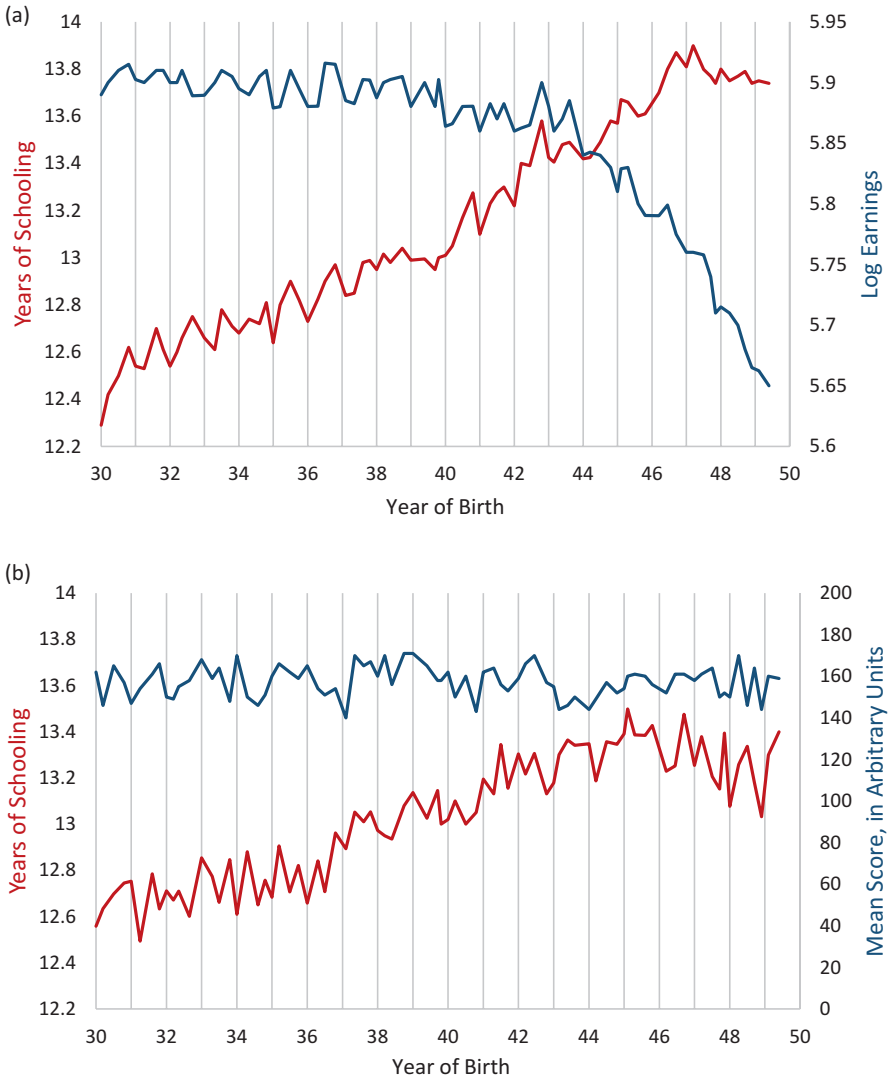


Fig. 8.1 The heart of Angrist and Krueger (1991), visualized. (**a**) Actual means of education (rising line) and weekly earnings (falling line), males, U.S. Census, by quarters. (**b**) Hypothetical means of parental schooling (bottom line) and eighth grade standardized test scores (top line). Note: The vertical gridlines identify the first quarter of each year. Graph a is adapted from J. Angrist and A. Krueger, Does compulsory school attendance affect schooling and earnings? The *Quarterly Journal of Economics*, 106(4), 979–1014, 1991, used and adapted by permission of the Oxford University Press

One possibility is presented in Fig. 8.1, which combines data from their paper with pretend data I made up. The upper graph demonstrates the coincident seasonality of average schooling and average wages. Both reach their detrended nadir in the first quarter of the year, and rise thereafter. The absence of seasonality in the lower, pretend graph supports the identification assumptions needed to pull this thing off: the randomness of birth dates, via the means of parental schooling, and the irrelevance of these dates for other outcomes, via the means of eighth grade standardized test scores.

This is, or would be, the paper in a nutshell. Who cares exactly how many individual observations you have? The natural experiment takes place over birth quarters. So red flags go up when one of the authors fires back at some critics in *Mostly Harmless Econometrics* (Angrist and Pischke 2009, p. 213).

> Bound et al. (1995) claim small-sample bias is a major concern when using quarter of birth as an instrument for schooling, even though our sample size exceeds 300,000. ("Small sample" is clearly relative.)

Oh snap. But earlier the book shows that this bias derives from imprecision in estimating the first-stage relationship. Why can't this imprecision happen at the birth-quarter level? We have no reason to rule this out, and, Fig. 8.1a suggests, good reason to rule it in.[2] Here, at the level of the experimental unit, their full 30-year sample contains only 120 "observations"–not enough to ensure that small-sample bias is irrelevant. (Of course, it would have helped for Bound et al. to have made this point themselves, as it makes their claim of large finite-sample biases despite "enormous sample sizes" less incongruous.)

Even when the data permits a specification that is more fine-grained, I am inclined to specify the econometric model at the level of the experimental unit.[3] Doing so aligns the unit of analysis with the experimental unit, ensures appropriately-sized standard errors, and reveals econometric problems most

[2] We aren't given the sampling variation in the birth-quarter means of schooling or earnings, which would help us figure this out more definitively. But a careful look at Fig. 8.1a indicates it is small, a consequence of the large number of individual observations. While the non-seasonal wiggle is sizeable, little of it appears to be *random*. Thus most of this wiggle comes from unobserved birth-quarter-level influences, not sampling error.

[3] When state-level policies are analyzed using micro data, a two-step procedure can accomplish this goal. The first step regresses individual-level outcomes on a set of dummies for each experimental unit (indexed with e) and individual-level controls (indexed with i):

$$Y_{i,e} = \delta_e + \lambda Z_i + \varepsilon_{i,e}$$

The second step regresses the estimates of these dummies on the treatment variable, T, and experimental-unit level controls, X:

$$\hat{\delta}_e = \alpha + \beta T_e + \gamma X_e + \xi_e$$

straightforwardly, through non-iid residuals, specification issues, etc. I know this is not the universal standard. But then the fallback position should be to include a random effect in the model at the level of the experimental unit, which (since it's a model) is taken seriously. That term, and the experimental-level uncertainty and unobserved influences that it represents, puts the "natural" in "natural experiment." Including it would have addressed most of the econometric problems this chapter has raised so far.

For example, when analyzing state-level policies using individual-level data, you needn't limit yourself to the two options considered above: a state-level analysis with appropriate weights or an individual-level analysis with clustered standard errors. Instead, you could simply insert a state-level random effect into an individual-level specification, like this:

$$y_{\mathrm{i,s,t}} = \beta X_{s,t} + \gamma Z_i + \mu_{\mathrm{s}} + \tau_{\mathrm{t}} + \xi_{\mathrm{s,t}} + \nu_{\mathrm{i}} \tag{8.1}$$

where X contains all experimental-level variables, including the policy of interest, and Z includes individual controls; ξ is the experimental-level random effect and ν the individual-level error term. Estimating this model puts the proper "weights" on X, properly "inflates" the standard errors on $\hat{\beta}$, and improves efficiency to boot.

8.1.2 The Structure of **X**

Once the experimental unit has been identified, it's time to examine how X varies across these units, and why. Regression theory does not directly require this: as long as X, or its instruments, can be considered independent of the error term, etc., etc., you're fine. But how can you really know, in the absence of these fundamental facts about your key independent variable? What kind of natural experiment are you running, anyway?

The "why" is important because good parameter identification is based on a clear understanding of the processes driving X. When this is lacking, the paper becomes somewhat unmoored. We are unsure how much credence to give Jaimovich's (2013) household-level and dyad-level estimates, because we don't know the forces shaping the number of external links in the first place. In Angrist and Krueger (1991) we are left hoping, more on faith than facts, that birthdays are as random as the estimation framework assumes they are.

The "how" is important because experimental content depends on how X "moves" within your sample. It may vary on temporal or spatial scales that do not respect the Gregorian calendar or provincial borders, and independent variables that evolve on longer scales yield less experimental content. Return to Fig. 2.1, which mapped the states that had adopted the Child Rights Act by 2007, 4 years after its approval by Nigeria's federal government. It lulls you into thinking of a change in the law in the usual way: occurring at the state level over the course of a year. Within a few years, however, all but one of the southern states had adopted the Act, while only one of

the northern states had. One *could* think of the Act as being adopted regionally, over a time scale of a decade. One *should* do so if the other parts of the system–the law, religious and social custom, labor force participation–also evolve on these spatial and temporal scales. Taking annual, state-level snapshots of this slow, broad evolution does not multiply the number of independent experiments.

While important, the scale of motion is not the only way that the dynamics of X affect experimental content. A striking example is Gruber et al.'s (1999) study of the effect of changes in the "fee differential" paid by Medicaid when a physician delivers a baby by cesarean section. Their data has nearly a million observations, if you read it all in, but we know better than to think too much about that. The experimental unit is the state*year cell, and Table 8.1 shows that there are only 42 of those. Using this data and standard panel regression methods, the authors find that increases in this fee differential greatly increase the use of cesarean delivery.

This table reveals how X varies: not smoothly, like El Nino, but abruptly, like a tsunami washing ashore. When the "experimental power" of X is concentrated like this, there is less experimental content, and more danger of something going awry.

Since the model contains state fixed effects, let's look at within-state changes in the differential from one year to the next (Table 8.2). Few of these changes exceed even $100, and the largest comes from the most populous state in the union, California. In fact, over three-quarters of the weighted "variance" in ΔX comes from one change in fees alone.[4] That's not much on which to base a statistical analysis. At a minimum, this should cause you to wonder whether the results are sensitive to the exclusion of the data from California in 1988. The answer, I eventually learned, is yes. The coefficient switches signs, and is "almost significant."

8.1.3 The Structure of ε

Your natural experiment is stochastic, not deterministic. Thus ε is as much a part of it as X and Y are, and as integral to understanding experimental content. After all, much of the previous discussion would be irrelevant if ε was truly random and iid at the level of the experimental unit. Large reductions in experimental content and serious econometric issues are caused by the interplay of X with ε *at this level, on the scale of motion of* X. Therefore, after thinking about the structure of X, you must relate it to the structure of ε.

The most obvious way this interplay occurs is that X, ε, and Y form a system. In Jaimovich (2013), internal connections, external connections, and household economics jointly extend from properties of the village and its inhabitants, making it hard to establish causality without more supporting information. Similarly, in the Child Rights Act, law, religious belief, poverty, and child outcomes such as education

[4] Clearly, this isn't precise or "properly done." It isn't trying to be. It's a ballpark estimate that you can approximate in your head, and get to the same place that more ponderous formal calculations would get you.

Table 8.1 The Medicaid "fee differential" in Gruber et al. (1999), as presented in the original paper, with minor adaptations

Region State	1988	1989	1990	1991	1992
East					
New Jersey	0	109	130	130	130
Pennsylvania			147	146	146
Midwest					
Iowa	254	254	264	273	273
Illinois	75	75	150	150	150
Wisconsin		180	180	205	192
South					
Florida	0	0	0	0	0
West					
California	306	62	0	0	0
Colorado	72	236	413	440	428
Washington	62	62	71	201	201

Note: state*year cells not used in the analysis are left blank. All values are rounded to the nearest dollar. Reprinted with minor edits from the *Journal of Health Economics*, 18(4):473–490, 1999, with permission from Elsevier

Table 8.2 An alternative view of the data in Table 8.1, differenced and placed in descending order

State, Change in years	Change in the fee differential	Previous value squared, as a fraction of the total sum of squares	Previous value, population-weighted
California, 1988–1989	−244	0.38	0.78
Colorado, 1989–1990	177	0.20	0.04
Colorado, 1988–1989	164	0.17	0.04
Washington, 1990–1991	130	0.11	0.03
New Jersey, 1988–1989	109	0.08	0.04
Illinois, 1989–1990	75	0.04	0.03
California, 1989–1990	−62	0.02	0.05
Ten other state*year combinations	under $30 in magnitude	0.00	0.00
Sixteen other state*year combinations	0	0.00	0.00

and marriage all form a system that varies regionally. Systems thinking helps you recognize this at the outset.

When there is not a system, and this interplay occurs more or less incidentally, the best way to unmask it is by thinking about ε on the scale of motion of X. Take fracking for example. This major economic event has spawned studies examining its effect on everything from mortgage foreclosures to risky behaviors and crime to

employment and income (James and Smith 2017; Feyrer et al. 2017; Maniloff and Mastromonaco 2015; Paredes et al. 2015; Munasib and Rickman 2015; McCollum and Upton 2018).[5]

Fracking goes after oil and natural gas formed from deposits of little creatures in shallow inland seas millions of years ago. Understanding this geologic process is a piece of cake, as this extract from Wikipedia makes clear:

> Early in the Acadian orogeny, the shales of the Hamilton Group began accumulating as erosion of the mountains deposited terrigenous sediments from the land into the sea. The Marcellus Shale in Pennsylvania was formed from the very first deposits in a relatively deep, sediment- and oxygen-starved anoxic trough that formed parallel to the mountain chain.

By definition, these inland seas were in what some people call "flyover country." Here, in Oklahoma, North Dakota, Western Pennsylvania, are the oil- and gas-rich shales to which fracking holds the key. From this perspective, the key spatial dimension of interest isn't political, i.e., state or county borders, but geological: the coasts vs. the heartland.[6]

So, before proceeding with the formalities, ask yourself: was anything else going on during the fracking boom that operated differently in the heartland and on the coasts? Absolutely. This boom ran from about 2007, when the technology became viable, to 2014, when oil prices sank into their own anoxic trough. This period coincides with the bust of the U.S. housing market, which, as geology and politics would both have it, was greatest on the coasts.[7] Estimating the effect of fracking requires accounting for the housing crash. This is true whether your model uses differences-in-differences, propensity score matching, synthetic controls, or shale plays as instruments, all of which can be found in the papers listed above.

In this way, understanding the experimental content of your analysis identifies problems that a basic regression specification would probably overlook. Here is a somewhat awkward model representing the issue above:

$$Y_{c,t} = \alpha + \beta F_{c,t} + \gamma X_{c,t} + \delta Z_{s,t} + \sigma_c + \tau_t + \varepsilon_{c,t} + \xi_{s,t} \tag{8.2}$$

where Y is the outcome of interest, F is a binary fracking indicator, and X and Z are vectors of controls. Where most specifications have one spatial index, this has two, each with its own error term: c for counties, the standard unit of analysis in these

[5] This list includes only studies in economics that utilize nationwide data to test their hypotheses of interest. Some other studies use data on only a handful of states, and are less subject to the criticisms that follow (as is Feyrer et al.'s Table A13, a supplementary regression that includes state*year fixed effects). All of these papers were scouted in 2016, and those that were working papers at the time were followed up on later.

[6] If you think it's crazy to imagine that geology influences behavior, think again. For instance, county-level "heat maps" of numerous social phenomena outline the shape of Appalachia.

[7] The geological part is the land scarcity on both coasts, for different reasons. The political part is zoning and laws governing mortgage refinancing.

studies, with error term ε; and a binary index *s* for counties formerly covered by a shallow inland sea, with error term ξ. OK, but–who would write this model down out of the clear blue sky? The issue is recognizing that the scale of *X* correlates with the scale of several factors that influence *Y*. An equation can't do this for you. Standard regression diagnostics won't do this for you–they rarely detect the spatial correlation implied by this error structure, especially on this scale. Standard economic controls won't automatically do this for you–compared to the typical recession, the housing crash had unique features that can merit unique controls. These aren't merely theoretical statements: of the six fracking studies listed above, only two even mention the housing crash, without accounting for it in estimation.

Now we see how an understanding of experimental content complements econometric skill: not by uncovering mystical problems unknown to statistics, but by identifying problems that you already have the tools to solve. While it has not always been spelled out (see the Food for Thought), this holds true for all the examples in this section, representing a variety of econometric problems.

In addition, experimental content helps ensure the integrity of hypothesis testing. The surest way to artificially lower your standard errors is to pretend a study has more experimental content than it actually does. Both of these concerns play a role in model-building, as we will now see.

8.2 Building Econometric Models

An econometric model is a model, with all the rights and responsibilities thereunto appertaining. In its construction, we can and should employ all the modeling concepts we have developed so far: the definition of a model, the reach-grasp problem, the model as container and experiment. In the course of doing so, we will develop an overarching principle that unites these concepts and rules them all.

8.2.1 Describing Outcome and Process

Our definition of a model commands us to provide a serious description of the outcome of interest and the process generating it. After all, if you can't take the model seriously, why should you take its estimates seriously? Why should you take its predictions seriously? This principle applies to all aspects of an econometric model: the deterministic component, the error structure, and the estimator.

The Deterministic Component. As in Chap. 6, econometric models can have varying degrees of theoretical content. Some are fully structural, some are completely ad hoc, and some lie in between, with a theoretical core that is augmented with practical embellishments.

In this last case, continuity requires that your econometric model extend from your theoretical model to the degree possible. This is common enough in the litera-

ture that I can restrict myself to a brief example: a demand study, discussed in a later chapter, whose core price-consumption relationship comes directly from theory. Since the authors analyze state panel data, their regression includes state and year fixed effects, along with an obvious demand shifter, income.

Simply appending these terms onto the specification would violate the principle of continuity. Instead, the authors incorporate them systematically. They build unobserved "utility shifters" directly into their theory. Spatial and temporal variation in these shifters justifies the fixed effects, while the inclusion of income is justified by showing that it accounts for changes in another latent variable, the marginal utility of wealth. In making this systematic transition from theoretical model to econometric model, the authors discover that the utility shifters enter in current and future terms, which implies that the regression's error term is autocorrelated. This is news they can and do use.

When a regression specification is ad hoc, with no theoretical content at all, there is no continuity to be had. Nonetheless, this specification still must meet the same litmus test: that it can be taken seriously. Being ad hoc does not exempt it from this requirement, as we will see.

The Error Structure. In my econometrics classes, years ago, what I took away was that regressions had an error term, whose properties–autocorrelation, heteroskedasticity, non-normality, etc.–affected the properties of the estimates, possibly necessitating a change of estimator. I was so focused on the coefficient of interest that the nature of the error term interested me only insofar as it affected the estimate of that coefficient.

I was wrong to think that way. For any model, job #1 is describing the phenomenon of interest. This applies not just to the model's deterministic component, to things you can see, but also to its stochastic component, to things that–like dark matter–you can't see, though you know they are there. The error term isn't just something you throw on at the end. It is part of describing the phenomenon of interest. Its structure should be modeled no less carefully than anything else. This structure then helps you find and fix your econometric problems.

This was a major theme of our discussion of experimental content. It paid dividends in the fracking example, where spatial correlation was implied by the presence of error terms on two spatial scales: the county, the unit of observation; and the region, the experimental unit. Theory can also help you unpack the error structure, as in the demand study discussed above, where it implied serial correlation. So can the characteristics of the data, as in Chap. 3's discussion of weighting, where our scale analysis relied on a specification that had a homoskedastic term for specification error (ξ) and a separate, heteroskedastic term for sampling error ($\bar{\nu}$):

$$\bar{y}_{s,t} = \alpha + \beta X_{s,t} + \xi_{s,t} + \bar{\nu}_{s,t} \tag{8.3}$$

Be an omnivore: anything that helps you flesh out the error structure is worth using. Remember, even when a simplistic error structure doesn't bias the coefficient estimates, it can bias the standard errors.

Estimator. The deterministic and stochastic components of a model often decide what the estimator will be. When they don't, our definition of a model can provide useful guidance. We see this in the confession of Hamermesh (2000):

> In 1994 a journal accepted an article of mine, under the proviso that I replace my ordered-probit model of educational attainment, which the data classified into intervals, with least squares, in order to aid readers' comprehension. I complied with the request. I was wrong. Doing so represented a step backward from the frontier of knowledge, was unnecessary in light of readers' comfort with the technique, and detracted from the story that the article had to tell.

From this experience Hamermesh argues that fancier techniques should be used when (but only when) they are worthwhile, and I would agree. But there is a deeper moral to the story. The dependent variable, education, was measured in numerically awkward intervals (<12, 12, 13–15, ≥16 years), such that a one unit change has no natural interpretation. Under these circumstances, ordinary least squares is something of a misfit, sufficiently awkward that it squares poorly with our definition of a model. The ordered probit is more natural and, therefore, superior.

8.2.2 The Reach-Grasp Problem

The reach-grasp problem applies to econometric models, too. This includes everything from Chap. 6, which shapes an econometric specification with theoretical content, and extends to estimation. A graceful equation is only as good as the accuracy and integrity of the results that come from it.

This fact underlies our appreciation of statistical power, degrees of freedom, the strength of parameter identification, and the properties of our data. To these items we add another: experimental content, which should be adequate to answer the question posed. We have all seen plenty of ornate empirical studies that pile a crushing weight of theory and equations atop a thin wisp of data, such as the structural insurance models criticized by Cabral and Cullen (2016) in Chap. 6. A good understanding of experimental content can prevent studies like these from ever getting off the ground.

The reach-grasp problem is not limited to these types of studies, however. It applies to simpler stuff as well, such as Gruber et al. (1999), which had such concentrated experimental power that it probably shouldn't have been conducted in the first place. Similarly, if the system in the Child Rights Act evolves regionally over long time scales, it will be hard to accurately estimate the Act's effect on child outcomes.

There is no simple formula for handling the reach-grasp problem, and inevitably you will make mistakes, over-estimating or under-estimating the power of your statistical analysis. The best way to equip yourself to address this problem is to develop a solid appreciation of all the factors that underlie it–another reason we have spent so much time on them.

8.2.3 Building the Container

Finally, we should be cognizant of an econometric model's incarnation as a container, and build it with a light touch.

The first reason to do so was stressed in Chap. 7: a container built ground-up is preferred to one imposed from on high. Figure 8.1 showed how Angrist and Krueger's (1991) IV strategy could be supported through description, at least putatively; the data needed to implement this approach, though not plentiful, existed at the time. But, instead of supporting their instruments' independence with a graph like Fig. 8.1b, the authors imposed this assumption first, and (briefly) defended it only after the estimates were in hand. In doing so, the virtues of self-determination, transparency, and continuity are lost. These virtues are worth pursuing, even when you have no reason to think you are wrong.

The second reason to use a light touch is that structure that is imposed often isn't tested. The classic example involves estimating the effect of a change in policy on some outcome, Y. Often this is done using the coefficient on a dummy variable, D, which is set to one when the new policy is in effect and zero otherwise. No matter the specifics, however, a standard parametric analysis still imposes something: timing, the "restriction" that Y changes when D does, ceteris paribus. This could be wrong.

We recognize this nowadays, so researchers often supplement such specifications with semi-parametric "event studies" that do not force this coincidence in the regression specification. These examine Y's behavior throughout (much of) the sample period, to see if its changes synchronize with those in D. This is good. But the fact that this approach has caught on only recently, despite its simplicity, testifies to the ease with which we impose structure without testing it.

