Preface

The world has changed in transformative ways. Data and evidence are ubiquitous. Quantitative information suffuses our talk of everything from policy to health care to job searches to politics to sports to education to dating to national security.

As a result, statistics and quantitative reasoning must no longer be the purview of only those who have a knack for mathematics or are headed for technical careers. Acquiring competence in foundational quantitative reasoning is now a fundamental responsibility of every educated human being and citizen. And this necessitates new ways of teaching and learning.

It was with that goal in mind that we decided to write *Thinking Clearly with Data*. But we didn't start with the book. Much of the material and ideas that ultimately found their way into the coming chapters were first developed for courses aimed at providing to students with little technical background the tools needed to be serious, thoughtful, and skeptical consumers of quantitative information. These courses include traditional university offerings, like introductions to quantitative reasoning taught to both undergraduate and graduate students at the University of Chicago. But they also include executive education courses offered to policy makers, military officers, national security experts, intelligence professionals, and journalists.

We learned a lot of lessons along the way that inform the choices we made in writing and organizing this book. Perhaps the most important was to create a shared language.

We knew we didn't want to teach a traditional statistics course. Such courses, in our view, are often too technical for many students and don't get to the most important and interesting issues, the ones that really matter for using quantitative information to make our lives and the world better. So, it was tempting to jump as quickly as possible to the exciting topics, like why correlation doesn't imply causation. But that would have been a mistake. A person can't understand why correlation doesn't imply causation until they understand what correlation and causation are.

For that reason, part 1 of this book is all about establishing a shared language. We define, conceptually and technically (but still accessibly), what we mean when we talk about correlation and causation—not in the sense of how to calculate a correlation coefficient or how to write down a causal effect in potential outcomes notation (though both will be covered), but in the sense of the questions, What do the words, properly understood and digested, mean in plain English? What's hard about correlation and causation? Why are they usefully separated? What are these two kinds of things, correlations and causal relationships, good for?

But what about the problem of motivation? If you don't put the good stuff up front, how do you keep people engaged? Well, first, who says a conceptual understanding of what causality does and does not mean isn't the good stuff? It is great stuff. But, more

to the point, our approach is this: if you want people to be engaged, make the material engaging. To us, this means several things.

The first is to tell stories. Throughout, you will find every conceptual discussion augmented by at least one extended, genuine, real-world example. Some of those examples will be about scientific studies. Many will be about personal experiences of ours where thinking clearly about quantitative evidence made a difference in the decisions we made. Others will involve reflections on the use of data and evidence in news, sports, policy, health care, and culture. This stuff really matters for how lives are lived and decisions are made in every realm of human endeavor. We want to keep that fact in the foreground. That is also why, despite the fact that this is a book by two political scientists, many of the examples are not drawn from politics.

The second way to engage readers is to emphasize ideas first and technicalities second. We love technicalities. But technicality can often be the enemy of understanding. When things get technical, lots of people stop thinking and start memorizing. We fervently wish to avoid that. So we always talk about the ideas and why they matter first. We treat things graphically whenever we can. And we do as little math as possible. But as little as possible isn't zero, for at least two reasons.

Familiarity with some technical matters is part of being a clear thinker. You can't understand mean reversion if you don't know what a mean or noise is. You can't understand publication bias and the replication crisis if you don't know what *statistical significance* means or what a p-value is. And it is hard to understand the problem of confounding or the answers offered by different research designs without being able to interpret a regression.

Moreover, sometimes being clear and precise requires a bit of math. We spend lots of time talking conceptually about counterfactuals and causality. But counterfactual talk can get a bit mystical. There is an extra degree of clarity that comes from writing down some potential outcomes and a proper definition of an effect that we think is indispensable. So we do not dispense with it. But, always, our emphasis is on clear thinking.

A third lesson for engaging writing is that it isn't enough for each chapter or lesson to tell a story. The whole course (or book) must do so. For us, the story is that making good decisions and doing good in our data-driven age requires clear thinking from each and every one of us. We can't just leave it to the experts, for many experts were never taught to think clearly about quantitative information. So we have to do it for ourselves or we will be frequently misled and may well make terrible mistakes.

Organization

That story informs the organization of the book. As we already noted, we open in part 1 by creating a shared language, focusing on the ideas of correlation and causation as the cornerstones of quantitative analysis.