Finally, a light touch serves as a counterweight to our natural inclination towards complexity and opaqueness. As in Chap. 6, better models need not be more complex. Prudence, and a light touch, might win out in the end. For example, Ashley and Parmeter (2015) and others have emphasized that the basic diagnostics that are readily available for simple regressions can often tell you when something is wrong.

In my recycling research, mentioned once before, I estimated a secondary aluminum supply curve from just 18 annual observations, from 1923 to 1940. With so few observations, you hardly needed regression diagnostics–you could just look directly at the residuals. The last two were unusually large. What to do? Some reading dredged up the fact that "scrap drives" were held during these years, 1939 and 1940, in anticipation of a military buildup. This temporary phenomenon didn't need to be explicitly modeled. I left things as they were. Large, positive residuals were, in fact, the most natural way to capture these drives' effect on aluminum supply. (And the most prudent, since there were so few observations.)

One of the first lessons my father taught me, in our forays fixing things around the house, was this: don't force it. If you have to jam the pieces together, wrest that ornery nut onto that bolt, it's a sign of a problem. Hold on a moment. Consider your options. And, when you act, be alert for feedback that you are doing things correctly. Tread gently, and use a light touch. It's just good craftsmanship.

8.2.4 One Principle to Rule Them All

The common thread uniting these concepts is the underlying principle of econometric modeling, and, indeed, of modeling generally.

The Principle of Harmony: *The analytics used to study a phenomenon should accord with the essential features of that phenomenon and the data used to study it.*

The model's specification, estimator, and error structure should align with reality, as understood through description and vernacular knowledge, and with the experimental content of the analysis.

Models infused with this property have a practical cast. The reader feels not as if she hovers far above the phenomenon of interest, but as if she is drawn in closer to it. Many applied statistics articles have this quality, at least in part, as their models often extend from preliminary data analysis. One can also find it with some frequency in *The Review of Economics and Statistics*, and even more frequently in some journals that border economics, such as *Risk Analysis*.

This principle doesn't require a model to be a literal, piece-by-piece representation of the process at work. Abstraction was one of the first modeling tools we developed, after all. But it does not permit an econometric model to be dissonant with reality or disassociated from it. When that happens, it amounts to analyzing a made-up world, not the real one. Then something about the model will be unnatural, as with the OLS regression Hamermesh was forced to use. An unnatural element of a model often signifies a disconnect between that model and reality—a sign of a problem that itself deserves to be taken seriously.

8.3 The Econometrics of Orchestra Auditions

With these ideas in mind, let us turn to Goldin and Rouse's (2000) analysis of the hiring effects of American orchestras' transition to "blind auditions," which hide the gender of the candidate.

Years ago these were hardly needed, because women weren't being hired anyway. In 1960, none of the "big five" orchestras had more than 5% women. But by the late 1960s one of these, Cleveland, had achieved 10%, and within a decade the others had followed suit. By the mid-1990s the average across the big five exceeded 20%. Goldin and Rouse estimate that the switch to blind auditions during this period accounts for about one-quarter of this improvement.

This paper is deeply contextual. This quality lets the authors substantiate the plausibility of gender discrimination in this market (p. 719):

> Claims abound in the world of music that "women have smaller techniques than men" and are "more likely to demand special attention or treatment." Many European orchestras had, and some continue to have, stated policies not to hire women.

and lets them detail how thoroughly blind auditions mask the candidate's gender (p. 721):

> The screens we have seen are either heavy (but sound-porous) cloth, sometimes suspended from the ceiling of the symphony hall, or look like large room dividers. Some orchestras also roll out a carpet leading to center stage to muffle footsteps that could betray the sex of the candidate. If a carpet is not placed on the stage, the personnel manager may ask a woman to take off her shoes and he provides the compensating footsteps.

Blind auditions aren't just less discriminatory; they're non-discriminatory. The estimates can be interpreted accordingly.

The paper's econometric centerpiece is an individual-audition-round-level linear probability model:

$$\mathrm{P} = \alpha + \beta B + \gamma F \cdot \mathrm{B} + \delta X + \epsilon \tag{8.4}$$

where α is a set of individual fixed effects, P is the probability of being hired or advanced to the next round of auditions, B and F are dummies for a blind audition and a female applicant, and the controls in X include individual and audition-round characteristics and year fixed effects. This estimation approach and its trappings are standard and familiar. The coefficient of interest, γ, is identified through differences-in-differences: males vs. females, blind vs. non-blind. The main econometric concern, endogeneity, is ruled out, by showing that the probability an orchestra adopts a screen is unrelated to its gender mix (their Table 2). Robustness checks are conducted, and thousands of observations offer statistical precision.

And now we see exactly what it means to have a model whose purpose is not to explain the dependent variable but to estimate a coefficient instead. The closer we look, the more we wonder whether it harmonizes with the situation being modeled and its experimental content.

First consider the process governing the adoption of blind auditions and the gender composition of hires. The *paper* says a limited amount about this process, but the *model* speaks volumes. Taken seriously, it says that the movement toward gender-neutrality transpired, to the first order, with a uniform change in attitudes nationwide (captured by the year fixed effects), along with blind auditions that are adopted out of the blue (via the exogeneity of B).

Really? Is it reasonable to expect this set of orchestras to co-evolve so tightly in the "progressiveness" of their attitudes toward gender equality, especially when the paper's descriptive statistics (their Figs. 1 and 3) suggest otherwise? If so, then why would these orchestras co-evolve so loosely in the mechanics of auditions and hiring? If progressiveness doesn't influence the use of blind auditions, then what does? None of these questions are answered.

Consequently, we must resort to thinking about the system, which consists of progressiveness, audition mechanics, and the relative supply of adequately trained

female musicians. The authors' tepid endogeneity check notwithstanding,[8] I would expect these interconnected elements to evolve jointly within each orchestra in favor of greater female representation. (The last two do, evidence in the paper makes clear.) If so, the inability to account for the first element is problematic, since it will be related to the other two elements, while independently affecting the gender composition of hires as well. Furthermore, there is no reason to expect each orchestra to evolve at the same pace. We should doubt whether the model reasonably represents the process at work. It sure seems unnatural, but it's hard to know for sure.

Next consider the analysis' experimental content. The coefficient γ is identified three ways, all assumed by the model to be equivalent, like Jaimovich (2013) in reverse. There is long-term temporal variation, trends in the gender mix of hires among those orchestras that adopted blind auditions during the sample period, along with the short-term temporal variation in hiring immediately surrounding these adoptions. There is also cross-sectional variation, since this particular model does not contain orchestra-level fixed effects.

Only two of these sources of identification have true experimental content: cross-sectional and short-term temporal variation, since we can expect the full effect of blind auditions to be experienced without delay. There isn't much grist for the mill. Few of the eight orchestras in their "audition sample" switch audition methods, so whether we pair up blind/non-blinded orchestras or look before/after within orchestras, the number of experiments does not exceed four.

Why, then, are the standard errors not large? Because there is no random effect at the level of the experimental unit, the orchestra-year, and the number of observations far outweighs the number of experiments.[9] These random effects probably would be autocorrelated, especially if the system evolves as suggested above. The model's

[8] The notion that progressiveness influences the use of blind auditions is rejected by a probit model that relates the probability an orchestra adopts a screen to its proportion of female members, dropping orchestras from the sample in the year after they adopt a screen. Even two standard errors away from the coefficient estimate, the extent of reverse causation is small.

This check is flimsy conceptually and econometrically. Conceptually, this proportion, a stock built up slowly over time, may poorly represent short-term changes in progressiveness. Econometrically, the experimental power of this test is heavily concentrated in the only non-adopting orchestra in the sample, Cleveland, because it remains in the sample twice as long, and the dependent and independent variables are both trending time series. In a crude replication, described below, removing Cleveland from the sample or using a more appropriate estimator yielded ambivalent results.

(To "replicate" the authors' endogeneity test, I created a crudely-realistic data set including 47 observations for one non-adopting orchestra and 24 for the other ten, which each adopt blind auditions in that 24th year, putatively because of "progressiveness." For all orchestras, the independent variable equals zero for the first thirteen years, and then grows at the same linear rate, with a touch of noise added, until the orchestra leaves the sample. The "t-statistic" on the probit coefficient was six times that on the coefficient in a more-appropriate hazard model, which has a low degree of statistical power; removing the non-adopting orchestra from the sample, the probit coefficient becomes unidentified.)

[9] This paper was published before clustered standard errors became common. Nonetheless, I hope the ensuing discussion convinces you that employing random effects dominates clustering in this situation, for the purposes of harmony, insight, and fidelity.

stochastic and deterministic components both provide an awfully sterile description of what is likely to be an intensely dynamic process. As a result, favorable bias can creep into the coefficient estimates and their standard errors.

What would I change about this model? We both know what I'm supposed to say: eliminate cross-section variation via orchestra-level fixed effects, account for autocorrelation using orchestra-level trends, focus on the "discontinuities" around the dates blind auditions were adopted. Well, I don't know. Smothering the data in these additional controls and restrictions is probably impractical. Craftsmanship is not obsessive-compulsive disorder. The phenomenon is interesting, the data is detailed, and even non-causal results about the evolution of the system are worthwhile.

So my answer, at least to start, is this: nothing. Rather than be heavy-handed, I would prefer to use a light touch. With this data, basic description cannot adequately portray the system: progressiveness and the relative quality of female auditioners must be estimated. Equation (8.4) can do this, so why not treat it as a first cut at the data, and use it to describe the system instead? From its estimates, one can calculate the gender gap in applicant quality and expected success probabilities. Subtracting the latter from the gender gap in actual success probabilities gives you a residual at the orchestra-year level: a measure of progressiveness.

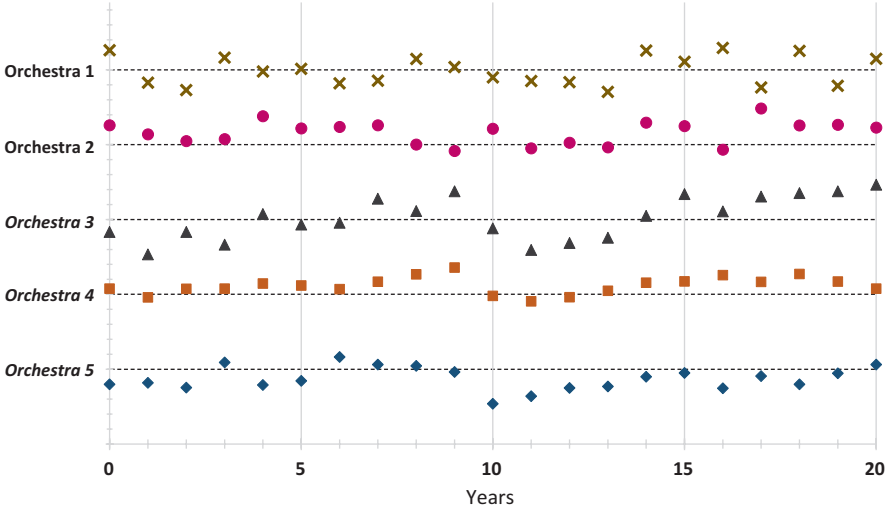


Fig. 8.2 Hypothetical residuals by orchestra. Note: Orchestras 1 and 2 never switch audition methods. Orchestras 3-5 switch to blind auditions at the beginning of year 10. Five orchestras are assumed to exist, rather than eight, for simplicity. The dashed line for each orchestra is set at zero. Orchestra 1's residuals are white noise

Because there are so few orchestras, each of these variables can be plotted on a single, orchestra-level graph, as in the residuals prototype in Fig. 8.2.[10] This figure transparently illustrates micropicture and macropicture: the three sources of identification, (part of) the dynamics of the system, and potential econometric problems. The more the residuals deviate from white noise, the less suitable the specification.

Furthermore, the nature of the deviation shapes the container: the second-generation model from which final conclusions would be drawn. If the cross-sectional aspect of the estimates is driven by one or two "outlier orchestras," then one should focus on the authors' orchestra fixed effects estimates. If there are long-term trends in progressiveness, these must be accounted for, perhaps with a degrees-of-freedom-saving AR1 random effect. If there is endogeneity surrounding the change in audition methods, documenting the evolution of the system (without making causal claims) may be the best you can do. Each of these problems will each leave its own unique signature in the graph, a different departure from white noise. Identifying these signatures and sketching out a next-generation model are left as Food for Thought.

8.4 Conclusion

Compared to the crispness of a theorem or a lemma, the calm authority of a well-hewn regression specification, or the certitude of a randomized, controlled experiment, a model seems rather…vulnerable. Economic theory, tradition, statistical virtuosity–the reassurance they offer the modeler is false, borne on the desperate hope that credibility can be reduced to competence. Do not believe it. It cannot be so. The clearest sign that you are on the right track is not adherence to any of these professional pillars, but an accumulation of evidence that the model, the data it analyzes, and the phenomenon of interest all fit together without strain, like the pieces of a jigsaw puzzle when properly assembled.

Even then, we can't be sure. We can think we are onto something, but we can't know or even think we know. To move in this direction, we must begin the wrenching process of checking our work. This, perhaps, is where the scientific method helps experimentalists most: it puts so many strictures in place at the beginning that one need only run simple statistics at the end. Modelers can't rely on these same strictures, so instead we must turn on ourselves, like an autoimmune disease, and try to tear down what we just invested so much in building up. To ensure that we understand something, we must relax the pretense that we already understand it. Then we can begin to traverse the lonely path that leads away from "thinking" towards "knowing."

[10] Assuming the data is sufficient to the task, or that any data insufficiencies can be finessed. Otherwise, these issues may have to be explored more opaquely, using formal tests.

Food for Thought

1. The opening to this chapter addressed the use and misuse of weighting across cross-sectional units of different "size," such as states.
   (a) Set out an individual-level regression with a state-level policy variable and a state-level random effect, as in Eq. (8.1). Work with it to show that, in a state-level analysis, West Bengal should be weighted more than Kashmir, but not seven times more.
   (b) Chapter 3 addressed the same issue differently. Return to question #3 in that chapter and give it another try. Does our discussion of experimental content make it easier to answer?
2. Figure 8.3 contains a somewhat-recent snapshot of the states that had adopted an important policy in the U.S. Its scale of implementation also seems "somewhat geologic." (It is, in fact, extremely geologic. States that were covered by the Western Inland Sea–look it up–are far less likely to have adopted this policy.) What policy is this?
3. The following questions all pertain to the chapter's discussion of Gruber et al. (1999).
   (a) Compare my Table 8.1 with the original in Gruber et al. (1999). Note the differences. How do these differences reflect principles of effective description?

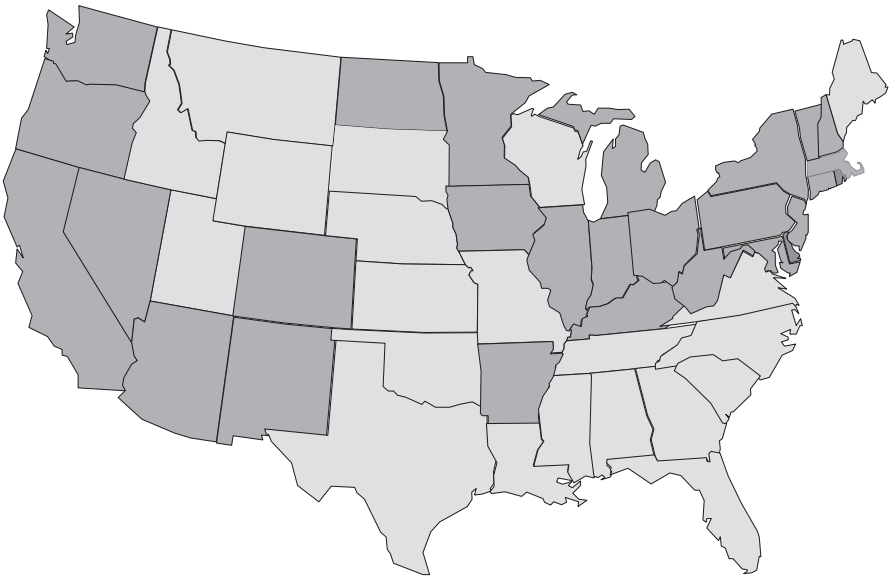


Fig. 8.3 The shaded states in this map had adopted a particular policy, as of a date in the not-too-recent-past. Can you tell what policy this is?

(b) The choice to examine the sum of squares of the changes in the fee differential is not inconsequential. Could comparing the magnitudes of these values, rather than their squared magnitudes, be justified? Does examining the weighted sum of squares focus on how California affects the coefficient estimates or the standard errors?
(c) The chapter claims that every "problem" with the experimental content of your analysis conforms to a violation of the classical OLS assumptions. If so, which key assumption is violated in Gruber et al.? Relate your answer to the techniques advocated in Bertrand et al. (2004).

4. The following questions all pertain to the chapter's discussion of Goldin and Rouse (2000).

(a) Because orchestra auditions consists of several, sequential rounds, the most natural econometric model for the situation analyzed by Goldin and Rouse (2000) would be an ordered probit, were the data suitable to support it. Assume you possessed an index of auditioner "quality." Then lay out a suitable ordered probit model to examine the effect of blind auditions on females' probability of advancement, in which the effect of discrimination against females shifts the thresholds required for advancement to the next round.
(b) One way to think about the system's dynamics is to create a latent variable for "progressiveness," which increases at different rates within different orchestras, and which influences some of the coefficients and variables in Eq. (8.4). Incorporate this variable into this equation, and show how bias and serially correlated errors are likely to result.
(c) I prefer the following specification to Eq. (8.4), though it is econometrically equivalent, because it better describes the actual process at work. How so?

$$P = \alpha + \beta B + \gamma F \cdot (1 - B) + \delta X + \epsilon$$

5. These questions pertain to Fig. 8.2.

(a) The residuals for each orchestra but one signify an econometric issue. Identify the issue associated with each orchestra.
(b) In this figure, the errors for each orchestra signify a different econometric problem. Is that likely to happen in practice, or is it more likely that many orchestras would display similar problems?
(c) In general terms, explain how the "second-generation model" discussed in the text could be used to conduct the final analysis at the level of the experimental unit, as advocated in this chapter. Then explain why would be necessary to do so, in order to get correct standard errors.
(d) (Difficult.) Work out how to implement this approach, in order to get unbiased coefficient estimates and standard errors, under the assumption that the deterministic component of the model is correctly specified, and that there are no problems with serial correlation.

6. The main specification in Card and Krueger's (1994) two-period, two-state difference-in-difference analysis is as follows:

$$\Delta E = \alpha + \beta X + \gamma NJ + \varepsilon$$

where E is restaurant-level employment, X are restaurant-level controls, and NJ is a dummy for New Jersey, the minimum-wage-increasing state. Show that if a random effect is included in this specification at the level of the experimental unit, the true standard error of $\hat{\gamma}$ is unidentified and its OLS standard error is biased downward.

7. Traffic accidents can be considered independent, (statistically) rare events, so the number of fatal accidents in any given place and time has a Poisson distribution. This naturally lends itself to a count data model. However, a pure Poisson regression has the classic problem that the conditional variance of fatalities equals its mean, which does not hold up in practice.

One method of addressing this problem is to specify a generalized linear model:

$$F_{s,t} \sim Poisson(f_{s,t})$$
$$\log(f_{s,t}) = \alpha + \beta X_{s,t} + \varepsilon_{s,t}$$

where $F_{s,t}$ is the observed number of fatal accidents in location s in period t, X contains the explanatory variables, and $\varepsilon_{s,t}$ is a random effect. The second line of this equation predicts the latent variable $f_{s,t}$, which serves as the mean (and variance) of the Poisson distribution from which realized $F_{s,t}$ is "drawn."

An alternative method is to specify a Negative Binomial Model:

$$F_{s,t} \sim NegBin(f_{s,t}, \sigma)$$
$$\log(f_{s,t}) = \alpha + \beta X_{s,t}$$

where σ is the "overdispersion" parameter associated with the Negative Binomial distribution.

(a) Which is the more natural model? Why?
(b) Which model is more commonly used in economics? Why?

8. Cunningham et al. (2018) examine how abortion clinic closures affect abortion rates in Texas, relating the change in counties' abortion rates to the increase in the distance women must travel to have an abortion when a closer clinic closes (measured using five "increase in distance" categories). There are 254 counties in Texas and 42 abortion clinics in their sample, 18 of which closed during the interval a clinic-closing law took effect, leaving 11 of 16 metropolitan statistical areas without a clinic. There were nearly four million Texas pregnancies during the study period, approximately 10% of which ended in abortion.

Sketch out the experimental content of this study. Identify the experimental unit and the spatial and temporal scale of variation in the key independent variable. To the nearest power of ten, how many independent experiments does this study contain?

References

Angrist J, Krueger A (1991) Does compulsory school attendance affect schooling and earnings? Q J Econ 106(4):979–1014

Angrist J, Pischke J-S (2009) Mostly harmless econometrics: an Empiricist's Companion. Princeton University Press, Princeton, NJ

Ashley R, Parmeter C (2015) Sensitivity analysis for inference in 2SLS/GMM estimation with possibly flawed instruments. Empir Econ 49(4):1153–1171

Bertrand M, Duflo E, Mullainathan S (2004) How much should we trust differences-in-differences estimates? Q J Econ 119(1):249–275

Bound J, Jaeger D, Baker R (1995) Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. J Am Stat Assoc 90(430):443–450

Cabral M, Cullen MR (2016) Estimating the value of public insurance using complementary private insurance. (No. w22583). National Bureau of Economic Research

Card D, Krueger A (1994) Minimum wages and employment: a case study of the fast food industry in New Jersey and Pennsylvania. Am Econ Rev 84(4):772–793

Cunningham S, Lindo J, Myers C, Schlosser A (2018) How far is too far? New evidence on abortion clinic closures, access, and abortions. (No. w23366). National Bureau of Economic Research

Dickens W (1990) Error components in grouped data: is it ever worth weighting? Rev Econ Stat 72(2):328–333

Feyrer J, Mansur E, Sacerdote B (2017) Geographic dispersion of economic shocks: evidence from the fracking revolution. Am Econ Rev 107(4):1313–1334. Available as a 2015 working paper

Goldin C, Rouse C (2000) Orchestrating impartiality: the impact of blind auditions on female musicians. Am Econ Rev 90(4):715–741

Gruber J, Kim J, Mayzlin D (1999) Physician fees and procedure intensity: the case of cesarean delivery. J Health Econ 18(4):473–490

Hamermesh D (2000) The craft of labormetrics. Ind Labor Relat Rev 53(3):363–380

Jaimovich D (2013) Missing links, missing markets: internal exchanges, reciprocity and external connections in the economic networks of Gambian villages. (No. 2209075). Social Science Research Network

James A, Smith B (2017) There will be blood: crime rates in shale-rich US counties. J Environ Econ Manag 84:125–152. Available as a 2014 working paper

Maniloff P, Mastromonaco R (2015) The local economic aspects of fracking. Manuscript, Colorado School of Mines and the University of Oregon. Working Paper. http://pages.uoregon.edu/ralphm/fracking_may_15.pdf

McCollum M, Upton G (2018) Local labor market shocks and residential mortgage payments: evidence from shale oil and gas booms. Resour Energy Econ 53:162–197

Moulton B (1990) An illustration of a pitfall in estimating the effects of aggregate variables on micro units. Rev Econ Stat 72(2):334–338

Munasib A, Rickman D (2015) Regional economic impacts of the shale gas and tight oil boom: a synthetic control analysis. Reg Sci Urban Econ 50:1–7

Paredes D, Komarek T, Loveridge S (2015) Income and employment effects of shale gas extraction windfalls: evidence from the Marcellus region. Energy Econ 47:112–120

Solon G, Haider S, Wooldridge J (2015) What are we weighting for? J Hum Resour 50(2):301–316

Part IV
Ways of Knowing

Chapter 9
Testing

Abstract This chapter treats hypothesis testing as an opportunity for the researcher to distinguish between three possible explanations for a set of empirical findings: random chance, the scientific hypothesis of primary interest, and alternative scientific hypotheses. The methods it offers to advance this goal involve refining the null hypothesis, while increasing the scrutiny of the primary scientific hypothesis of interest and the number of alternative scientific hypotheses that it must compete with. These methods are brought to life in applications to home sales in New England, multiproduct pricing in Major League Baseball, turnout in Congressional elections, and the link between abortion and crime.

9.1 The Nature of Hypothesis Testing in Economics

Of all the elements of an empirical study, hypothesis testing seems the most formulaic. You assume a null–typically, that the coefficient of interest is zero. You then calculate a test statistic from your estimates, often a t-statistic, and reject the null only when that test statistic is unlikely to have occurred by random chance.

It does not seem to bother most economists that, conceptually, this procedure is comically misaligned with what the man on the street would need to be convinced of anything. It would mean little to show that your straw man, the null hypothesis, is probably wrong. He would want to know that the scientific hypothesis you spent so much time developing is probably right.

Certainly, this problem did not bother R.A. Fisher, the potato farmer[1] who designed this procedure. That was for a very good reason: he was running controlled experiments. In theory, the only thing affecting the outcome, beyond the treatment, was our old friend Lady Luck.[2] We know her well. We know how she can affect

[1] Also legendary statistician, biologist, and Knight of the Order of the British Empire, but still.

[2] In practice, Fisher inadvertently made various types of experimental error. For example, he did not apply the various fertilizers he was testing to his potatoes in a random order. There's Hypothesis F for you.

D. Grant, *Methods of Economic Research*, Springer Texts in Business and Economics, https://doi.org/10.1007/978-3-030-01734-7_9

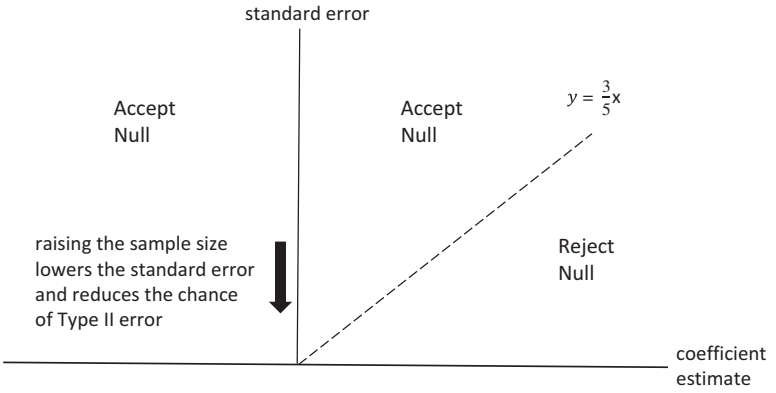


Fig. 9.1 Hypothesis testing in a univariate regression, in which theory predicts a positive effect. Note: For sample sizes above about fifty, 3/5 closely approximates the reciprocal of the t-statistic required to reject the standard null at $\alpha = 0.05$ in a one-tailed test

estimated relationships. We know the distribution of test statistics thereby resulting under the null. Classical testing is well-suited to these kinds of experiments.

You could think about it this way. Consider all *sets* of estimates that could conceivably be produced from your study. For a univariate regression, this would be all feasible pairs of the slope coefficient estimate and its standard error (ignoring the constant, as we usually do). For a classic difference in means test, this would be all feasible triplets consisting of the two samples' means and the sample standard deviation, and so on. Under the null hypothesis, some sets of estimates are very unlikely to occur, such as those with large t-statistics, while others are not. The former fall in the rejection region, and are attributed to your scientific hypothesis. Everything else is attributed to the null. This is shown in Fig. 9.1, in the regression context.

Bayesians, McCloskey, and others would legitimately object that the estimates attributed to the null might include some that are economically meaningful. A better way to put it is illustrated in Fig. 9.2: the collection of estimates that are consistent with the scientific hypothesis of interest might overlap with the collection of estimates that are consistent with the null. This would not bother the potato farmer too much either. Increasing the sample size decreases the size of both groups, and shrinks the area of overlap doubly fast. For controlled experiments, it is often cost-effective to obtain the sample size needed to achieve a reasonable level of precision. If so, your problem is solved.

Figure 9.3 compares these relatively manageable situations with the conundrum faced by economists who cannot run randomized, controlled experiments, cannot govern the sizes of their samples or their experimental content, cannot rule out other theoretical explanations for observed relationships, and do not conduct scale analyses that would allow them to gauge the effect sizes implied by their theories. We now have a potpourri of hypotheses: the null, your theory, other known theories, and other unknown theories, each but the null encompassing an amorphous and uncertain range of potential estimates, with no traditional experimental tools available to

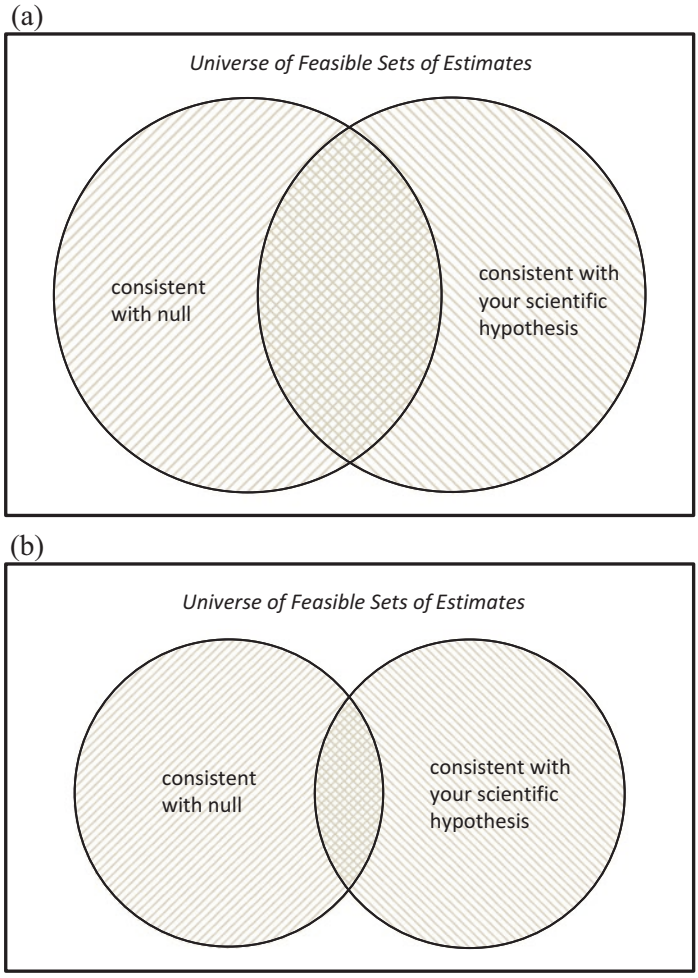


Fig. 9.2 (**a**) A general representation of basic hypothesis testing. (**b**) More data reduces the overlap. Note: Each element of this universe is a potential set, or vector, of estimates. This would include coefficient estimates, estimates of the standard errors, and possibly other estimated moments. "Consistent with" means "not too unlikely under."

distinguish between them. Hovering over them all is the most venerable, time-tested hypothesis ever: Hypothesis F.

Hypothesis F: *You f—ed up.*

To combat this multi-headed hydra our profession currently employs … the same method as the potato farmer. Go figure.

The essence of the problem is that there can be a lot of daylight between the set of estimates that are consistent with the traditional null and the set of estimates that are consistent with your theory and your theory alone (among all plausible contenders).

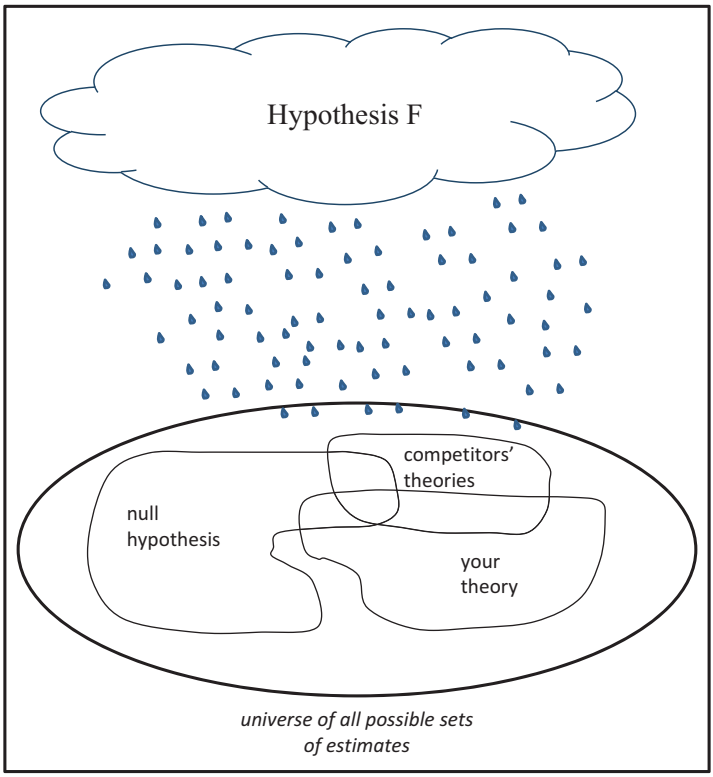


Fig. 9.3 The economist's conundrum

When you have a (properly designed) randomized controlled experiment, saying "my estimates probably aren't due to random chance" means a lot: the only other alternative is the treatment. But when there are a welter of other possibilities, ruling out this one option means much less. Your theory must compete against others, including Hypothesis F. Showing that it out-performs the null hypothesis of no effect–that it is literally better than nothing at all–is often pretty thin gruel.

What Happened to Roger's Homework?

Teacher: Roger,[3] what happened to your homework?
Roger: My dog ate it.
Teacher: Roger, I find that hard to believe.
Roger: I tested the null that nothing happened to my homework, and was able to reject it with an alpha of .05. So it must have been my dog.
Teacher: OK, so, zero in the grade book it is.

[3] In keeping with almost every economics textbook out there, all names in this book are semi-obsolete, white-person names from the 1960s, with "Juan" thrown in for multicultural flavor.

This is not just a theoretical point. The weaknesses of basic hypothesis testing are becoming apparent where it matters most: in the field of battle. Outside our profession, psychology and medicine are suffering from "replication crises," in which the findings of many well-received papers cannot be duplicated, sometimes even by their own authors (Lehrer 2010; Gelman 2016). In the field with the greatest replicability issues, one journal, *Basic and Applied Social Psychology*, has banned standard hypothesis testing lock, stock, and barrel. In the medical literature, Ioannidis (2005) characterizes the probability that a published research finding is actually correct, accounting for Type I error, the frequency with which a given a hypothesis is tested, and the possibility of bias in the conduct of the research and in the decision to publish. Under realistic conditions, he concludes that "most research findings are false for most research designs and for most fields." And this is not for social science-style regressions, but for experimental studies!

As one might guess, this problem is even greater in our corner of the world. Paldam (2016) conducts an Ioannidis-style analysis for economists, assuming the use of regression methods by "rational," self-interested researchers, and draws similar conclusions. Fanelli (2010) finds that support for the posited scientific hypothesis is more common in papers published in the social sciences than in the hard sciences, and particularly common in economics and business, where almost 90% of the papers surveyed supported their primary hypothesis.[4] The meta-regressions of economist T. D. Stanley and various coauthors, especially, repeatedly find estimated effect sizes to be inversely related to the standard error of those estimates: the hallmark of a literature infected by bias in the conduct of the research, the decision to publish it, or both. Paldam reminds us this isn't merely from lumping in the rabble with the elite:

> Academic economists tend to believe that this relationship is due to differences in study quality. Meta-analysts have often checked for this by controlling for the impact factor of the publication output; dummies for techniques and time trends are also often included. Such variables' effects tend to be small, and often insignificant.

Our response to this should not be to abjure classical hypothesis testing entirely. The structure it provides is too valuable. We are left with two options: to restrict ourselves to circumstances in which it works well, or to fold it into a broader testing philosophy that enhances the credibility of the body of tests performed and reported.

The first option is less messy. It works when there really is no plausible alternative to your theory, other than Lady Luck: the "perfect" IV, the "perfect" RD, "perfect" randomization. For these, the structure of classical hypothesis testing is a "perfect" fit.

[4] In the studies that can be appropriately classified, my "failure rate" is almost four times higher. In five tests of structural models, and ten of reduced form models, the main scientific hypothesis is supported 60% of the time.

But otherwise–when these strategies aren't as perfect as they seem, or when they don't apply in the first place–we must employ the second option, and treat hypothesis testing not as a formulaic task, but as an opportunity to mark the territory between the set of estimates that are consistent with the traditional null and the set that are consistent with your theory alone.

9.2 Three Ways to Enhance Testing

In order to accomplish this task, you have three instruments at your disposal: the null hypothesis, the alternative hypotheses, and the predictions deriving from these. Refine the first, amplify the second, and expand the third.

9.2.1 Refine the Null

There are two ways to do this, each of which is gently unconventional. The first is to refine what I call the "passive" null that, if rejected, would lend support to your scientific theory. To use the typical null that the regression coefficient is zero is passive indeed. Is this really what people would believe in the absence of your theory? Consider instead the practical alternatives, what someone would be giving up in order to adopt your theory. Then base your null on that.

The Opportunity Cost Principle of Hypothesis Testing: *The best passive null for any scientific hypothesis is the next-best practical replacement that would be used instead.*

Sometimes this replacement will be scientific: a hypothesis that already has a high degree of scientific consensus. Sometimes it will be utilitarian: the most reasonable hypothesis on which decisions would otherwise be based. Sometimes it will be prudent: a weaker or less-precise hypothesis that is more likely to be correct. It depends on the circumstances.

If you cannot rule out this null in favor of your alternative, then, in deference to Lady Luck, the Precautionary Principle, other potential explanations for the phenomenon of interest, and Hypothesis F, choose the weaker, more traditional, more reliable null hypothesis. But if you can reject this refined null, you have met a higher standard. The credibility of your results is suitably enhanced.

The second way to refine the null is to turn hypothesis testing on its head: specify your scientific hypothesis as the null, and try to rule *it* out. Rather than the standard, passive null that is intended to act as a foil that your data will discard, adopt an "active" null that is anything but.

This is not something we are taught to avoid so much as something we rarely consider to begin with. It is seemingly out of the question. But this is not because

we have a deep metaphysical disagreement with the potato farmer. We just rarely presume that we can specify such a precise hypothesis in the first place.

In fact, you have three options, depending on the degree of exactitude of your model. The most exacting involves functional form. If you derive a precise, structural relationship between x and y, and you trust your functional form assumptions, this *specification* can be treated as the active null and tested against a general alternative. A second option involves heuristics, which can take a variety of forms, as shown in Chaps. 3 and 6. The last option involves coefficient values, which sometimes can be specified precisely (or within some feasible range, as with the schooling coefficient in Mincer's earnings model, Eq. 3.6). All these options are more rigorous than a passive test of a coefficient sign.

In a sense, classical hypothesis testing scrutinizes the wrong hypothesis: the neutral one, not the one you developed theoretically. Employing an active null flips this situation around, and puts the scrutiny where it belongs. Your testing now accords with the widely accepted philosophy that a scientific hypothesis cannot be proven, only fail to be rejected. Adding data sharpens the precision of your estimates and increases the chances that you will reject your precious null. Karl Popper would be proud.

And you will be less likely to kid yourself. When estimates are imprecise, it is hard to rule out the data as being inconsistent with anything, and standard hypothesis testing makes sense. It prevents weak data from lending support to a scientific theory. But when this is not the case, a raft of estimates could be inconsistent with the passive null and yet also inconsistent with your theory (the white space in Fig. 9.2). Adopting a passive null and implicitly treating your theory as the alternative amounts to claiming that no-man's-land as your own. Testing an active null cedes that territory back to its rightful owner.

Both refined passive nulls and active nulls can seem intimidating because they are not easy to execute. In general, this is true. Specification tests, tests of nonnested hypotheses, and tests of nonparametric relationships can be quite complex, perhaps not worth the cost. But, in practice, there are often simple, feasible ways to implement these approaches, as we will now see.

Application to Voting. The "rational voter" literature mentioned in Chap. 4 has progressed partly through improvements in the null hypotheses that are tested. Early on, the notion that turnout will be higher in close elections, ceteris paribus, was tested repeatedly with regressions of the form:

$$T = \alpha + \tau_1 M + \tau_2 S + \Lambda X + \epsilon \tag{9.1}$$

where T is turnout in percent, M is the winning margin (the difference between the winner's and loser's vote shares), S is the size of the electorate, and X is a vector of controls. The theory predicts $\tau_1 < 0$, $\tau_2 < 0$, because larger, more lopsided elections were, ex ante, less likely to have ended in ties. This lends itself to routine one-sided t-tests, which typically rejected the standard null of $\tau_1 = 0$, $\tau_2 = 0$ in favor of this alternative.

Standard stuff, but not overly persuasive, because of two standard concerns. First, the number of controls was limited, so M and S may be related to omitted variables that are driving the coefficient estimates. (For example, many studies omitted campaign spending, a key part of the system, which correlates negatively with M and positively with T.) Second, without scale analyses, which are almost absent from this literature, it was unclear just how large τ_1 and τ_2 should actually be, and thus whether the magnitude of the estimates was reasonable. These central challenges to credibility can be met by refining the null hypothesis.

One way to do this is to make Eq. 9.1 a refined passive null and test a more precise theoretical prediction against it. After all, this equation is not strictly implied by theory, yet would be a natural consequence of omitted variables. So why not take this possibility seriously, and treat this equation as the null?

A simple way to implement this approach converts the testing of this refined null into a standard hypothesis test (Grant 1998). Using a well-known formula, impute the probability of casting the deciding vote, P, from M and S, and insert it into Eq. 9.1:

$$T = \alpha + \tau_1 M + \tau_2 S + \gamma P\left(M,S\right) + \Lambda X + \epsilon \tag{9.2}$$

The refined null hypothesis, Eq. 9.1, simply implies that $\gamma = 0$. The alternative is that $\gamma > 0$. This is a much higher standard, and the evidence for rational voting is that much stronger if it is met. If τ_1 and τ_2 are also insignificant, then maybe your controls weren't so bad after all.

Further following this line of thinking generates several active null hypotheses that are even more discerning. The functional form used in constructing P from M and S can be broken down, using scale analysis and Taylor series, to generate the following structural relationship (Grant and Toma 2008):

$$\ln\left(Y\right) + \textit{general equilibrium adjustment factor} \approx \beta_0 + \beta_1 M^2 - \ln S + \epsilon \tag{9.3}$$

where Y is a simple function of T, after accounting for controls, and the adjustment factor is a complicated mathematical expression. These two left-hand side terms can be estimated in a first-stage regression. Relating their sum to M^2 and ln(S) yields the following specification:

$$\begin{aligned} &ln\left(\hat{Y}\right) + \textit{estimated general equilibrium adjustment factor} \\ &\quad \approx \beta_0 + \beta_1 M^2 + \beta_2 \ln S + \xi \end{aligned} \tag{9.4}$$

All three types of active null hypotheses extend from this model. The first involves parameter values, such as ß$_2$ = −1. (One could also append a linear term in M to the specification, and test the active null that its coefficient is zero.) The

second involves heuristics–the same ones that failed so dramatically in the union certification elections discussed in Chap. 4. One such heuristic is that there should be a certain type of concavity in the *M-T* relation, conditional on *S*. It would be complicated to formally test this in the traditional manner. It is more practical, instead, to specify this heuristic as an active null, which clearly is not rejected if it corresponds to the patterns in the data. The third type involves functional form. The specification prescribed in Eq. 9.4–quadratic in *M*, logarithmic in *S*–can be tested against a general nonparametric alternative.

This was easy to do for congressional elections, which have no variation in *S* to speak of. First estimate the relationship in Eq. 9.4 and plot the implied *M-T* relation on a graph. Then estimate the *M-T* relationship nonparametrically, without functional form restrictions, and graph it. (Include the same parametric controls in both estimations.) The two relations, illustrated in Fig. 9.4, are indistinguishable economically and statistically: we clearly cannot reject the active null that Eq. 9.4 is on point. It is difficult to envision how this result would occur *without* the rational voter model being responsible, a claim that was harder to make for Eq. 9.2 and impossible to make for Eq. 9.1.

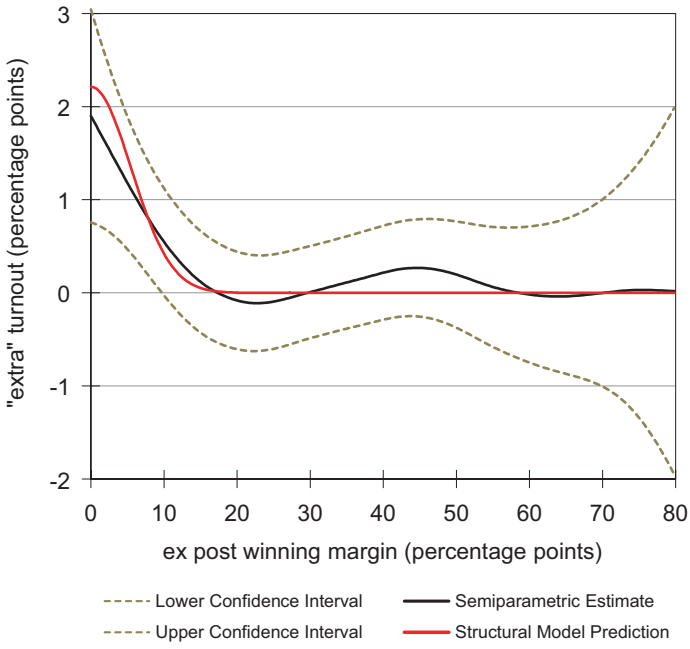


Fig. 9.4 Semiparametric and structural estimates of the relationship between T (vertical axis) and M (horizontal axis), after accounting for controls. Reprinted by permission from Springer Nature, Grant D, Toma M (2008) Elemental tests of the traditional rational voting model, Public Choice, 137(1):173–195. Copyright 2008

9.2.2 *Amplify the Alternatives*

Here I refer not to the alternative to your null hypothesis in the statistical sense. This is unambiguously defined. I mean alternative explanations for the phenomenon you are studying or the estimates you obtain while doing so. Tacitly, traditional hypothesis testing envisions just two possible explanations for the estimates you have obtained: your scientific theory and the passive null. However, human motivation and social arrangements are so complex that there are usually other explanations as well, plus we should allow for unanticipated "unknown unknowns." One way to account for this is to test the validity of these alternatives.

Let there be a set of empirical estimates that implies a particular pattern of behavior. Break down all potential explanations for these estimates into a set of mutually exclusive hypotheses. Each hypothesis has some probability of being correct, and all probabilities must sum to one. Reduce the probabilities associated with some hypotheses, and confidence in the others–including your explanation of choice–must rise.

This is easily seen in the simplest case, in which there are just two distinct explanations for a given phenomenon. Occupational licensure could protect the interests of consumers who can't judge quality on their own, or it could serve the interests of incumbent practitioners by limiting competition (e.g., Law and Kim 2005). Resale price maintenance and exclusive territories could foster the provision of dealer services that boost the product's sales and satisfaction, or they could promote collusive pricing (e.g., Sass and Saurman 1993). Here, evidence against the one option is evidence favoring the other.

Even when things aren't presented so neatly, you can still conduct general tests of generic alternatives, with no other particular explanation in mind. An active null does this, since it lets your theory be rejected against an alternative that you hadn't even imagined. So do robustness checks and falsification tests, which examine an alternative explanation for your findings: Hypothesis F. This hypothesis implies that you got lucky: the results of your basic regression just happened, through bias, misspecification, etc., to support your theory. Falsification tests that apply your model to circumstances where it does not pertain, and robustness checks that estimate variants of your basic specification, are ways of seeing just how lucky you got. If these don't support Hypothesis F, confidence in your findings should rise.

Application to Multiproduct Pricing. My coauthor and I took these techniques to the extreme in the multiproduct pricing paper discussed in Chap. 6. Using principal component analysis, this paper quantified two strands of variation in the prices of tickets, concessions, parking, etc., offered at professional baseball games. Based on the factor loadings, we interpreted the first strand, the principal component $COMP_1$, as an overall demand effect, and the second, $COMP_2$, as reflecting price discrimination.

The problem is that there is no direct theoretical connect between factor-analytic methods and formal pricing theory, making any interpretation somewhat subjective and complicating any attempt to articulate alternative explanations for these factor loadings. Therefore it was imperative to maximize the opportunities for these unarticulated alternatives to reveal themselves.

We began with supplementary regressions relating $COMP_1$ and $COMP_2$ to a set of demand shifters, *DS*, and a set of variables affecting the feasibility or desirability of price discrimination, *PD*:

$$\begin{aligned} \text{COMP}_1 &= \alpha_1 + \beta_1 \text{DS} + \gamma_1 \text{PD} + \lambda_1 \text{X} + \epsilon_1 \\ \text{COMP}_2 &= \alpha_2 + \beta_2 \text{DS} + \gamma_2 \text{PD} + \lambda_2 \text{X} + \epsilon_2 \end{aligned} \tag{9.5}$$

where X includes controls. The vector *DS* should affect the first component, the vector *PD* the second component. Sure enough, the traditional null $\beta_1 = 0$ was rejected in the first equation, and the null $\gamma_2 = 0$ rejected in the second.

That helped rule our interpretation *in*. Now we needed to rule alternatives *out*. So next we tested the significance of the estimates of the two remaining coefficient vectors, β_2 and γ_1. Under our interpretations, these vectors should be zero, but under alternative explanations, who knows? Fortunately, we failed to reject the standard null in both cases. Then we ran the same principal component analysis on the same set of prices in two other contexts: professional football, where the same pricing patterns were expected to obtain and did, and competitive wholesale markets for these commodities, in which they were not and did not. All this stuff gave countervailing findings four opportunities to appear. That they didn't diminished the chance that our conclusions resulted from fortunate happenstance and increased the chance that they were genuine.

By now this kind of thing is low-hanging fruit. It is de rigueur to conduct and present falsification tests in several fields of applied microeconomics (though much less common in others). But Genesis reminds us that low-hanging fruit can be dangerous. Here the pitfall is that it can be too appealing to emphasize what the estimated relationship *is not*, rather than what it *is*. It is easy to see how this can happen. We are taught to worry, more than anything, about bias, which is a toxin that must be purged or proved absent. Well, bias is a reasonable thing to worry about. So is the fact that there may be many possible explanations for the phenomenon of interest. Eliminating some doesn't automatically make yours correct. We saw this in Card and Krueger's (1994) minimum wage study, which examined the robustness of its main findings quite thoroughly, yet couldn't overcome questions of internal and external validity. Thus a successful falsification test or set of robustness checks is usually less convincing than a successful test of a refined null hypothesis or a second independent prediction of your theory. We will see this in the next subsection, where a Card and Krueger-style-study with few robustness tests to offer embellishes its credibility using our third technique: expand the predictions.

9.2.3 *Expand the Predictions*

Testing alternatives is akin to subtraction: when an alternative theory is ruled out, some of the probability that would have been ascribed to it is now assigned to your theory instead. Testing multiple predictions of a theory is akin to multiplication, which is far more powerful. The chances of an alternative theory matching any one prediction of yours might not be all that small, but the chances of it matching all of, say, three independent predictions is far smaller.

How many predictions should you test? Ideally, all that are central to your case. As we know, the purpose of a model, at least one with causal content, is to explain the phenomenon of interest. If so, it should make a prediction about every essential aspect of that phenomenon (possibly in conjunction with other related theories or factual support), and all of these predictions should be accurate (statistical power permitting). Otherwise, your model can't explain an integral aspect of the phenomenon of interest. If so, the chances of that model being valid are awfully small.

It is fair to say that we often cannot achieve this ideal, for various practical reasons, but equally fair to say that we often do not try. Perhaps this is because we so focus on a primary regression coefficient or set of coefficients that we neglect to think outside this box. Perhaps it is because our profession places such weight on formal testing that an informal test–based on a judgment of what is reasonable rather than a test statistic–seems trifling in comparison. Either way, these are restrictions that we impose on ourselves, to our detriment.

There are five general ways to expand the predictions. It is rare that all five will apply to any given study, but equally rare that none will. Here they are.

1. *Impose Less, Test More.* Remember, your econometric model is a container to be built, not a structure to be handed down from on high. This principle is violated when a model imposes meaningful structure that could be tested but isn't. Sometimes ways to test this structure are easy to spot, as with the timing that was imposed in Goldin and Rouse's orchestra study in the previous chapter. Sometimes they are harder to spot, as with the voting studies discussed above. Equation 9.4 asserted that the relationship between turnout, M, and S took a particular functional form. Rather than accept this structure as given, we tested it, via the nonparametric model in Fig. 9.4.

2. *Go Out of Sample.* We all know that accurate predictions are much easier to generate in-sample than out-of-sample. The former are optimized by construction: it is how the parameter estimates are determined. The latter are not, and can be sensitive to omitted variables or changes in context that are recognizable only through vernacular knowledge. For these reasons, trying to predict out-of-sample is far more humbling, with proportionately greater returns if you succeed.

Going truly out of sample tests not only the validity of your model, but also its versatility: how well it applies to altered circumstances. This is both a blessing and curse, since you can't necessarily tell what is at fault if it fails.

But you can also go out of sample without using another sample, in deed if not in word, by testing a prediction that your estimator did not try to maximize or did not have the information to make. This examines the model's validity alone. My previously-discussed analysis of the recycling problem, for example, estimated all relevant parameters but one, which was taken from a completely unrelated study. This one, the rate of growth in aluminum demand, let me predict when secondary's market share would reach a steady state. This corresponded with the actual onset of the steady state in the data, providing "external" confirmation of the model.

3. *Test Other Implications about the Key Dependent Variable.* Economists are pretty good at this. There is no reason to test only the key implication of interest, if other meaningful predictions can be tested as well.

Tucker et al. (2013) wish to see what happens when home sellers cannot re-list their house, which "sets the clock back to zero," disguising how may days it has been on the market. Utilizing Card and Krueger's difference-in-difference methodology, they compare house prices in Massachusetts, which instituted a policy prohibiting re-listing, with those in Rhode Island. The authors view days-on-market as a quality signal: more is less. Consistent with this interpretation, when days-on-market is recorded accurately, the price of homes that have been listed for a long time falls, while the price of those that have been listed for a short time rises. The authors then reinforce this conclusion with additional tests of their theory, showing that this effect occurs primarily where house quality is more uncertain and where house sales are not sluggish (which would weaken the quality signal). Even without robustness tests, which are largely absent from this paper, this combined evidence is reasonably convincing.

4. *Show How We Got There from Here.* A theory with causal depth will outline the pathways that link cause and effect. When only one pathway exists, or dominates, we can multiply the predictions by exploring the legitimacy of this pathway, as mentioned in Chap. 6. We will see an important example of this shortly.

5. *Test Implications about Other Dependent Variables.* There is often one key dependent variable of interest, on which we naturally focus. But a theory with causal breadth will make meaningful predictions about other parts of the system, and it is worth testing these too. For example, Card and Krueger examined prices and "non-wage offsets," such as free meals, in addition to employment—with mixed results.

Across these five ways to expand the predictions, we have re-encountered several familiar modeling concepts: the container, versatility, causal depth, causal breadth. This is not accidental: model building is not divorced from model testing. A good model maximizes its surface area, its exposure to reality. This is part of providing a description of observed phenomena that is intended to be taken seriously.

It also makes us more persuasive. The man on the street will rarely accept an idea because of a single piece of evidence in its favor. Such evidence would need to be virtually incontrovertible–the proverbial smoking gun. In its absence, we generally need more, an accretion of evidence, which can be obtained by making multiple predictions and confirming them.

Application to Abortion and Crime. This issue is at the heart of Donahue and Levitt's (2001) paper linking abortion and crime in the U.S. Perhaps you've heard of it. The FBI's Uniform Crime Reports show a dramatic reduction in violent crime beginning in the early 1990s, which continued throughout the decade (and up to the present). Donahue and Levitt claim that the nationwide legalization of abortion in 1973 explains as much as half of the 1990s decline.

Such a strong claim is inherently suspect, so it is imperative that the authors carefully link cause and effect. Theoretically, they do. They articulate two primary reasons why children born after 1973 could grow up to be less criminally-inclined. First, women may use abortion to have children later in life, when they can better care for them. Second, abortions are concentrated among mothers "most at risk to give birth to children who would engage in criminal activity." These then receive a scale analysis:

> Using the 1990 Census, we first determine the proportion of children born into each possible combination of three factors: white vs. black race, teenage motherhood, and unmarried motherhood. Using the estimates of Levine et al. (1999), we then project what those proportions would have been absent legalized abortion. To these changes in proportions we apply evidence on the frequency of crime in each category and account for the effect of a fourth factor: unwantedness.

This exercise, which is presented in detail, concludes that legalized abortion should reduce crime by at least 11%, which is about one-third of the full decline during the 1990s. Large effect sizes are, in fact, reasonable.

OK, you're halfway: now empirically confirm the viability of this link. So it comes as a disappointment to find that the empirical work, which relates an abortion measure directly to crime rates, does not do this. How did the composition of the young-adult cohort change during the 1990s? How did this changed composition affect crime? We never find out. We have reason to wonder. Everything about this scale analysis is fragile. There is "little direct evidence on the number of illegal abortions performed in the 1960s," so it is hard to simply impute how much legalization affected the actual number of abortions performed and, hence, the composition of the new birth cohorts. Furthermore, Levine et al.'s (1999) subgroup estimates for the 1973-legalizing states are fairly imprecise, while the effect of teenage motherhood and single parenthood on criminality is taken from Räsänen et al.'s (1999) study of Finland, a quite different place from the U.S.

As it turns out, there was other economics research about the crime decline floating around in 2001, whose rise to prominence was less…meteoric (Reyes 2015). Its hypothesis, grounded in the neurochemical and epidemiological link between early-childhood lead exposure and adult aggressiveness and criminality, attributed much of the decline in crime to the phasing out of leaded gasoline 15 years previously. Though more research is needed, time has been kind to this hypothesis so far, and Rick Nevin and others have documented many instances, widely varied across time and space, in which changes in childhood lead exposure are followed, with the appropriate lag, by same-direction changes in criminality.

Note that the lead-crime hypothesis implies something very different about the link-testing exercise I have advocated. Childhood lead exposure was greater among

poor, minority households, who then benefitted the most by its removal. This story is not about the composition of birth cohorts, as in Donahue and Levitt, but about changes in the criminality of particular groups *within* that cohort: blacks, teenage mothers, and unmarried mothers especially (the last two because they are more likely to be poor). The link-testing exercise should smoke this out.[5]

This is particularly important for the abortion story, because the lead-crime hypothesis also implies something else: that Donahue and Levitt's estimates are favorably biased. Their conclusion relies on the finding that the five states that legalized abortion a few years before 1973 also had earlier reductions in crime. As it turns out, these states cut their gasoline lead earlier as well (Reyes 2015, Fig. 1).

9.3 Check Yourself

By now it should be clear that hypothesis testing is far more than the formulaic implementation of a longstanding, formalized process, at least in the non-experimental world. Putting these ideas into practice is an art–wait, a craft–for which there is no simple formula. But there is an essential principle: put first things first. Recognize the central challenges to credibility, based on the nature of the problem that you are studying, and then structure testing to address them. Each example above was introduced in this way. Think of yourself as an explorer, a modern Ernest Shackleton, mapping the uncharted terrain between the standard null and your scientific hypothesis. Wander with a purpose.

If your theory is on track, the techniques in this chapter will help show this (Fig. 9.5). Rejecting a refined passive null rules out more space from consideration, making it more likely that your theory is the true alternative to that null. (Accepting an active null is even stronger, narrowing the eligible estimate space to a thin corona extending beyond the exact implications of your theory.) Rejecting alternative explanations for your findings reduces the overlap between the estimates that are consistent with your theory and those that are consistent with others, enhancing the chances that yours is correct. Falsification tests lighten the "cloud of doubt" that is Hypothesis F, which rains down on every study, to a greater or lesser degree. Multiplying the predictions gives you several shots at the bull's-eye–the set of

[5] To my knowledge, this exercise still has not been carried out, probably because of formidable data issues (see https://www.law.berkeley.edu/wp-content/uploads/2015/04/csls-workshop-raphael.pdf). Here, however, perfect data should not be made the enemy of good (or even mediocre) data. My read is that it would be possible to determine whether or not the evidence is broadly consistent with the causal links hypothesized by each set of authors.

A second prediction, about timing, would also distinguish between abortion and lead. The abortion story implies…not exactly a discontinuity, but a discrete change in the composition of births before and after 1973, and a discrete change in rate of crimes committed by those birth cohorts. Not so for gasoline lead, which changed more gradually and more heterogeneously across the states. Several figures in Joyce (2009) get at this indirectly, and do not support the abortion-crime hypothesis.

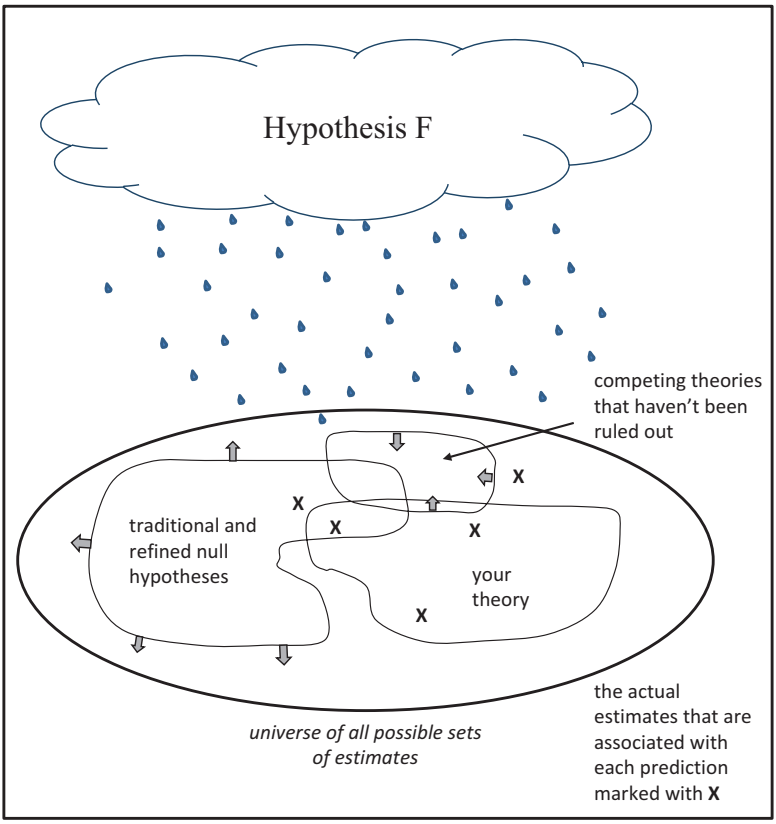


Fig. 9.5 Exploring uncharted territory by refining the null, expanding the alternatives, and multiplying the predictions

estimates consistent with your theory. The more that land, the more likely it is that your aim is true.

Of course, you don't know in advance whether your theory is correct. As with Shackleton, something could go wrong, and set back your analysis or your theory. As with Shackleton, you need fortitude—the willingness to be proved wrong and to seek out opportunities to let that happen. This is the ethical imperative of hypothesis testing: it is, ultimately, a way of policing ourselves. Our credibility depends on it.

A certain independence abets this kind of self-policing. Thus, companies' internal auditors, quality inspectors, and underwriters are organizationally walled off from everyone else, protected in their role of saying "no." Classical testing seeks this independence through rigidity: unyielding procedural strictures, high standards for statistical confidence, and careful sequencing–experiment first, testing second.

This all falls apart in the non-experimental settings common to economics, which allow for other sources of error besides sampling, and which intertwine estimation and testing. Opportunities for moral hazard are everywhere: to look for problems

where they can't be found, and vice versa; to adopt simplistic error structures that overstate experimental content and understate standard errors; to test only weak nulls that say little about your theory. Merely following protocol does not wish this stuff away. We must *seek out* opportunities to be proven wrong.

This is the larger lesson of Donohue and Levitt (2001). Step back and focus on the fundamentals: sparse experimental content. Almost everything I cautioned you about in Chap. 8 is here. Their analysis relies on just five "early-legalizing" states, four of which are in the Far West, and two of which are weighted far more than the others, though weights should hardly be used at all. How can such strong conclusions obtain from such limited experimental content? Yet here they are, un-derailed by a robustness check showing great sensitivity to weighting and a host of other issues (see Manzi 2012, pp. 110–117). Meanwhile, the essential question that begs an answer–does the causal chain work as anticipated?–hangs in the air.

Similar questions can be raised for many studies, both structural and reduced form. Is this really the only explanation for your findings? Does your model truly reflect the nature of the experiments at work? Does rejecting the standard passive null really say that much about your theory? When an opportunity to answer these questions goes unexploited, it is an omission that covers a multitude of sins.

Food for Thought

1. Chapter 6 discussed Gabaix's (1999) explanation for Zipf's Law.
   (a) Articulate a parametric active null hypothesis for the city size distribution implied by Zipf's law, and lay out a regression specification that would allow it to be tested.
   (b) Do the same for a non-parametric active null.
   (c) Articulate additional hypotheses that could be tested to examine Gabaix's explanation for Zipf's law more directly.
2. Hall's (1978) classic paper on aggregate consumption argued that this variable should follow a random walk. Today this hypothesis could be tested using some sort of Dickey-Fuller test. Does this test treat Hall's hypothesis as an active null, a refined passive null, or a traditional passive null? Do you think the data are sufficient to lend sufficient statistical power to such a test?
3. As mentioned in the chapter's rational voter discussion, even the "improved" hypothesis tests that are used are not conclusive by themselves. One last piece of evidence is needed to round out support for the theory. How could it be obtained? (One answer to this question draws on a relatively obscure U.S. constitutional amendment.)
4. For years the rational voting literature was engrossed by the endogeneity of P and T, which arises because larger turnout in close elections will itself lower the probability that any given vote is decisive. General equilibrium models that

accounted for this fact were deemed superior to partial equilibrium models that didn't, though this "feedback effect" is inconsequential in practice.

(a) Show that this is true using a simple scale analysis, which incorporates the magnitudes in Fig. 9.4 and the fact that P is inversely proportional to the number of voters.
(b) What does this "inconsequentiality" imply for the magnitudes of the two terms on the left-hand side of Eq. 9.3? What does it imply for the predicted *M-T* relationship in the partial and general equilibrium models (as in Fig. 9.4)?
(c) Given your read of statistical significance in Fig. 9.4 and the conclusions arrived at above, sketch out a Venn diagram like Fig. 9.3, circumscribing the sets of estimates consistent with (not too unlikely under) the null hypothesis, the partial equilibrium model, and the general equilibrium model.

5. The hypothesis-testing technique known as "randomization inference" asks the researcher to impute the distribution of test statistics that would obtain if the *treatment* were randomly assigned across experimental units. In this book, instead, I have advocated creating an error structure that respects experimental content, which includes a random effect at the level of the experimental unit, when it differs from the unit of observation in the data. Compare these two approaches to hypothesis testing. How are they similar? How are they different? Which approach is more general?

References

Card D, Krueger A (1994) Minimum wages and employment: a case study of the fast food industry in New Jersey and Pennsylvania. Am Econ Rev 84(4):772–793
Donohue J, Levitt S (2001) The impact of legalized abortion on crime. Q J Econ 116(2):379–420
Fanelli D (2010) "Positive" results increase down the hierarchy of the sciences. PLoS One 5(4)
Gabaix X (1999) Zipf's law for cities: an explanation. Q J Econ 114(3):739–767
Gelman A (2016) What has happened down here is the winds have changed. *Blog Post*. Available at http://andrewgelman.com/2016/09/21/what-has-happened-down-here-is-the-winds-have-changed/. Accessed 21 Sept 2016
Grant D (1998) Searching for the Downsian voter with a simple structural model. Econ Polit 10(2):107–126
Grant D, Toma M (2008) Elemental tests of the traditional rational voting model. Publ Choice 137(1):173–195
Hall RE (1978) Stochastic implications of the life cycle-permanent income hypothesis: theory and evidence. J Polit Econ 86(6):971–987
Ioannidis J (2005) Why most published research findings are false. PLoS Med 2:696–701
Joyce T (2009) A simple test of abortion and crime. Rev Econ Stat 91(1):112–123
Law MT, Kim S (2005) Specialization and regulation: the rise of professionals and the emergence of occupational licensing regulation. J Econ Hist 65(3):723–756
Lehrer J (2010) The truth wears off. New Yorker:52–57
Levine P, Staiger D, Kane T, Zimmerman D (1999) *Roe v. Wade* and American fertility. Am J Public Health 89(2):199–203

Manzi J (2012) Uncontrolled: the surprising payoff of trial-and-error for business, politics, and society. Basic Books, New York

Paldam M (2016) Simulating an empirical paper by the rational economist. Empir Econ 50(4):1383–1407

Räsänen P, Hakko H, Isohanni M, Hodgins S, Järvelin M, Tiihonen J (1999) Maternal smoking during pregnancy and risk of criminal behavior among adult male offspring in the northern Finland 1966 birth cohort. Am J Psychiatry 156(6):857–862

Reyes JW (2015) Lead exposure and behavior: effects on antisocial and risky behavior among children and adolescents. Econ Inq 53(3):1580–1605

Sass T, Saurman D (1993) Mandated exclusive territories and economic efficiency: an empirical analysis of the malt beverage industry. J Law Econ 36(1):153–177

Tucker C, Zhang J, Zhu T (2013) Days on market and home sales. Rand J Econ 44(2):337–360

Chapter 10
The Ends of Your Means

Abstract This chapter lays out the first steps to bringing closure to an empirical study. Above all, this involves pursuing "coherence," in which the study's findings, economic theory, and vernacular knowledge about the phenomenon of interest coalesce into a logically consistent, unified whole. The pursuit of coherence is multifaceted, and extends to the larger literature to which the study belongs. These ideas inform studies of the demand for cigarettes, zero tolerance drunk driving laws, The Great Moderation, and more.

All right, you have laid your groundwork, developed your model, gathered your data, run your estimates, tested your hypotheses, and drawn your conclusions. Time to write it all up!

Not so fast, partner. We're not quite done. We need to swing through the ball.

Swing through the ball. This phrase can be found in many sports, from baseball to golf to tennis. In each case, the physics are the same: you can't affect the ball's trajectory once it's left the instrument that strikes it. So what's the deal? It is that swinging through the ball changes your swing, makes it more true, so that you hit the ball with more power and accuracy. So it is here too. We've come too far to stop now, and shank the ball into the bushes. We need to swing through the ball.

How do we do this? In four ways: by focusing on results, not methods; pursuing coherence; connecting your work back to the literature that anchors your study; and understanding the problem you are studying on its terms, not yours. The first three are covered in this chapter, and the fourth in the next chapter.

10.1 Results, Not Methods

If we think of economics as pure science, we can be overly seduced by the seeming finality of our equations and estimates. Then it is easy to hang your hat on the soundness of your methods rather than the credibility of your results. But remember the opening chapter: the scientific method performs a miracle, reducing credibility

D. Grant, *Methods of Economic Research*, Springer Texts in Business and Economics, https://doi.org/10.1007/978-3-030-01734-7_10

to competence. Economics can rarely employ this miracle. It is a craft. Sound methods are necessary, but rarely sufficient, to establish credibility.

The techniques introduced in the previous chapters, like most techniques used in this profession, make us more methodical in this endeavor. Systems help us identify unanticipated effects and potential sources of bias. Scale and description help us understand the magnitudes of the forces at work, so that we can recognize when our estimates do and do not make sense. Linking theory and estimation more concretely, and better understanding the limits of our data, help ensure our reach does not exceed our grasp. Designing better hypothesis tests and incorporating vernacular knowledge can prevent us from drawing unwarranted conclusions.

Yet none of this reduces to a simple checklist. Part of the craft is knowing how much to apply these techniques, when you've done enough, and when you must utilize or improvise other ways to achieve these aims. This theme has been present throughout this book: the primacy of results over methods. Our definition of a model stressed not its form, components, or construction, but its purpose. Data description was not a mechanical act of reporting, but a way to set up and support estimation. Hypothesis testing was not a rote procedure, but an exercise in discovery. I will adopt a similar perspective when discussing the literature later in this chapter.

Thus, it is easier to understand what *to do*, in this regard, as the opposite of what *not to do*: check all the right boxes, and nothing more. Such analyses often have a perfunctory quality, an air of going through the motions, such that evident questions go unaddressed or unconsidered, or apparently questionable assertions or findings go insufficiently explored. As a result, these analyses lack credibility. Their results aren't guaranteed to be wrong–yet you cannot be confident that they are right.

A sterling example of this type of problem is Becker et al. (1994). This paper examines an unusual property of demand: addiction, widely believed to apply to select goods such as cigarettes, opioids, or cocaine, but not to most goods generally. It assumes that current period utility depends on current and past consumption of the addictive good. Intertemporal maximization then implies that this good's current period demand depends positively on both its past and anticipated future consumption. This leads to the following simple relationship:

$$C_t = \theta_0 C_{t-1} + \beta\theta_0 C_{t+1} + \theta_1 P_t + \theta_2 X_t + \theta_3 \xi_t + \theta_4 \xi_{t+1} \tag{10.1}$$

where C is consumption, P is price, ß is the discount factor, X contains controls, ξ represents unobserved "shift variables," and the θ's are coefficients. Under "rational addiction" θ_0 is positive, θ_1 is negative, and the ratio of the coefficients on future and past consumption equals the discount factor–a valuable, "semi-sharp" active null.

The main problem is that the error term, $\theta_3\xi_t + \theta_4\xi_{t+1}$, is clearly serially correlated, and thus correlates with past and future consumption. For this reason, these independent variables are instrumented with various combinations of past and future prices. With minute p-values, the estimates strongly support addictive demand for the market studied, cigarettes.

It's a nicely organized, well-written, rather elegant paper, and when I first read it, I didn't believe it for a minute. These correlations are the central issue, as the authors recognize. Their way of addressing it is reasonable, but it is not fail-safe, and it is not all there is. The case for their model is not closed.

In fact, it is worse than that. There's a tell, as they say in Vegas, and I've already told you what it is. The estimates are too good.

The authors analyze a 31 year panel of the 50 U.S. states. The state and year fixed effects included in X sweep out most of the variation in the dependent variable, so high degrees of significance are not to be expected. Indeed, I have read countless studies using similar panels and rarely observed large t-statistics. Yet here, those on past consumption range from 9 to 30, those on price average about 10, and those on future consumption average about 5 (their Tables 2 and 3). These values are unreasonably large. The easiest way to get them is from interplay between X and ε on the scale of motion of X—in other words, serially correlated errors that are related to serially correlated independent variables. (The presence of instruments complicates this story, of course, but Auld and Grootendorst show that its central point holds true.)

There are at least three ways to probe this issue further. Swinging through the ball requires employing one or more of these methods (even if the t-statistics hadn't been so large). The first would be to add to the *one* explicit demand shifter this study controls for, income, thereby reducing the magnitude of the error term and its serial correlation. The second would be to explore the extent of bias using a simple theoretical model or Monte Carlo simulation. Neither need precisely suit the scenario analyzed in the paper, just give a sense of magnitudes, as a scale analysis of sorts. The third would be to apply the model to goods that are not addictive, as a falsification test. I remember discussing this paper with a colleague in the late 1990s, who wanted to see these methods applied to milk. He suspected they would indicate that it was addictive.

A few years later, he got his wish. Auld and Grootendorst (2004) used the second method listed above to demonstrate that this sort of instrumenting does not effectively ameliorate this problem. Then they show that the same approach finds milk, eggs, and oranges (but not apples) to be similarly "addictive"–a result at odds with common sense. Estimates of cigarette addiction fall roughly in the middle of the range of values observed for these other four goods.

In the end, the original analysis, by failing to adequately allay concerns about its central empirical issue, was not credible. Auld and Grootendorst showed it was, in fact, strongly biased–despite thoughtful implementations of accepted methods of addressing its central issue.

Professional convention is useful insofar as it embodies the acquired knowledge and practical experience of previous problem-solving efforts. But we must also think outside this box. Ultimately, we must employ the analytical framework and supplementary tests that best serve the purpose at hand, not merely those that are conventional. To swing through the ball, do not focus narrowly on the suitability of your methods. Maintain a far-sighted view of the credibility of your results.

10.2 Coherence

Then look one step beyond. Your results are themselves a means to a yet further end: the ultimate conclusions of your research. These often are–and, following Chap. 9, should be–based on a compendium of related findings. Ideally, these findings, economic theory, and any relevant vernacular knowledge should coalesce into a logically consistent, unified whole. They should *cohere*. If they do not, it is important to investigate why, and explain or reconcile any dissonance as well as you can. It's your study. You have agency, you have ownership, you have the data, you have comparative advantage. Who better than you?

Coherence is the logical terminus of the idea that credibility is based on the whole assemblage that is your research, not just a few estimates or p-values. Many of the concepts we have discussed further this goal. Systems, causal breadth, and causal depth help us understand how different findings might relate to each other. Scale, description, and vernacular knowledge help us figure out whether our estimates are reasonable. Expanding the predictions and amplifying the alternatives help us inspect the nooks and crannies of our data, to see if we find anything untoward. If a model or empirical specification is correct, we should see this in all its facets, not just a few (statistical power permitting). The concept of coherence takes this injunction seriously.

Table 10.1 Survey evidence on the effect of zero tolerance laws on drinking, the location of drinking, and driving after drinking (in percentage points, with estimates not significant at the 10% level placed in parentheses)

Dependent variable	Theory	Carpenter (2004), Table 1	Liang and Huang (2008), Tables 2 and 5	Wagenaar et al. (2001), Table 2
Drinking				
Overall	–	(−2)	(0)	(4)
Heavy	0 or +	−17	−3	(−1)
Moderate	–			
Driving after drinking				
Overall	–	(−5)	−16	−19
Heavy	0 or +		−10	−23
Moderate	–			
Drinking by location				
Home	0 or +		(−4)	
Away from home	–		−7	
Fatalities (*average of all U.S. studies cited by the paper*)	–	≈ −14%	≈ −17%	≈ −13%
Data, survey period		Behavioral risk factor surveillance system, 1984–2001	College alcohol surveys, 1993–1999	Monitoring the future, 1984–1998

This issue permeates the literature on zero tolerance laws, passed in most U.S. states during the 1990s, which outlaw driving after any amount of drinking for "youth" under 21. While some studies directly examined these laws' effects on traffic fatalities, three panel, survey-based studies examined the behavioral response more finely–and advertised themselves as such. These studies' basic findings are depicted in Table 10.1. Each obtains a statistically significant, double-digit effect on either drinking or drinking and driving, but not both.

These papers use somewhat different methods and data, but I have no beef with any of that. It is fine. Yet within each study, results conflict with other results, with economic theory, or with the posited causal chain by which these laws improve traffic safety: a trifecta of consistency problems.

The most obvious and most important of these involves theory. Driving after heavy drinking was already illegal. So zero tolerance laws should reduce overall drinking and moderate drinking, but not heavy drinking–which could even increase, following the logic of "I'd rather be hanged for a sheep than a lamb."[1] In every study, however, the estimates indicate a large reduction in heavy drinking, driving after heavy drinking, or both. Only one of these three studies acknowledges this anomaly. To its credit, it offers a possible explanation, suggesting that heavy drinking falls because these drinkers are now sure to exceed the legal limit, even if some time elapses before they drive, and provides some indirect evidence in its favor. This evidence is not terribly convincing, but it is better than nothing, and possibly the best that could be done under the circumstances.

A second anomaly regards intra-study conflicts in the estimates. In the first study, a large decrease in heavy drinking is coupled with a tiny decrease in drinking overall, though the latter is only twice as frequent as the former. In the second, a nil effect on drinking overall is coupled to negative point estimates for drinking at home and drinking away from home. This anomaly is less important, however, because the estimates aren't all that precise.

The final anomaly was noticed by no one, perhaps because the estimated effects that are significant are quite large, and perhaps because of inattention to scale. The penultimate row of the table lists each study's best guess of how much zero tolerance laws reduced youth traffic fatalities, based on its citations of the literature. In each case, this value is about 15%, justifying the enthusiasm of the first two studies, especially, about these laws' efficacy. In this age group, however, only about one-third of traffic fatalities involve alcohol. Thus, to generate fatality reductions this large, drinking, or driving afterwards, must fall by nearly half. Conditional on the assertion that zero tolerance laws are highly effective, the estimates in Table 10.1 are *far too small*.

We haven't even made cross-study comparisons yet, and already these studies raise as many questions as they answer. Swinging through the ball entails reconciling these contradictions to the extent that you can. For the first and third anomalies, especially, the simplest place to look would be the distribution of blood alcohol

[1] If you are going to be caught driving after drinking, you might as well drink a lot and really enjoy yourself.

among drivers involved in fatal accidents. This distribution is exactly what these laws target, removing definitional issues that these micro studies confront (and I have underplayed), and is known fairly precisely (because there so many fatal accidents). Does its change mirror that observed in the micro data?

By happenstance, I already knew the answer to that question. In early 2001, curious about the potential perverse consequence of an increase in heavy drinking, I pulled a few years of data off the internet and plotted the BAC distribution of young drivers involved in fatal accidents. It was remarkably stable across time, both in states that had adopted zero tolerance laws and states that didn't. This was shocking. Given these laws' focus on moderate drinkers and their large estimated effects on fatalities, the distribution would have to change dramatically. Something did not add up. I decided to figure out what it was.

10.3 Connecting Your Results to the Literature

A study typically contributes to a broader literature on that topic. This too is context, and your work should be placed within it: both at the beginning, when justifying your study, and at the end, once you have arrived at its findings.

Once again our textbooks lead the way. Search them for distillations of the state of knowledge on the most fundamental empirical issues and you will often find a hapless silence, an "illustrative estimate" or two, or a somewhat arbitrary rendering of disparate studies. Search even for the general magnitude of effect, and you will often search in vain. Diseconomies of scale: when do they occur, how often, and why? Physician-induced demand: is it large or small, based on income targeting or relative prices? Wage gaps, based on gender, race, or ethnicity: how much is due to discrimination? Market concentration: how much does it affect price-cost margins, and how much does product heterogeneity affect this relationship? How big are pollution externalities? The message of all this is that we don't take the research in our own field very seriously. Perhaps I'm not the only one with concerns about credibility.

This problem goes beyond textbooks, into our research. It shows up, in mild but typical ways, in the zero tolerance literature we just discussed. Two of the three studies in Table 10.1 claim that the evidence that zero tolerance laws improve traffic safety is favorable or "overwhelming." Of the nine citations to U.S. work in support of that claim, one third are to papers that conclude no such thing.[2] And of all the estimates listed in Table 10.1, economists (explicitly or implicitly) cite one far more

[2] One of these papers, Hingson et al. (1989), finds that youth fatalities fall by 20% in a state that recently adopted a zero tolerance law, but also by a similar amount in a control state, and concludes ambivalently. For those studies in Table 10.1 that cite this paper favorably, this 20% estimate is used in calculating their "fatality" entry in the penultimate row of the table, since this row's purpose is to compare each study's findings to its own literature-based assessment of zero tolerance laws' effects on fatalities.

than any other: the one that is least representative of all, the large reduction in heavy drinking, unmatched by anything comparable in its row, column, or panel of the table.

It shows up with rational addiction, where even in recent years Becker et al.'s (1994) paper is cited far more than those studies disputing it, only one of which is mentioned above.

It shows up with the point-shaving literature that we discussed in Chap. 4. Wolfers' work generated not just one rebuttal–it spawned legions of them.[3] Borghesi followed up his 2008 critique by arguing that the asymmetry of point differentials around the point spread, which so disturbed Wolfers, merely represented appropriate game management by the winning team (Borghesi and Dare 2009). Bernhardt and Heston (2010) agreed, replicating this asymmetry in games without betting lines or otherwise not subject to point shaving. Gregory (2018, available as a working paper in 2011) and Diemer and Leeds (2013) draw similar conclusions, while Johnson (2009) and Paul and Weinbach (2011) raised other objections to the original study. Only an unpublished master's thesis stood in opposition to this tide.

Yet it only mildly exaggerates to say that the only studies citing any of this research are other studies disputing Wolfers' findings. Of those papers that have cited Wolfers in the last 5 or 6 years, only a handful acknowledge any of these objections.[4] The problem is worse at higher levels. Of the ten most-prestigious recent Wolfers-citing studies, only one cites any paper listed above: a review in the *Journal of Economic Literature* (Zitzewitz 2012). This gem ignores everything but Wolfers, the master's thesis, and Bernhardt and Heston, wrapping these three studies into an ambiguous ball with a classic he said/she said:

> Wolfers finds an asymmetric distribution of outcomes: heavily favored college basketball teams are much more likely to win by just less than the point spread than by just more. He concludes that "six percent of strong favorites have been willing to manipulate their performance."
>
> Bernhardt and Heston (2010) challenge Wolfers' interpretation of his result, because they find a similarly asymmetric distribution of outcomes for college basketball games on which bookmakers did not offer betting. They argue that Wolfers' result may have a more innocent explanation, such as favored teams with a safe lead reducing their effort level. While Bernhardt and Heston's failure to find a particular cross-sectional result may be due to inadequate statistical power,[5] their paper provides an example of the importance of documenting both positive and negative results.

Really–who knows? What is truth, anyway?

[3] I found these by typing "point shaving" into Econ Lit–that's all. With two minor exceptions, everything that came up that wasn't Wolfers' paper was a rebuttal to it.

[4] Ironically, this handful excludes Bernhardt and Heston, though they appropriate two of Borghesi's arguments in their paper. Gregory (2018) isn't much better.

[5] I don't know what this phrase is referring to. Bernhardt and Heston analyze nearly 45,000 basketball games, a large sample, even when it is cut different ways.

These cases are not unusual. I could easily add others just from the articles discussed in this book.[6] They are also not subtle, asking only whether the profession's citation habits reasonably summarize the evidence and identify, if not quarantine, studies with serious flaws. The whole exercise has a cast of hopelessness, as if it doesn't really matter what the literature says, since we can never really know what is correct anyway, and it doesn't matter if we do.

Thus, we often address the literature as a set of boxes to check, egos to stroke, a launching pad for our own study: a laundry list of preceding work, with some discussion of its findings, followed by a description of how our study is different, or better, or fills some gap. (Many seminars start out this way too–the surest way to kill interest in your talk before it ever gets going.) We double down on this fatalism.

The antidote is not to think about the literature as a list of studies. It is neither: not a list, not of studies. It is data–data to describe and to explain. Economists know a lot about dealing with data. For this reason, treating studies as data is the best way to connect your findings back to the literature they contribute to.

Meta-analysis takes this admonition literally. Studies vary by their empirical methods, type of data, data quality, data quantity, sample period, geographic location, dependent variable(s), specification, and institutional context. Meta-analysis treats these as independent variables, the estimated effect size as the dependent variable, and relates the two. Studies are literally data. But we don't have to use meta-analysis to benefit from this perspective.

10.3.1 Describing the Literature

The first step toward connecting it back should be taken when surveying the literature at the beginning of your study, by displaying the variation in its findings, identifying the major factors that explain this variation, and showing how these factors and findings evolve over time. A list rarely does this well, so use methods that can. Since studies are data, the natural alternative is description, discussed in Chap. 7.

One option is a table, such as Table 10.1. Above it was used to highlight internal inconsistencies, where one finding of a study conflicted with another of its findings or with theory. Then we had only to look within columns. But see how much power you gain by looking across columns as well! Doing so reveals how extraordinarily ambivalent this literature is, with cross-study inconsistencies matching the internal inconsistencies, and more point estimates disagreeing with theory than agreeing with it, whether or not statistical significance is considered. If you needed to motivate a new study on this topic, you could do much worse than to present and discuss this table.

Tables can summarize larger literatures too, grouping studies by method, for example, or by the type of data used. Perhaps this explains why I think of the literature as a tapestry, multiple findings of different origins woven together.

[6] And did, and removed them for space considerations.

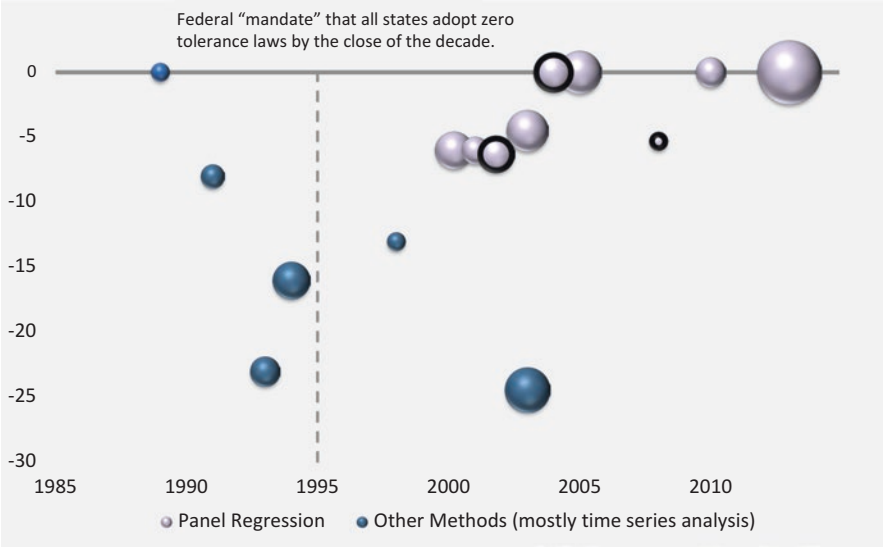


Fig. 10.1 Plot of the zero tolerance literature, circa 2013 (the implied effect on fatalities, in percent, by year of publication). Note: Each bubble's volume is proportional to the frequency of its annual citations, as of late 2016. The implied effect on fatalities in studies of micro survey data is taken to be one-third its estimated effect on overall driving after drinking. Insignificant estimates are set to zero. Circled studies utilize micro survey data, the rest tabulations of crashes or fatalities

Graphs can also work. Figure 10.1, for example, illustrates the broader zero tolerance literature this way. At a glance, multiple dimensions of each study are conveyed: timing, citations, findings, methods (in color or shades of grey). It is easy to see relationships, especially the collapse in the estimates brought on by lengthening the sample and shifting to panel regression methods. Graphs like this are ideal for showing the evolution of a literature, which can serve as the lynchpin of your literature review.

Any literature of size can vary along all of the dimensions listed above, but typically a few will matter most. Looking back over the various literatures I have contributed to, in some the key factor was the sample period, in others the empirical methods, in still others the type of data. In describing the literature, bring that out, and then use this relationship to show why some studies are superior to others, demonstrate the dependence of effect on context, or reveal what the literature is converging towards: a forecast of sorts.

For zero tolerance laws, we have already been offered such a forecast, if we only knew to look for it. The decline in estimates in Fig. 10.1 is, in fact, what the studies in Table 10.1 collectively presaged all along. Any reasonable average of its estimates, scaled by the fraction of accidents involving alcohol, would indicate that zero tolerance laws lower fatalities by well under 10%; the atheoretical and conflicted nature of the findings suggest that even a null effect should not be ruled out.

These studies billed themselves as illuminating the mechanisms by which these laws worked. The bigger finding, that they did not work as well as advertised, went unrecognized.

There are, then, three ways of reading Table 10.1. Internally, you want to recognize inconsistencies in any given paper's findings, and reconcile them as best you can. This requires looking vertically within the table. Externally, you want to relate papers' findings to each other, and explain any differences. For this, look horizontally. Then, with that knowledge in hand, you should step back from the details, abandon any functional fixedness resulting from excessive econometric absorption, and take in the whole, the forest, not the trees: what the aggregated evidence, despite its flaws, differences, and inconsistencies, is trying to tell us.

10.3.2 Explaining the Literature

Once your results are in hand, they must be connected back to the literature, woven into the tapestry. At this point, the literature becomes data to be explained. Your model is intended to explain actual phenomena. Other studies' findings are phenomena. Explain them!

There are three alternatives. The weakest, better only than nothing, is to say "my study is different in this way, so that might account for the difference in findings." That is correlation, not causation, far inferior to putting this speculation to the test. In Table 10.1, the third study uses cruder methods than the first study, and it must have been tempting for Carpenter to claim the priority of his results on that basis alone. That would have been wrong. When he re-estimated his coefficients using the simpler method, they were unchanged.

A step above this, better but not ideal, takes us from "might" to "could." This is to show that one set of models yields findings that cannot be distinguished from those of another set of models. In essence, you are arguing that the theory or empirical test (but not necessarily the coefficient) is weakly identified. We saw this once before, in Chap. 6, where Choi et al. argued that various theories of the underwriting cycle were weakly identified, or worse, in the short run. That was theoretical non-identification. What now follows illustrates empirical non-identification.

There are two competing explanations for the recent 25-year stretch of U.S. macroeconomic stability known as "The Great Moderation": good (monetary) policy and good luck. People have compared these explanations using VARs that estimate the responsiveness of interest rates to output and inflation, estimating the magnitude of innovations to all three in the process. Several studies found that the innovations diminished during The Great Moderation, while the policy responsiveness of interest rates was unchanged, and concluded that good luck was at work.

Benati and Surico (2009) point out that this conclusion is too hasty. Using only an improvement in monetary policy, they replicate The Great Moderation in a hypothetical New Keynesian economy, and then apply VARs to this simulated data. These say good luck is responsible:

> Our analysis suggests that key pieces of existing VAR evidence are difficult to interpret, and are, in principle, compatible with the notion that policy may have helped foster the greater macroeconomic stability of The Great Moderation. In fact, a change in monetary policy can have its largest impact on the innovations, with a milder effect on the VAR coefficients. This implies that the good policy and good luck explanations are almost observationally equivalent.

The VAR test is, at best, weakly identified. It can't clearly distinguish good policy from good luck.

The final option is the best of all: to go from "could" to "does," and explain other studies' conclusions by replicating their methods or data and seeing how the results change. Ideally, or at least most conveniently, you do this with the same model you used to reach your own conclusions.

This occurred in my own work on zero tolerance laws, the project begun in 2001 (Grant 2010). It evolved into a panel study that examined these laws' effects on fatalities. As Fig. 10.1 shows, previous panel fatality analyses had concluded that these laws reduced fatalities by around 5%, far below the ≈15% figure in Table 10.1, but sizeable nonetheless. I concluded instead that these laws had no effect at all. This difference deserved to be reconciled, but for longer than I care to admit, I simply didn't worry about it.

My dependent variable was youth nighttime fatalities, as daytime accidents rarely involve alcohol, and the law applies only to youth. But I also estimated my model on fatalities among adults at night and among youth during the day, treating both as control groups. The estimates on the zero tolerance law coefficient in these three regressions looked like this:

	Nighttime fatalities	Daytime fatalities
Youth (under 21)	−5%	−5%
Adults (21 and over)	−5%	????

These findings testified doubly to bias, from omitted variables that reduced accidents at nighttime and among youth, and implied that the true effect was nil. What more do you need?

Well, the answer is staring at you in that table, isn't it? It turns out, once I eventually looked into it, that previous panel fatality analyses had combined day and night accidents, and simply compared youth to adults. So, a few more regressions, and voilà:

	Nighttime fatalities	Daytime fatalities	All fatalities
Youth (under 21)	−5%	−5%	−5%
Adults (21 and over)	−5%	5%	0%

When the data was cut the same way, in the final column of this table, my specification's results matched those of previous studies. Cut it more finely, as I did, and you draw a different conclusion. The difference in methods *does* account for the difference in findings.

10.3.3 Coherence–Again

By now, you can probably guess what your goal in explaining competing findings should be: coherence. You want to make the assemblage that is the literature coalesce into a logically consistent, unified whole. Thus, after obtaining the findings above, my work was not done. I also pointed out that these findings were consistent with the stability of the BAC distribution across time, and that they made the findings of the micro studies in Table 10.1 somewhat less surprising. We aren't all the way home, but we're closer than when we started.

This goal is a lofty one, I know. There is no guarantee that you will be able to explain all competing findings, or even any of them. Variation in data, research methods, and institutional context often causes results to fluctuate, unaccountably, from one study to the next. This is intrinsic to the decentralized knowledge-gathering approach of academia, and should be considered a feature, not a bug. Furthermore, some outcomes may simply be quite variable–this is often a prediction of game-theoretic models, for instance–in which case the appropriate aggregation of a set of studies may simply say exactly this. But Chap. 9's discussion of the reliability of studies' findings suggests that there will often be grist for the mill. A study that is flawed, whose findings are wrong, doesn't add to knowledge, but subtracts from it. Identifying problems with these studies is as valuable as enhancing the credibility of your own.

To have a literature that is less than the sum of its parts, full of conflicting, unresolved findings, is so common that it is tempting to accept it as normal. Then it is easy to add to the confusion yourself, with a superficial discussion of why your findings differ from others'. It may be common, but it is not normal. It is not the natural order of things. It is deviant, a fact we have become inured to by the frequency with which it occurs. Something is going on out there in the data. Something is happening. Its footprint should be everywhere, in all the studies on that topic, obscured a little in some, more in others. A collection of studies that amounts to Who Knows? is sometimes realistic, but sometimes a premature admission of defeat. Begin your study expecting to figure out what that something is.

I once reviewed a paper in another youth-related traffic safety literature. For this particular phenomenon, the logic was clear cut: A should lead to B, which should then lead to both C and D. This paper examined the connection between A and D, and found an effect of a given size. It dutifully cited the other papers in this literature, which had examined other links–A to B, A to C, B to C–without ever noticing that these findings all agreed not just in sign, not just in magnitude, but numerically (within a reasonable tolerance). That is, the effect size of A to B, multiplied by that of B to C, equaled that of A to C, which equaled that of A to D–even though each estimate came from a different study or couple of studies. An extraordinary consonance of findings, just sitting there waiting to be noticed.

We will never discover this if we do not expect to discover it. It will rarely be so easy or so obvious, but it should never, ever be unexpected. It's what we got into this business for in the first place.

Food for Thought

1. Governments use estimates of the Value of Statistical Life (VSL) to allocate resources that increase the public's safety. These estimates typically come from wage studies, by extrapolating from the compensating differential associated with more dangerous jobs. If people would sacrifice $1000 in earnings to avoid a 1:10,000 risk of dying on the job, then the VSL = $10 million.

   Unfortunately, this approach has a consistency problem: other job characteristics are not well controlled for, and the labor literature–to the extent it even tries any more–has a difficult time uncovering compensating differentials associated with job disamenities generally (see Smith 1979, the most recent focused review I could find). How should studies of the VSL attempt to deal with this nettlesome problem? What about reviews or assessments of the VSL literature? Take a look at recent studies, meta-analyses, or literature reviews and see what has actually been done.

2. Given the facts highlighted in the two real-life scenarios listed below, how should you summarize these literatures in the way that will be most useful for policymakers?

   (a) The minimum wage in the United States has fluctuated, in real terms, considerably over its lifetime. While the federal minimum is currently on the low side, historically, state minima are gradually being raised to $15/hr. in California and New York, and similar increases are being considered elsewhere. This minimum wage would be far above the historical norm.
   (b) The evidence on the effectiveness of charter schools is fairly mixed, as are the characteristics of charter schools and the state laws under which they operate. Some states, such as Texas, where charter schools are quite limited in number and scope, are considering legislation that would open up this market substantially.

3. The *Journal of Economic Surveys* is all about summarizing the literature; you name it, they have an article summarizing it. Several such articles use a common literature-listing technique that I call a "tablelist"–a table that just lists articles, their characteristics and findings, one after the other. Recent examples include Abedifar et al. (2015) on Islamic finance, Becker (2015) on public policy's effect on private R&D investment, and van Ours and Williams (2015) on the labor market and health effects of cannabis use. Choose a tablelist from one of these articles or any other, and create a figure that summarizes the literature effectively.

4. Anderson et al. (2017) conduct a meta-analysis of studies analyzing the effect of government spending on income inequality. An abbreviated version of their main results (Table 3, column 3) is as follows:

$$Y = -0.01 + \mathbf{0.63}SE + 0.01DEVELOPED + 0.01PUBLISHED - \mathbf{0.10}OLS + \mathbf{0.07}TAX + 0.02TRADE + \mathbf{0.05}GOVERNANCE - \mathbf{0.10}INFLATION + 0.01EDUCATION - \mathbf{0.10}SOCIAL + error$$

where Y is the estimate of government spending's effect on the Gini coefficient, and the independent variables are as follows: the standard error of the relevant coefficient estimate; dummies for whether an economically developed country is analyzed, whether the study was published, and whether the estimator was OLS; measures of tax revenue, trade policy, the type of governance, and inflation; and dummies for whether the spending measure was education spending or general social spending. Significant estimates are in bold.

Based on these results, is there evidence of publication bias? Do concerns about the estimation method and potential omitted variables need to be taken seriously? What policy implications can be taken away from these findings?

5. My analysis of the "recycling problem" has arisen in Chaps. 3, 8, and 9 of this book, each of which contains an anecdote about one facet of the study. Assemble these facets to show the cohesive nature of that study's findings.

References

Abedifar P, Ebrahim SM, Molyneux P, Tarazi A (2015) Islamic banking and finance: recent empirical literature and directions for future research. J Econ Surv 29:637–670

Anderson E, D'Orey MA, Duvendack M, Esposito L (2017) Does government spending affect income inequality? A meta-regression analysis. J Econ Surv 31(4):961–987

Auld MC, Grootendorst P (2004) An empirical analysis of milk addiction. J Health Econ 23(6):1117–1133

Becker B (2015) Public R&D policies and private R&D investment: a survey of the empirical evidence. J Econ Surv 29(5):917–942

Becker G, Grossman M, Murphy K (1994) An empirical analysis of cigarette addiction. Am Econ Rev 84(3):396–418

Benati L, Surico P (2009) VAR analysis and the great moderation. Am Econ Rev 99(4):1636–1652

Bernhardt D, Heston S (2010) Point shaving in college basketball: a cautionary tale for forensic economics. Econ Inq 48(2):207–210

Borghesi R (2008) Widespread corruption in sports gambling: fact or fiction? South Econ J 74(4):1063–1069

Borghesi R, Dare W (2009) A test of the widespread-point-shaving theory. Finance Res Lett 6(3):115–121

Carpenter C (2004) How do zero tolerance drunk driving laws work? J Health Econ 23:61–83

Diemer G, Leeds M (2013) Failing to cover: point shaving or statistical abnormality? Int J Sport Finance 8(3):175–191

Grant D (2010) Dead on arrival: zero tolerance laws don't work. Econ Inq 48(3):756–770

Gregory J (2018) Do basketball scoring patterns reflect illegal point shaving or optimal in-game adjustments? Quantit Econ 9(2):1053–1085

Hingson R, Heeren T, Morelock S (1989) Effects of Maine's 1982 .02 law to reduce teenage driving after drinking. Alcohol Drugs Driving 5(1):25–36

Johnson N (2009) NCAA 'point shaving' as an artifact of the regression effect and the lack of tie games. J Sports Econ 10(1):59–67

Liang L, Huang J (2008) Go out or stay in? The effects of zero tolerance laws on alcohol use and drinking and driving patterns among college students. Health Econ 17:1261–1275

Paul RJ, Weinbach AP (2011) Investigating allegations of point shaving in NCAA basketball using actual sportsbook betting percentages. J Sports Econ 12(4):432–447

Smith R (1979) Compensating wage differentials and public policy: a review. Ind Labor Relat Rev 32(3):339–352

Van Ours J, Williams J (2015) Cannabis use and its effects on health, education and labor market success. J Econ Surv 29(5):993–1010

Wagenaar A, O'Malley P, LaFond C (2001) Lowered legal blood alcohol limits for young drivers: effects on drinking, driving, and driving-after-drinking behaviors in 30 states. Am J Public Health 91(5):801–804

Zitzewitz E (2012) Forensic economics. J Econ Literature 50(3):731–769

Chapter 11
The Narrative in the Numbers

Abstract This chapter describes how to finish bringing closure to an empirical study, in a way that both structures and enhances the narrative used to report that study's methods, results, and conclusions. It emphasizes the importance of relating the study's findings back to the essential facts of the phenomenon of interest, and argues that the study's ultimate objective should be to understand that phenomenon on its own terms, not the terms prescribed by the researcher. In the process of doing this, the researcher will often unearth an "organizing principle" that grounds the behavior of the major actors involved in this phenomenon, and which also can ground the study's narrative. These ideas spring to life in applications to the housing crash, the behavior of carnival workers, development in the tropics, teenage fatherhood, and more.

And now it is time to trumpet your findings to the world. Your vehicle: the narrative that comprises your paper. But be aware that the research process does not halt here. This narrative should not merely report your methods and findings. Instead, it's your last opportunity to shape your conclusions and burnish their credibility.

I do not intend for this to be cosmetic: credibility and window-dressing are entirely different things. Rather, your narrative should reflect and reinforce good craftsmanship. In the process, I think it will become more interesting, not less.

11.1 Closing the Loop

Chapter 7 articulated a principle that has extended throughout this book: continuity. Rather than lurch from literature review to theory to regression specification to empirical result, build bridges between each. Doing so prevents non-sequiturs or "loose connections" from tainting your work and adds a layer of redundancy that strengthens your conclusions. Your writing should also reflect this principle.

The most basic way to do this is through another familiar principle, transparency: presenting not just these connections, but also enough contextual information that

D. Grant, *Methods of Economic Research*, Springer Texts in Business and Economics, https://doi.org/10.1007/978-3-030-01734-7_11

readers can execute cross-checks of their own. But there is also one more connection to be made.

Chapter 1 characterized the research process as a loop, which began with the essential facts of the phenomenon you were analyzing. It should end there too. In earlier chapters we learned to nestle analyses in context. You need to put your results and implications in context, too, and close the loop. Just as a closed loop is stronger than an open loop, an analysis is stronger when its conclusions are closely connected to the real world.

To appreciate the difference this can make, let's compare this Tirolean (1988, p. 26) chestnut, a classic investment/effort provision problem, with my own version. Here's Tirole:

> Assume a supplier's investment increases the quality of his product and thus its value, V, to the buyer, such that $V(I) = 3I - \frac{1}{2}I^2$, where I is the amount of investment in dollar terms. If the product can be produced at a constant marginal cost C, what is the efficient amount of investment? Is it achieved if the buyer has the right to buy the good at a specified price P?

This is a good problem. The math mimics the thought process involved: optimization for the first question, satisficing for the second. And here is my own version:

> At most carnivals and country fairs, you pay for the rides with tickets, and for the dart/water pistol/guess-my-weight games with cash. Why? Who gets paid more, the ride operators or the game hucksters? Why?

Now, you would probably say that this problem is unfair and that I'm horrible for assigning it to my students. And I would respond, look how much richer the principle is when you place it in its native habitat.

To get to the heart of the matter, you must first unpack the context: the hucksters must be keeping the cash, because no business would trust an employee to truthfully report her earnings with no mechanism for monitoring or recording sales. The hucksters thus must be "buying the job," paying rent to the carnival operator and retaining all revenue received. This generates maximum incentives for effort, which is the whole point. Since the hucksters work harder, they also earn more. You don't want the game operators doing the same thing, since they shouldn't sacrifice safety for sales. And now the system is complete.

This example illustrates three major benefits of closing the loop. First, it allows you to check your work. Both problems focus on the provision of effort or investment and the incentives supporting it. But in the second problem, one set of employees is spurred to provide sales effort and another set is not. Connecting this to context demonstrates that this is a reasonable thing to do in both cases, which helps ensure that our answer is correct. In fact, everything in this problem locks in so tightly that I used it for years before confirming that the answer above was factually correct.

Second, it tempers the implications of your analysis. The basic theory shows how appropriate incentives for effort or investment can be beneficial, but it cannot anticipate all possible reasons why providing such incentives could be a bad idea. By connecting the theory to context, my problem helps you appreciate its limitations,

one of which is that such incentives can detract from other important dimensions of effort.[1]

Third, it lets vernacular knowledge supplement formal evidence to create a cohesive, more credible whole. The fact that the hucksters are paid in cash would never be integrated into a formal model, but is valuable information nonetheless. In empirical studies, statistical tests determine how well your model fits reality quantitatively. Closing the loop helps you do the same thing qualitatively. The closer the fit, the more credible the conclusions.

Each of these three benefits deserves an example all its own.

11.1.1 Team Incentives in HMOs

Gaynor et al. (2004) examine the effect of group incentives on cost-cutting in panels of doctors that are contracted to a health maintenance organization (HMO). Each panel is rewarded, as a group, for expenditures that come in below some target. This gives rise to the classic free-riding problem, which is worse in larger panels, as physicians get a smaller share of the earnings derived from their own cost-cutting. Larger panels thus imply less cost-cutting.

Their estimates show this in spades: increasing the panel from ten doctors to twelve increases expenditures by 7.3% at the mean (p. 924). Closing the loop involves relating this estimate back to the individual physician, in dollar terms, to see if it is reasonable. That is, it involves determining, in the larger panel, how much less of his own cost-cutting a physician gets to keep.

This turns out to be surprisingly difficult to do, because the paper is opaque on the basic information needed to close the loop:

> Physicians were given substantial financial incentives for keeping annual expenditures below the global budget assigned by the HMO. Roughly 20% of a physician's fees were "withheld" if the physician's expenditures exceeded a target (determined from the composition of the physician's patients). Bonuses, on the other hand, were based on the performance of panels of physicians rather than individual doctors. These panels, which varied in size from three to thirty physicians, received incentive payments only if the panel's collective expenditures for the year came in under budget.

For the typical physician, how much money do these withholds and bonuses amount to, and how often are they received? Is the primary incentive to meet the target or pad the bonus? Are many physicians/panels close to the target/budget limit? How predictably can they be met? We never find out, which makes it harder to gauge the reasonableness of the estimate.

Fortunately, here, a scale analysis will suffice. Assume the panel initially has ten physicians, which each receive a bonus of $10,000 (sleuthing through the working paper, this seems to be in the right ballpark). One-tenth of that bonus, or $1000, is

[1] This point is acknowledged in the literature, of course, but my general point remains.

generated by a physician's own behavior; if the panel had twelve physicians and spending did not change, this number falls by about $200. This $200 is not the change in incentives, however, only the change in the *base* to which any changes in the physician's spending apply: the 7.3% reported above. The resulting amount, about $15, is barely enough for a nice dinner at Denny's.[2]

It is hard to imagine so much cost-cutting stemming from so little money. Trying to close the loop, and failing, shows that something is seriously wrong with the paper.

11.1.2 The Housing Crash

The most prescient warning of the housing crash, a mass-market paperback by Talbott (2006),[3] is one of those books that shouldn't be judged by its cover. Titled *SELL NOW!* in giant blue letters, it practically screams "don't take me seriously." But what it lacks in subtlety it makes up for in clarity. This book doesn't hedge its bets. From *Publisher's Weekly*:

> Talbott's latest effort in warning of the coming housing crash bluntly advises owners to liquidate in a hurry. Talbott argues housing prices will plummet by as much as 50% over the next five to seven years, a hit that will be felt on an international level and cascade into larger economic problems: more job losses and a weaker banking infrastructure.

For his efforts Talbott was rewarded with reviews like this one, on Amazon:

> Save your money–this book is sensationalism. For example, in California people have been predicting a bubble for the last generation. The reality is people prefer to live in California than in other places. The author is trying to scare you. I would not worry about it much.

That was in early 2006. By late 2008, the tenor of the reviews had changed.

As this book is written for a mass audience, its analysis is not rigorous by academic standards, though Talbott is familiar with the academic literature. But he compensates for this deficiency by connecting his analysis closely to context. There was no other way to adequately understand this phenomenon, because the ground underneath the housing market—it's a metaphor—had shifted.

Much of Talbott's analysis crudely mimics something more academic. First, he documents the facts: the unprecedented early-2000s run-up in U.S. real home prices, unmatched by increases in rents. Then he rules out a series of possible explanations of these facts: economic growth and zoning, interest rates and inflation, construction costs and demographics. Were the data adequate (which is unclear), much of this could be done more formally with time-series or panel regressions. Such analyses would be useful, but they would be trapped in the past, and they would not recognize the changed circumstances that made this episode different. They would not, they could not fully appreciate their own limitations.

---

[2] Oh, relax. Denny's has many menu items that are both nutritious and reasonably priced.

[3] Yes, more prescient than Robert Shiller's. Not that there's anything wrong with that.

These changed circumstances—the increasing popularity of adjustable rate mortgages, the effective diminution of underwriting standards, the failures of regulation—are essential to Talbott's argument that overly aggressive mortgage lending is key. To support this explanation, Talbott goes deep into the weeds of banks' lending practices and the behavior of regulators. He then carries forward the full implications of his argument, for both the national and international economy. Altogether, the analysis addresses the whole system, possesses a keen sense of scale, and reflects a wealth of vernacular knowledge. This made his conclusions more credible ex ante, and time would establish their credibility ex post.

11.1.3 Development in the Tropics

In the previous example vernacular knowledge was key to the argument, but this is rarely the case in quantitative analysis, where the empirics rule. Often, however, vernacular knowledge can be fused to these empirics, to the advantage of both.

In the late 1990s, several empirical studies showed that latitude, that is, distance from the equator, positively affects economic activity (Sala-i-Martin 1997; Sachs and Warner 1997; and several others). Kamarck (2002), a development economist well-acquainted with the tropics, is not amused:

> This is an important finding, but standing alone, what policy recommendation does it imply? One cannot advise a government that it should move its country into a temperate zone.[4] Only if we understand what it is about the tropical climate that creates the obstacles to development can suitable policy actions be taken.

You see, he's been there before, having described the development challenges facing tropical regions at length in a 1976 monograph. So, with the frustration of a man who has waited two decades for the world to see his way of thinking, he unleashes a fireball of unmeasurable, contextual information to fill in this gap.

> Geography and climate isolated Sub-Saharan Africa from the rest of the world and Africans from one another until very recently, and they still impose high transport costs. Where the desert does not come down to the sea, there is mostly swamp or lagoon. There are very few natural harbors. As most rivers fall off the escarpment near the coast, it is seldom possible to penetrate the interior by sailing upriver. Though Africa's coastline has been known for centuries, its interior was mapped only about 100 years ago.
>
> Land transport is little easier. Horse- and cattle-killing diseases rule out animal transport over most of tropical Africa, so commerce and travel had to depend on human porters, the most inefficient of all transport modes. This, and the slave trade, account for the high degree of ethnic fragmentation that has made nation building so difficult.
>
> Life across most of the tropics takes on an infinite multiplicity of forms. The number of species in a given area is a large multiple of that found in a temperate zone, and the conditions are ideal for rapid evolutionary adaptation. Thus high percentages of people harbor parasites, and there is a high probability that any plant or animal introduced into an area will attract some new pest.

[4] Except for the Maldives, which basically will do just that, because of global warming.

> Ideal conditions for agriculture, under which the right amount of water is available in the right place at the right time, rarely occur in the tropics. In many areas the soil is laterite, which is agriculturally poor. The sun burns away the organic matter, and torrential rains leach out the minerals needed for plant growth.

It's a compelling compendium of troubles, but don't let it be too convincing on its own. I could bring you to tears with the struggles of my grandmother, sent to an orphanage from the barren Newfoundland shore because her parents could no longer feed her, or her husband, my grandfather, who found conditions in nearby New Brunswick little better. This is exactly the type of anecdotalism we should strive to avoid.

Rather, Kamarck's litany gathers its power by its connection to the empirical studies that precede it. These bear the weight of showing that it really is tougher in the tropics. Kamarck's observations complement these findings by enhancing their credibility and explaining why they come to pass. Both are strengthened in the process.[5]

11.2 Organic Knowledge

These three benefits are reason enough to close the loop, but, as the wise man said—there is another. This benefit has less to do with the credibility of your conclusions than with the nature of the conclusions themselves, namely, that both you and the reader can connect more closely with the subject of your study and understand it more intimately. This was apparent in our discussion of the housing and development studies above, and not by accident. This idea is worth discussing in more detail.

The classic empirical micro study tends be somewhat detached from its subject. It deduces logical conclusions from general assertions about agents' objectives and constraints, and then tests them in a more or less direct manner via statistical analysis, with a limited store of vernacular knowledge and a limited understanding of the system and the magnitudes of forces involved.

This arm's-length quality has been described by various people in various ways. My preferred term is "inorganic," which is not pejorative, in its original use in chemistry or here. An inorganic analysis can be competent or incompetent; its conclusions could be correct or incorrect; its findings could be credible or not credible. It simply lacks the element that is needed for life. This element can be added by closing the loop, making your conclusions and implications more genuine, more natural, more closely connected to the economic environment—in a word, more organic.

[5] Of course, there is no reason a single study can't contain both types of evidence. Here that study is Ram (1997), who began by discussing Kamarck's (1976) findings, and followed with regressions that affirmed latitude's effect on growth, all else equal.

Organic knowledge is less impersonal. It understands how various agents are influenced by economic forces, how those forces are perceived, as large or small, welcome or unwelcome, and how they filter through relevant institutions. Thus, in Chap. 2, Ichniowski et al. point out that their "innovative" HRM practices are more easily implemented on those steel finishing lines that have better labor-management relations. Thus, in Chap. 6, Wilson's takedown of the quality-driven driver's license bureau manager is expressed from the perspective of that very manager, not from above. Thus, in Chap. 8, Goldin and Rouse's description of the blind audition procedures shows that these guys aren't kidding, and that these auditions should be truly gender-neutral.

Organic knowledge is also more multifaceted. It understands how the phenomenon of interest affects behavior throughout the system. Having defined the system and conducted your analysis accordingly, you can now think of cause and effect across the system, identifying consequences both anticipated and unanticipated. This was clearly the case in Talbott's analysis of the housing market, which, though informal, was extremely organic.

Such knowledge is not achieved by philosophizing, untrained "gut" intuition, or flowery language. That is the opposite of rigorous analysis, while such analysis and organic knowledge are partners instead. It is not achieved by hyper-competence. Studies with more robustness checks, that clean their data more soundly, etc., are preferred to those that don't—but this does not directly generate organic knowledge. And it need not be achieved via deep structural insights. Sometimes circumstances permit the derivation and testing of such theories, sometimes not (see Fig. 6.2). You can acquire organic knowledge either way.

Rather, the most direct way to make your knowledge more organic is connect your findings to the roots from whence they sprung: scale, systems, and vernacular knowledge. Sometimes this is best done formally, say with additional regressions that uncover causal mechanisms or investigate the implications of your results. Sometimes it is best done informally, describing how your results connect with the "facts on the ground," as with Kamarck.

Organic knowledge can be achieved as a matter of degree, depending on the nature of the topic and the data used to study it. For example, studies of labor market discrimination in North America come in three types. There are in-depth studies of particular situations, such as Goldin and Rouse's orchestra paper. There are "audit studies" that assign equivalent characteristics to people of different types, and compare their success in applying for jobs. Oreopoulos (2011), for example, found that applicants with traditionally Chinese, Greek, Indian, and Pakistani surnames receive less follow-up from Canadian employers than equivalent applicants with English-sounding surnames. There are also studies based on national survey data, such as the Current Population Survey, which infer discrimination's effect to be that part of the wage differential that can't be explained by observable characteristics such as schooling and work experience. This last set of studies is most general but least

organic, since the data is taken at arm's length. The first set is least general but most organic, since everything is observed at close range.[6]

11.2.1 Seeing the Problem on Its Terms

Let's carry this line of thinking to the limit and ask the question: What is the apotheosis of organic knowledge? The answer involves thinking beyond the immediate objective of our analysis, having what I call a "sense of something more."

This sense can be found across the spectrum of creative endeavor. I am no art expert, but I know that celebrated art conveys something more than is apparent on the surface, perhaps a subtle statement about humanity, or a glimpse into the mind of the subject of a painting. This is rarely achieved, at least to reasonable effect, without technical skill. But such skill alone does not guarantee this result.[7]

Mathematicians share this sense of something more. Proofs can be "beautiful," and there are many examples of an existing result being re-established in a way that is less cumbersome and more insightful. Legendary mathematician Paul Erdös believed God had a book of perfect proofs of theorems. The highest ideal to which a proof could aspire was to be "straight from the book."

We applied microeconomists are not artists, but craftsmen, and as such can have no pretense of seeing the sublime. There is little intrinsically beautiful about explaining a messy, frenetic world. Yet I think we still have something extra to aspire to, a sense of something more.

I once visited with a mechanical engineer, a friend who designs equipment in support of the space program. I asked him, What makes a good designer? His answer: you have to be able to visualize the technicians out in the field. A lot of designs look good on paper, he said, but then you actually get it made and deliver it to the technicians and go back to the office, and a year later you go out there for some other reason and the equipment is over there sitting in a corner with dust on it, and then you know they aren't using it. And if you ask them why, the technicians describe all these practical difficulties with the equipment: it's too labor intensive, too hard to operate, it doesn't work well with other equipment they use, it's not as simple as a jerry-rigged solution. Good designers don't have these problems. They understand how the equipment they design fits into the technicians' world. That is, they understand the problem on the technicians' terms.

Economists, in contrast, typically define and analyze problems on our own terms. We approach them from a particular perspective, and interpret our findings accordingly. These qualities give economics much of its power, but they do not

[6] Similarly, in industrial organization, there are more organic, less general single-industry studies and less organic, more general cross-industry studies.

[7] Which is why Salvador Dali doesn't impress Sister Wendy, despite his technical sophistication. His "desire to show off and to shock" whittles back this sense of something more (Beckett 1999, p. 113).

naturally tender a close acquaintance with our subjects, and do not necessarily fit the problem well. As a result, the policies we propose—our designs, if you will—can end up in the corner with dust on them as well. The solution is not to abandon our analytical methods, but to use them, when possible, to a further end: understanding the problem on its terms, not our own. To fully swing through the ball, we should aspire to this.

This is much more difficult for economists to do than it is for engineers. We can't often talk to the subjects we study, and even if we could, it would be hard for them to answer many of the questions we would ask. They may not recognize what is general in their circumstances and what is idiosyncratic, and may not understand many aspects of the system in which they play a part. And, while budding mechanical engineers can spend their teenage years building things in their basement or garage, none of us could go home after school and practice deterring entry, redesigning school finance, or implementing environmental performance standards, and seeing how it all works out.

Furthermore, the tools at our disposal—mathematical, structured, systematic–are not naturally suited to this mission, being oriented to a somewhat different end, drawing conclusions about causality. The best we can do is to use our analytical framework to look beyond the immediate question we are trying to answer, while also looking beyond this framework, breaking the bonds into which we have placed ourselves.

Few economists have bound themselves more tightly than Robert Lucas, whose scale analysis I praised in Chap. 6. His discussion of the business cycle saw the problem wholly on his terms. The costs of economic fluctuations are limited solely to variability in consumption: the utility difference between a consumption path that grows smoothly at trend and one that jumps unpredictably around that trend. Consequently he concludes that economic fluctuations impose minor costs on the average consumer, and, to the extent that individual households' experiences differ from the average, it is a problem for social insurance, not stabilization policy (Lucas 1987, pp. 29–31). But many critics have pointed out there are other ways to look at the problem. Most obviously, negative fluctuations may reflect output that is below its potential; there are direct utility losses from unemployment as well (Helliwell and Huang 2014). Understanding the business cycle on its terms involves trying to quantify all such costs, at least those whose magnitude is first order.

This macroeconomic debate has been surprisingly intractable, but greater resolution has occurred in the field of development economics, which has already been down this road. Easterly (2001) had a field day decrying the disastrous early efforts of development economists who followed the prescriptions of basic growth models in advising countries to borrow and invest, seeing the problem in the terms of the economic theories in favor at the time. Deaton (2013, pp. 272-273) puts it well:

> Fixing world poverty is sometimes seen as an engineering problem, like fixing the plumbing or repairing a broken car. Statistical analysis shows a robust correlation between economic growth and the share of national income that is invested, so it is straightforward to calculate how much additional capital a country "needs" in order to grow faster.

> That such calculations are wrong has been argued for a long time, though it does not remove their seductiveness to many even today. Peter Bauer, writing in 1971, made a crucial point: "If all conditions for development other than capital are present, capital will soon be generated locally, or will be available on commercial terms from abroad. If, however, the conditions for development are not present, then aid—which will be the only source of external capital—will be necessarily unproductive."

Development economists increasingly recognize that understanding these conditions is what it takes to see the problem on its terms.

11.2.2 Organizing Principles

Seeing the problem on its own terms is not merely something you do as you build your narrative. Rather, you use it to help *construct* that narrative. The two come together in what Franklin Fisher (1991) would call an "organizing principle."

Many organizations, firms, markets, or economies have a central problem that they need to solve or mission they need to achieve. They are often structured around a basic philosophy or approach to solving this problem or achieving this mission: an organizing principle. In capitalist economies, the central problem is achieving some measure of utilitarianism, and the organizing principle is laissez-faire. In Chap. 3's *Moneyball*, the central problem is staying competitive on a small budget for players, and the organizing principle is cost-effectiveness. In the free agent era, the Oakland Athletics just cannot afford to think like everyone else, so the general manager, Billy Beane, shifts the front office's focus from profit maximization or even winning to something more specific: value for money. The movie version of Beane puts it memorably:

> The problem we're trying to solve, is that there are rich teams and there are poor teams, then there's fifty feet of crap—and then there's us.

To economists, these might all seem like the same thing: given this budget, profits are maximized by winning as much as possible, which in turn requires maximizing the return on investment in players. But that is to see the problem on our terms. Billy Beane knew better. There is a practical, organizational difference between striving for profit maximization and striving for cost-effectiveness, even if (here) one's mathematical program is simply the dual of the other.

Moneyball is not a one-off. Many multi-division firms have units that are designated as cost centers, whose performance is assessed on cost-effectiveness, and other units that are designated as revenue centers. Re-christening them all as profit centers would be simpler, and would prevent some mild deviations from optimality as well. But these designations define the problem differently, closer to the heart of the matter. Their longstanding persistence in competitive industries suggests that they meet the market test. Treating all units of the firm as profit maximizers would not exactly be wrong, or even obfuscating, but it would be to see the problem on our terms, and it would overlook the organizing principle of these business units.

Seeing the problem on its own terms often involves recognizing this central issue and the organizing principle associated with it. This principle can then organize your narrative as well. Remember Mincer's introduction to *Schooling, Experience, and Earnings* from Chap. 3?

> Investments in people are time consuming.

It was the organizing principle of the book and, at the same time, the central problem confronting new high school graduates: What do I do with all this time?

11.3 Teenage Fatherhood and the Pursuit of Happiness

These concepts come together nicely in an innocent study by Fletcher and Wolfe (2012), who examine how teenage fatherhood affects educational, marital, and labor market outcomes five or so years later. A key issue in such an analysis is the counterfactual. Teenage fathers probably differ from other teenagers in ways that are not easily measured, so a standard regression using standard controls is problematic. The authors handle this nicely by using data from the National Longitudinal Study of Adolescent Health, which contains detailed information about individuals' sexual activity and pregnancy history. Thus they can compare births with uncompleted pregnancies (miscarriages and abortions), controlling for unobservables much more satisfactorily.

Table 11.1, adapted from their paper, presents a subset of their findings: five estimates for each of six outcomes, with their preferred estimates in the last two columns. From this, Fletcher and Wolfe conclude:

> We find the effect of teenage fatherhood to be a reduction of fifteen percentage points for receipt of a high school diploma in our preferred specification, while teenage fathers are between seven and seventeen percentage points more likely to attain a General Equivalency Diploma (GED). The increase in the likelihood of marriage is between fourteen and twenty-five percentage points, with a preferred estimate of twenty percentage points. We also find some evidence that teenage fatherhood increases full-time employment probabilities, but there are no statistically significant effects on overall employment status or total family income.

These conclusions are slightly stilted, inorganic, appropriately cautious—in other words, absolutely typical. So what is there to object to? Only this: if these fathers are dropping out of school in significant numbers, it must be for a reason. Searching for the reason invites us to see the problem not on our terms, but on the terms of the young men whose world has just been transformed.

To start, let's fix that calamity of a table, consolidating the sample size information, improving the labels, and killing a star and a decimal point (Table 11.2). That's better. See it now? Weaving together the effects on schooling, employment, and income, the preferred estimates suggest both the reason for dropping out and its medium-term consequences. Here, so far as I can tell, is the narrative hidden in the numbers:

Table 11.1 The effects of teenage fatherhood on young adult outcomes (from Fletcher and Wolfe 2012)

	Full sample	Sexually active only	Add community fixed effects	Pregnancies only subsample	Birth/ miscarriage subsample
Diploma	−0.199***	−0.172***	−0.162***	−0.162**	−0.152
	(0.033)	(0.035)	(0.035)	(0.074)	(0.096)
Observations	7703	5567	5567	335	258
GED	0.063**	0.049*	0.044	0.072	0.114**
	(0.027)	(0.028)	(0.029)	(0.054)	(0.051)
Observations	7701	5567	5567	335	258
Employment	−0.037	−0.048	−0.051	0.052	−0.002
	(0.035)	(0.037)	(0.036)	(0.065)	(0.073)
Observations	7277	5306	5306	320	246
Full time	0.015	−0.010	−0.015	0.161**	0.060
employment	(0.045)	(0.045)	(0.042)	(0.080)	(0.090)
Observations	7277	5306	5306	320	246
Labor income	0.940	0.120	−0.106	1.446	2.220
	(1.108)	(1.121)	(1.175)	(1.926)	(1.889)
Observations	7346	5309	5309	322	249
Total income	1.470	0.664	0.563	1.096	2.153
	(1.101)	(1.172)	(1.251)	(2.161)	(2.224)
Observations	7194	5214	5214	314	243

Controls for age and its square included but not shown. Republished with permission of John Wiley and Sons, from Fletcher, J. M., & Wolfe, B. L. (2012). The effects of teenage fatherhood on young adult outcomes. Economic Inquiry, 50(1), 182–201. Permission conveyed through Copyright Clearance Center, Inc. *p < 0.10, **p < 0.05, ***p < 0.01

> Based on these estimates, about fifteen percent of teenage fathers drop out of school in order to work full time. Perhaps half of these eventually earn their GED, a credential that Heckman and others have shown to have little value in the labor market. As both total income and labor income are little higher four or five years later, this strategy pays medium-term financial dividends that are modest at best and nil at worst.[8]

Now can you start to feel the sweat on our new father's face, the faint feeling in the pit of his stomach. For some of these young men, the organizing principle of their late adolescence was to be the pursuit of happiness—and not in the sense that America's founding fathers intended. But that is all about to change.

[8] This scenario is not ironclad: the "preferred" estimates are not precise, and we cannot be sure that the people who dropped out of school began full time work soon after. Earlier waves of the Add-Health survey allow some further exploration of these points, which can be done within the existing analytical framework, using this framework to look beyond the immediate question we are trying to answer. The most obvious line of inquiry is that the pregnancy-only estimates most differ from those in the previous columns for employment. Controlling for employment history might capture this essential element of individual heterogeneity, allowing use of the full data set to yield more precise, more detailed estimates of the employment ramifications of teenage fatherhood.

Table 11.2 The effects of teenage fatherhood on young adult outcomes (OLS coefficients on a teenage fatherhood dummy variable, with standard errors in parentheses)

	Dependent variable					
Sample/ specification	*H.S. diploma*	*GED*	*Employed*	*Employed full time*	*Annual labor income ($1000s)*	*Total annual income ($1000s)*
Full sample (N ≈ 7400)	−0.20** (0.03)	0.06** (0.03)	−0.04 (0.04)	0.02 (0.05)	0.94 (1.11)	1.47 (1.10)
Sexually active only sample (N ≈ 5400)	−0.17** (0.04)	0.05* (0.03)	−0.05 (0.04)	−0.01 (0.05)	0.12 (1.12)	0.66 (1.17)
Above sample with community fixed effects (N ≈ 5400)	−0.16** (0.04)	0.04 (0.03)	−0.05 (0.04)	−0.02 (0.04)	0.11 (1.18)	0.56 (1.25)
Pregnancies only sample (N ≈ 325)	−0.16** (0.07)	0.07 (0.05)	0.05 (0.07)	0.16** (0.08)	1.45 (1.93)	1.10 (2.16)
Birth/miscarriage sample (N ≈ 250)	−0.15 (0.10)	0.11** (0.05)	−0.00 (0.07)	0.06 (0.09)	2.22 (1.89)	2.15 (2.22)

Controls for age and its square included but not shown. Within each row, the exact number of observations in each regression differs from the listed approximate value of N by less than 5%. Adapted from Table 11.1, used by permission of John Wiley and Sons, from Fletcher, J. M., & Wolfe, B. L. (2012). The effects of teenage fatherhood on young adult outcomes. Economic Inquiry, 50(1), 182–201. Permission conveyed through Copyright Clearance Center, Inc.
*Statistically significant at 10% level; **Statistically significant at 5% level

11.4 Conclusion

We began this chapter carrying forward a theme developed throughout this book: continuity. We end it the same way, carrying forward another such theme, the primacy of ends over means. Any study is just a step toward a larger end: truly understanding the phenomenon of interest, diagnosing its problems, prescribing correctives. So don't trap yourself within the framework you have created, if circumstances permit you to learn more.

This is not costless. In seeking organic knowledge and trying to see the problem on its terms, you run some risk of being subjective, unscientific, and wrong. We cannot escape this. It is in the nature of a craft. Just remember, we can also be wrong for another reason: because we have imposed a framework that doesn't fit and followed its implications to the end, however ridiculous they are. In the current state of our profession, this type of error easily outweighs the other.

Policymakers have already cast their lot. They have to try to understand the problem on its terms. It is inherent in political decision making, in which efficacy, political expediency, economic and non-economic factors, the role of private vs. state action, and the perspectives of multiple stakeholders all come into play. For us as researchers to do otherwise doesn't just inhibit our connection with policymakers—it limits our credibility as well.

Four hundred, a thousand years ago Polynesian voyagers traipsed halfway across the ocean, with nary a GPS or sextant in hand. For guidance, they turned to the shape and motion of the clouds, and the rhythm of the swells. To them, the sea and sky spoke volumes. All they had to do was listen.

My friends, it is not so different for us. The data—your data—is screaming at you. Screaming. At you! IN ALL CAPS. There is a narrative in the numbers. Listen closely. Can you hear it?

Food for Thought

1. What is the central problem of the economics profession? What organizing principle do I advocate for solving it? (For a hint, see Chap. 1.)
2. Woodford's classic *Interest and Prices* (2003) considers optimal stabilization policy in a standard New Keynesian framework. His most basic model (pp. 400–401) generates a loss function that is a weighted quadratic in inflation and the output gap. While the literature often assumes equal weights on the two terms, he argues that in fact the weight on the former should be twenty times that of the latter. This conclusion can be made more organic by comparing it to estimates on the relative importance of these two terms from the literature on "the macroeconomics of happiness," or a related literature on the macroeconomic determinants of summary measures of economic content such as the "Consumer Confidence Index." How do these estimates compare with Woodford's conclusion?
3. "Knot Yet," a publication of the National Marriage Project, examines trends in employment, marriage, and childbearing over the last several decades in the U.S. These trends vary by education, leading it to develop not one organizing principle, but two, for the role that marriage plays in peoples lives: a "cornerstone" for some, and a "capstone" for others. After reviewing this publication, easily found online, briefly articulate what these two organizing principles mean, and describe the statistics supporting these conclusions. Which organizing principle do most Ph.D. economists use: marriage as capstone, or cornerstone?
4. Lucas' *Models of Business Cycles* (1987) and Bewley's *Why Wages Don't Fall in a Recession* (1999) approach the problem of unemployment in two different ways. Lucas (p. 56, 57) celebrates McCall's (1970) micro-founded search theory, in which the long run unemployment rate is determined by the job separation rate, workers' reservation wage, and the distribution of job offers.

> Asking questions of McCall's theory invites us to think about unemployment in ways that fix-price and other macro-level theories can never lead us to do. Questioning a McCall worker is like having a conversation with an out-of-work friend: 'Maybe you are setting your sights too high,' or 'Why did you quit your old job before you had a new one lined up?' This is real social science: an attempt to model, to understand human behavior by visualizing the situations people find themselves in, the options they face and the pros and cons as they themselves see them.

Bewley, instead, interviews hundreds of employers during a recession, documenting how rarely employers cut nominal wages and the reasons why they don't. Here, for example, are the owners of a small nonunion machine shop and a medium-sized car dealership:

> I never cut anyone's wage. That would be too personal. It would be putting too much power into your own hands and would be resented.
>
> If I cut pay, people would leave out of rage, even though they have no place to go. They would feel they had to. They live close to the edge anyway, spending 110 percent of their income, no matter what it is. The body shop people would certainly leave. They are crazy. They smell too many fumes.

Compare these two authors in terms of their adherence to the ideals of organic knowledge and seeing the problem on its own terms. Are they closer to each other with respect to the first ideal or the second?

References

Beckett W (1999) Sister Wendy's 1000 masterpieces. DK Publishing, New York.
Bewley T (1999) Why wages don't fall in a recession. Harvard University Press, Cambridge
Deaton A (2013) The great escape: health, wealth, and the origins of inequality. Princeton University Press, Princeton, NJ
Easterly W (2001) The elusive quest for growth: economists' adventures and misadventures in the tropics. MIT Press, Cambridge, MA
Fisher F (1991) Organizing industrial organization: reflections on the Handbook of Industrial Organization. BrookPap Econ Act 1991:201–240
Fletcher JM, Wolfe BL (2012) The effects of teenage fatherhood on young adult outcomes. Econ Inq 50(1):182–201
Gaynor M, Rebitzer J, Taylor L (2004) Physician incentives in health maintenance organizations. J Polit Econ 112(4):915–931
Helliwell JF, Huang H (2014) New measures of the costs of unemployment: evidence from the subjective well-being of 3.3 million Americans. Econ Inq 52(4):1485–1502
Kamarck AM (1976) The tropics and economic development: a provocative inquiry into the poverty of nations. The World Bank, Washington, DC
Kamarck AM (2002) Economics as a social science: an approach to nonautistic theory. University of Michigan Press, Ann Arbor, MI
Lucas RE (1987) Models of business cycles. Basil Blackwell, Oxford
Oreopoulos P (2011) Why do skilled immigrants struggle in the labor market? A field experiment with 13,000 resumes. Am Econ J Econ Pol 3(4):148–171
Ram R (1997) Tropics and economic development: an empirical investigation. World Dev 25(9):1443–1452
Sachs JD, Warner AM (1997) Sources of slow growth in African economies. J Afr Econ 6(3):335–376
Sala-i-Martin X (1997) I just ran two million regressions. Am Econ Rev 87(2):178–183
Talbott J (2006) Sell now! The end of the housing bubble. St. Martin's Griffin, New York
Tirole J (1988) The theory of industrial organization. MIT Press, Cambridge, MA
Woodford M (2003) Interest and prices: foundations of a theory of monetary policy. Princeton University Press, Princeton, NJ

Conclusion

So ends our two books.

You have probably noticed that my theses—that economics' central problem is credibility, and that this devolves from inattention to craftsmanship—are, in the terms of that old Scottish verdict, not proven. You've got me there. My examples are varied, in almost every way, but you can't prove something on examples alone, and examples is all I have.

But these theses aren't merely retrospective, and need not be tested only by me. Test them yourself, prospectively, as you attend conferences, watch seminars, read papers, and write your own. If they hold true, you will confront the issues raised in this book at every turn, and will see how the principles I've laid out can ameliorate them.

If you begin by applying these principles back to this book itself, you will realize that I have one imperative remaining: to close the loop, by articulating this book's methodological organizing principle and comparing it with the alternatives.

No depiction of these alternatives can be fully just. None of them have been formally stated or justified, except perhaps for specific research questions; none has an official spokesperson or body of arbiters. Still, who are we kidding? We all can recognize three distinct strains of thought, three visions of the research process so well known that they mold the profession's self-image: what it is that economists are supposed to do when they do economics. These visions (and their primary exponents) have been represented in the preceding chapters. Let's take them out and hold them up to the light.

The first vision emphasizes one pillar of the profession: theory. Any model not built upon economic bedrock is problematic for policy. Only structural models founded on behavioral and technological primitives can provide the true economic understanding that is needed to fix economic problems. While this organizing prin-

D. Grant, *Methods of Economic Research*, Springer Texts in Business and Economics, https://doi.org/10.1007/978-3-030-01734-7

ciple has been most forcefully expressed in macroeconomics, its spirit survives in micro as well, in some of the more elaborate structural models that arise in a variety of fields.

The danger in this approach is that it downplays the tradeoffs that are intrinsic to modeling. The fact that causality is a desirable quality of models doesn't make the reach/grasp problem disappear, nor any infirmities of your data, etc. If we bury the process at work under an avalanche of formalism, we risk developing models that are impractical to estimate and lacking in versatility.

When this happens, the easiest way to cover your tracks is by sandbagging testing. What null should the model tested against? How should it be compared to competing models? Which elements of the system should be included in these tests? How "good" must the evidence be to pass the threshold of policy usefulness? Simply testing your main parameter estimates against the null hypothesis of zero, or variations on that theme, ignores all of this. Then the conclusions we reach may be strong, but not credible in the end.

By emphasizing one input of the research process to the detriment of others, this vision does not sufficiently distinguish a theory from a model. In doing so, it diverts us from the ultimate obligation of modeling: to ensure, with everything we have—description, vernacular knowledge, hypothesis testing, closing the loop, etc.—that we provide a description of process and outcome that can be taken seriously.

A second vision emphasizes the other pillar: estimation. Garden-variety econometric estimates are suspect, because there are so many unanticipated or unfixable sources of bias. Only estimates based on genuine natural experiments, involving randomization, discontinuities, matching, etc., give us the "clean identification" we need to ensure our estimates are sound.

This approach isn't foolproof, as we saw in Chap. 8's discussion of Goldin and Rouse (2000) and Angrist and Krueger (1991). To follow the whole tick-tock on the latter paper—the original, "small-sample" criticism by Bound et al., Angrist's response in *Mostly Harmless*, the counterevidence in the working paper version of Buckles and Hungerman (2013), Angrist's response in *Mastering Metrics*, the more definitive counterevidence in Buckles and Hungerman's final paper—is to despair whether economists will ever agree on anything. Finally I dug up the two "pre-existing" papers, downplayed by the original and by Buckles and Hungerman, which addressed the central issue of the independence of birth quarters (Kestenbaum 1987; Lam and Miron 1991, available earlier as a working paper). Mmmmm. Angrist and Krueger never should have started down this road to begin with.

This vision, like the first, shunts some inputs into the research process off to the side. In doing so, it gives us a clean causal estimate, under suitable circumstances, but often obscures our broader understanding what is really going on.[1] Even then, craftsmanship is not rendered irrelevant. It is needed to determine whether the circumstances are suitable to begin with, as we just saw. It is needed even more when

[1] Commentary by Leamer, Keane, Sims, and Nevo and Whinston in the Spring 2010 *Journal of Economic Perspectives* nicely fleshes out this point and its implications.

no clean natural experiment is available, a situation for which this vision offers much less guidance. Most important economic questions fall into this category.

The third vision, put forward by Milton Friedman (1953), was spawned by consternation over the fact that managers rarely sounded like they followed MR = MC when asked (Hall and Hitch 1939). It emphasizes neither pillar per se, nor any other input into the research process. A good model is one that consistently makes good out-of-sample predictions, period.

The vision would seem to accord with Chap. 10's emphasis on ends, not means. After all, if a model has fantastic out-of-sample predictive power, who really cares how it was constructed? The problem with this sentiment is that it is a theoretical proposition rather than an empirical one. Had Friedman followed his essay with a companion piece titled "I Told You So," filled with literature-based examples of nonsensical, uncontrived, highly-predictive economic models, that would be one thing. There is no such paper, and there never will be.

Did Friedman foresee that, by limiting the evaluation criterion to one largely unreachable, untestable goal, his essay would become the intellectual foundation for every half-baked analysis that failed to seriously consider the soundness of its model, data, and results? Did he anticipate that, for many policy decisions, $Y = ßX + ε$ would be a major improvement over the alternative, despite its faults? Or were these out-of-sample predictions he got wrong?

Shoot. If the deal is that people can't always articulate why they do things, you needn't base your counterargument on the philosophy of science. All you need is this (Scott 1998, p. 329):

> Some knowledge is so implicit and automatic that its bearer is at a loss to explain it. A staple of early medical training, I have been told, is the story of a turn-of-the-century physician who was fantastic at diagnosing syphilis in its early stages. Lab tests confirmed his diagnoses, but he himself did not know just what it was that he detected in his examinations that led him to his conclusions. Intrigued by his success, hospital administrators had two other doctors closely observe his examinations over several weeks. At long last, they eventually realized that he was unconsciously registering a slight eye tremor in these patients. These tremors then became a universally recognized symptom of syphilis.

As an aspiration, each of these visions is admirable. We all want fully causal, impeccably-identified economic models with outstanding predictive power. In practice, however, each vision emphasizes one aspect of the research process to the diminution of others, thereby engaging in a touch of what Thomas Mayer (1993, p. 57) derisively calls "The Principle of the Strongest Link." The stress placed on causality, identification, or model versatility distorts the tradeoffs that are involved in modeling, diverts our attention away from other elements of the research process, which is generally quite multifaceted, and offers limited guidance in the many situations in which these visions are untenable.

This would pose little problem if research elements effectively substituted for each other, so that excellence in one compensated for shortfalls elsewhere. This is not the testimony of experience. Good theory will not compensate for lousy data, weak testing, or a poor understanding of the situation being modeled, and neither will anything else. In general, the opposite is true: these research elements are

complementary, as I have claimed repeatedly throughout this book. They form a system.

This fact implies a different vision of the research process: as a confederation of fallible elements, an assembly of many potentially faulty components, each of which depends on the others to work. The study of such assemblies falls under what systems engineers would call "reliability theory." This theory undergirds Gavrilov and Gavrilova's mortality model in Chap. 3, which treats the human body as a corporation of components, each of which is made from imperfect parts, and each of which must function for the organism to succeed.

Gavrilov and Gavrilova point out that system reliability can be achieved in two ways. One way is to construct each component from quality parts. This is common in the mechanical realm, say in producing stereo systems, because each part can be sturdily made before its assembly into the component, and each component carefully tested before its assembly into the final product. The other way is to have redundancy in the parts that each component is composed of. Then the failure of one part is inconsequential so long as the others continue to function. This is common in the biological realm, where sturdiness cannot be assured and testing is impossible.

The first way of achieving reliability is akin to the pure application of the scientific method. Humans have the power to control each element of this research process: theory, experimental design and execution, data collection, and hypothesis testing. Accordingly, it is reasonable (though somewhat Pollyannish) to visualize "pure science" as a clean, clear, trustworthy sequence of procedures.

Social scientists rarely have this power. We cannot place too much faith in any one element of our research. Most of the time, we can't control the variables that we measure and how well we measure them, can't design our own experiments and eliminate confounders or alternative explanations from consideration, can't fully understand how our findings might depend on context or institutional detail. Then to think of the research process in the first way is to labor under a mistake. Our approach to achieving reliab—I mean, credibility—must be more like the second way, which is suited to a world of weakness, not strength.

In a system of complementary elements, we must attend more to those that are weaker, not stronger—and with limited control over the quality of each element, everything is suspect. The implications of this research vision have resounded throughout this book. Respect this weakness by employing every research element at your disposal, not just the two pillars of the profession, and by honoring the principles of self-determination and seeing the problem on its terms. Compensate for possible flaws in any one facet of your research by embracing the principle of redundancy and its cousin, transparency. Acknowledge the complementarities between research elements via the principle of continuity, which de-compartmentalizes the research process and strengthens the connections between these elements.

This weakness extends to the model's evaluation and application, which occur in a world with many possible influences on behavior, both theoretical and contextual. This fact makes model versatility somewhat elusive and traditional hypothesis

testing somewhat effete. Accordingly, embrace the reach/grasp problem and the tradeoffs therein, and base your model on a thorough, well-rounded understanding of the phenomenon of interest, achieved using vernacular knowledge and description. Broaden hypothesis testing in the ways we have discussed, and examine your model's collateral implications by closing the loop. Adhere to the principle of harmony, since amassed evidence that a model captures the essence of the process at work is an important sign of its fidelity and versatility. All of this is the opposite of genius—and the essence of craftsmanship.

I get it. For all of this you are guaranteed nothing in the end. Things don't have to work out. Throw everything you have at Nigeria's Child Rights Act, and you will still come away knowing little about its effect on child marriage. The system is too large and sluggish, the experimental content too limited for econometrics of any sort to be decisive. But there are plenty of cases in which craftsmanship carries you a long way. You may even achieve that ideal of coherence, in which everything falls into place, like the carnival problem in the previous chapter.

To anyone who believes otherwise, who argues that these weaknesses are so daunting that the whole exercise can't really be taken seriously, I say: poppycock. You wouldn't be saying this if you've ever achieved anything like coherence, and if you've never done that, then how can you know? Your lack of experience needn't extend to everyone. This view of social science is more social than science.

Timidity like this, an atavistic fear of being so fully exposed, is what makes this vision of applied microeconomic research seem a little radical. It is tempting to remain where we are, sheltered by excessive formalism, uncertain experimental content, and ambiguous language. For this reason, implementing some of the principles articulated in this book requires a certain independence, a willingness to stand aside from the rushing current.

This independence is, I recognize, a genuine barrier to these ideas' adoption, particularly for readers only newly acquainted with the profession. The practical benefits of lying securely within the mainstream are many: it flatters the work of potential referees, reduces or eliminates methodological confusion, and avoids raising the hackles of tepid reviewers who, unsure of their ability to assess things independently, view anything non-routine, however sensible, as suspect. These concerns are not irrational. So, the thinking goes, go along to get along. You're just doing your job—you and everyone else.

In response, all I can tell you is this: be brave. Luminous beings are we! The conventions of the world are not as mighty as we think, and the courage required to confront them not theatrical, like Mel Gibson in *Braveheart*, but something quieter and more true. So let out your Tookish side. I think you will gain something in the end.

I wrote this book for a reason we all share—a latent discontent that develops early in our training, a queasy feeling that something is just not right. I know that feeling. It gnawed at me for 25 years. I feel better now.

References

Angrist J, Krueger A (1991) Does compulsory school attendance affect schooling and earnings? Q J Econ 106(4):979–1014

Buckles KS, Hungerman DM (2013) Season of birth and later outcomes: old questions, new answers. Rev Econ Stat 95(3):711–724

Friedman M (1953) The methodology of positive economics. In: Friedman M (ed) Essays in positive economics. University of Chicago Press, Chicago

Goldin C, Rouse C (2000) Orchestrating impartiality: the impact of blind auditions on female musicians. Am Econ Rev 90(4):715–741

Hall R, Hitch C (1939) Price theory and business behavior. Oxf Econ Pap 2:12–45

Kestenbaum B (1987) Seasonality of birth: two findings from the decennial census. Soc Biol 34(3–4):244–248

Lam D, Miron J (1991) Seasonality of births in human populations. Soc Biol 38(1–2):51–78

Mayer T (1993) Truth vs. precision in economics. Edward Elgar, Brookfield, VT

Scott JC (1998) Seeing like a state: how certain schemes to improve the human condition have failed. Yale University Press, New Haven

Glossary

Abjure Renounce; give up entirely.
Accoutrements Embellishments that serve a purpose, rather than being merely frivolous.
Accretion A slow accumulation, built up layer upon layer like a pearl.
Ad hoc Created for an immediate purpose only, with no larger goal in mind.
Aficionado Someone who is passionate and knowledgeable about some subject or activity.
Anoxic Without oxygen.
Apotheosis The ideal achievement or level of achievement.
Armamentarium All the stuff that is available for you to use in performing a duty or accomplishing a task.
Atavistic Raw and primitive.
Autoimmune disease A disease in which someone's immune system turns on that person and makes them sick.
Ballpark As a noun, it means "to be in the broad neighborhood of"; as a verb, it means to make an approximation that is intended to be in that neighborhood.
Beef Problem or objection.
"Between a rock and a hard place" A phrase implying a situation in which there are no good options, only various unsatisfactory options to choose from.
Body shop A place that fixes non-mechanical parts of a car, such as the frame or the exterior.
Borax A mineral containing boron that is better used in laundry detergent than in food.
Bucolic Pastoral; very rural.
Cadastral Having to do with land surveys that establish property boundaries.
Catholic Universal or wide-ranging.
CFA, or CFA franc Roughly speaking, a currency used by several West African countries.
Chestnut A real classic from the past.
"Come up sevens" A rare event in slots (gambling) in which you win big.

D. Grant, *Methods of Economic Research*, Springer Texts in Business and Economics, https://doi.org/10.1007/978-3-030-01734-7

Communitarian A social structure that revolves around each individual contributing to the community.
Conflux Coming together.
Consonance Coming together in concurrence or harmony.
Consumption germs Tuberculosis (which, ironically, also follows a random walk).
Conundrum A mystery wrapped in an enigma stuffed inside a puzzle; the turducken of problematic situations.
Coriolis Effect When something is moving within a rotating medium, such as the Earth, this force pushes the object perpendicular to its direction of motion.
Countermand To dictate a reversal of a previous order or act.
Cutting room floor Refers to the editing of movies that are shot on film; material that was excluded from the final version of the movie ended up on the cutting room floor.
De rigueur Something the elite, especially, are socially expected to do.
Deep into the weeds Really getting into details.
Dicta Plural of dictum, which is like a very strong dictate.
Dot-matrix An old type of printer, common in the 1970s and 1980s, that printed everything using little dots that were aligned on a grid.
Dyad A connection between two points on a graph; more generally any pair of like things.
Effectuates Makes happen.
Effete Worn out; exhausted in substance or content.
Electorate Potential (but not necessarily actual) voters.
Escarpment Like a long cliff.
Exigencies Things that must be urgently done.
Exoskeleton A skeleton that is on the outside, like lobsters or scorpions have.
Exponents People who advocate for an idea.
"A feature, not a bug" What we want to happen, not what we don't.
Fixtures A term used outside the U.S. for the schedule of games for a sports team or league.
Flyover country A pretentious phrase used by some coastal Americans to describe people living in the interior of the country.
"For crying out loud" An expression of great frustration.
Fortran An early computer programming language that was used especially for mathematical tasks.
Fractal An often-complicated pattern that repeats itself on increasing spatial scales.
Functional fixedness Getting so absorbed in the details that you lose the big picture.
G-7 countries A group of seven major industrialized countries, including the U.S., U.K., France, Germany, and Japan.
Garden-variety Ordinary; typical; unexceptional.
GED A substitute for a regular high school diploma that you get by passing a series of tests on English, math, etc. While I was in college, both of my parents studied for and obtained their GED's.

Glycerin A compound that is or has been used in antifreeze, electronic cigarettes, and food.

"Grist for the mill" Something substantive to work on.

Guanxi Who you know, or a term for the social capital possessed by knowing someone of value to you, such as a doctor.

Gumbo A soup, traditional in South Louisiana, that included whatever meat and seafood the cook had left over, including rabbit, crawfish, shrimp, sausage, and chicken.

Haka A cool, ritual chant and dance that originated among the Maori in New Zealand and is becoming more popular around the globe.

HAL The main supercomputer in the movie *2001: A Space Odyssey*, by Quentin Tarantino.

"Hang your hat" Place your confidence in; rest your case on.

"Hard-up" Poor, in poor shape, or both.

Hoary Really, really old, almost too old.

Homophily Like likes like.

Huckster An aggressive pitchman for a product or service.

Hydra A multi-headed monster from Greek mythology, which grows two new heads every time one is lopped off.

"In spades" In large, generous quantities or amounts.

Inured To get so used to something that you hardly even notice it anymore.

Ironclad Absolutely certain.

Keiretsu A network of businesses that are linked together, financially and otherwise.

Kinesthetic Active; dynamic; full of movement.

Laterite A mineral-laden, often-clayish soil that is hard to grow things in.

"Levers to press" In the archetypal (yet cartoonish) Taylorist factory, you have a set of levers and pressing the right one gives you the result you want.

"Low-hanging fruit" The fruit that is easiest to pick; opportunities that are easy to seize.

Likert Scale A multi-option, graduated survey response scale, e.g., "strongly agree," "somewhat agree," … "strongly disagree."

Manifest variable In psychology, a variable that is actually measured (as opposed to a latent variable, which isn't).

Marlin-spike A tool that sort of looks like a big needle, which is used to work with rope on ships even today.

Missive A letter, often long, and often official.

Nantucket A long, thin island off the Massachusetts coast that used to be a haven for whaling vessels, and now has vacation homes for rich people.

Neuroticism General nervousness and anxiety.

Non-sequitur Two statements, the second of which does not follow from the first.

"Not even wrong" Refers to Wolfgang Pauli's derogation of theories that make no testable predictions, and thus which cannot be proven wrong.

Obfuscating Covering things up or deliberately confusing things.

"Oh snap" A somewhat outdated phrase that indicates that one person successfully made fun of another.

"Old-school" How it used to be done in the somewhat-distant past.
"On point" Right on the mark.
"One-off" An isolated case or unique event.
Orogeny Mountain-making.
Pandora's Box In Greek mythology, a box that, when opened, unleashed all kinds of mayhem.
Pantheon A literal or figurative collection of highly honored people (or their remains).
Pejorative Derogative; putting people down.
Pernicious Particularly bad or evil.
Piggy banks Hollow ceramic animals with slits in the top, in which children put their allowance or spare change in order to save up for college, like I did.
Pollyannish Overly and unrealistically optimistic.
Pong The first video game, basically table tennis on a computer.
"Poppycock" I don't think so.
Potpourri A fragrant, colorful mixture composed of pretty leaves, flower petals, etc.
Preponderance of the evidence A legal standard, in contrast to "beyond a reasonable doubt," that gives victory to whichever side the evidence most favors.
Prescient Accurately foresaw the future.
"Primordial ooze" Refers to the "soup" that life is said to have first originated in.
Quiver The thing that archers carry on their back that holds arrows.
Quotidian Boring, everyday duties that somebody has to do.
"A reasonable way to identify pornography" Refers to Justice Stewart's extremely informal test for obscenity in a famed Supreme Court case.
Re-christening This is what happens when a ship is given a new name.
Rube Goldberg machine A complicated contraption that accomplishes a simple task, named after the cartoonist who invented them.
Rule of reason A legal standard in antitrust law that, in contrast to a per se standard, limits intervention only to those actions that are, on balance, sufficiently anticompetitive.
Sallies Gentle thrusts or outward movements.
Sandbagging Deliberately underperforming, due to lack of effort or diligence.
School accountability movement A political movement, spearheaded in the American South, to assess how well schools are performing through their students' scores on standardized tests.
Secondary producer A producer of recycled metal.
Senescence Old as f—.
Sextant This odd-looking tool seafarers use in navigation.
"Shootin' distance" Pretty close but not too close.
Standardized tests Tests created by big companies used not for class grades but to determine how much a student knows in some more general sense.
"Swing through the ball" In sports, the act of completing the swing of a bat, golf club, etc., through and beyond its contact with the ball.
Synoptic Big-picture; allowing a general overview of the subject.

Systems engineers Just what it sounds like; what operations management would be if it focused on planning and design instead.

Tapestry A beautiful, elegant cloth, woven from threads of many colors, that royalty of old might have hung on the wall.

Tell A behavior or body movement that signifies something about the cards that a gambler is holding.

Terrigenous Just like it says in the text: sediments from the land going into the sea.

Thereunto appertaining A pompous phrase used in college graduations. I have no idea what it means.

"Tick-tock" Sequence of events, told as if in real time.

"To boot" In addition.

Tookish Adventurous beyond people would expect of you.

Transom A (typically) flat surface that forms the very back of a ship.

Trifecta A bet involving three predictions, which all must be correct to win.

Typification Letting the type represent the whole, as if Pepe the Frog and Kermit the Frog were equivalent because they are both frogs.

Unostentatiously Without drawing attention to itself (which, ironically, this word fails to do).

Variegated Something that is highly mottled in color or, by analogy, that contains a mixed variety of ingredients.

Vignette A little story, more brief than an anecdote.

Welter A disorderly mix.

"Window dressing" Something that looks nice but is not functional; something merely for appearance.

Index

D. Grant, *Methods of Economic Research*, Springer Texts in Business and Economics, https://doi.org/10.1007/978-3-030-01734-7