With those ideas in hand, part 2 focuses on how we use data and evidence to figure out whether a correlation, causal or not, exists between features of the world. One of our goals in this part of the book is to convince everyone that there is plenty of good stuff, even before causal inference. Chapter 4 gets us motivated by explaining the incredibly common mistake of selecting on the dependent variable, by showing how trying to establish correlation without variation is impossible, and by illustrating the staggering number of instances when this mistake really matters. Chapter 5 turns to measuring correlations, focusing on a graphical explanation of regression. Chapter 6 introduces

statistical significance and hypothesis testing, framing everything in terms of a device we call our favorite equation, which recurs throughout the book:

$$\text{Estimate} = \text{Estimand} + \text{Bias} + \text{Noise}$$

If chapter 4 didn't already achieve this goal, chapter 7 makes it clear that there is plenty at stake in thinking clearly about what it means to establish a relationship in data by discussing the problems of p-hacking, publication bias, and related issues. Finally, chapter 8 covers the under-discussed topic of reversion to the mean and then uses it in conjunction with our prior discussion of publication bias to reflect on the replication crisis and the common phenomenon of scientific estimates shrinking over time.

Part 3 turns to causal inference, reminding readers of how important knowledge of causality is for making decisions about how to intervene in the world. Chapter 9 explains why correlation need not imply causation, discussing both confounders and reverse causality. Chapter 10 addresses the issue of statistical controls and provides some graphical explanations in the context of regression. Chapters 11–13 provide an overview of how scholars use research designs to try to learn about causality. Chapter 11 covers both randomized and natural experiments, introducing instrumental variables as a method for dealing with issues of noncompliance. Chapters 12 and 13 cover regression discontinuity and difference-in-differences designs, respectively. Chapter 14 closes this part of the book with a discussion of the challenges of learning about causal mechanisms.

Part 4 points out that we are not done once we've tackled causality. Even reliable knowledge of causal effects is not, on its own, sufficient to ensure that we are thinking clearly about how to use quantitative information to make good decisions. Chapter 15 points out how easy it is to fool yourself into thinking that a piece of quantitative information that answers one question in fact answers an entirely different one, encouraging readers to avoid this mistake by translating information presented technically into substance. In the course of so doing, we introduce Bayes' rule. Chapter 16 turns to issues of measurement, external validity, and extrapolation, which also leads us into a discussion of sample selection bias. And, finally, chapter 17 confronts some of the fundamental limits that quantitative analysis, no matter how clearly thought about, faces in informing decision making.

At the end of each chapter, there are exercises that readers can work through on their own to make sure they are grasping the material. Some of these exercises involve analyzing data, which can be tackled by readers and students who have learned (or are learning) how to use statistical software like Stata or R. The end of each chapter also has a "Readings and References" section that will allow curious readers to find the sources that are mentioned in the main text and dive more deeply into a particular topic.

Who Is This Book For?

We hope this book is for everybody interested in learning to think clearly about data, evidence, and quantitative reasoning. As we've mentioned, we have used these materials for a wide range of audiences, from undergraduates to highly accomplished professionals.

In our view, to prepare for living in our data-driven age, every undergraduate should meet material like this, ideally in their first couple of years of college. So we wrote the book in the hope that it would be helpful to instructors in many different disciplines who

teach quantitative reasoning, whether in general education courses or in an introductory course inside a department. We believe this will be especially true for instructors who want to take an approach that is more conceptual than the traditional statistics- or methods-based approach, while still covering some fundamental technical content.

We think the book works equally well for professional students. We have taught it to graduate students earning master's degrees in public policy. Some go on to take more technical courses in econometrics or program evaluation. But, for many, the essential skill is to learn to think critically and clearly about quantitative information. Our approach fits the needs of these students while, at the same time, providing the conceptual foundations that more technically inclined students will need in future courses.

Colleagues at other universities have also employed these materials in more advanced courses for social science majors who, for instance, must learn quantitative methods in preparation for writing a thesis. In that setting, our book may benefit from being coupled with another text that is somewhat more technical or places more emphasis on issues of statistical computing. In all of these settings, we hope the exercises at the end of each chapter will be helpful. These include applied data analyses, for which data are available to download online.

Finally, we also believe this book will be useful to many doctoral students. Often, in doctoral training, statistical material is taught quickly and at a high level of technicality. This can be productive; mastering advanced techniques is both challenging and important. But, in our experience, even the best doctoral students can lose sight of what really matters—how we learn about the world from data—as they focus on proving theorems and programming estimators. We very much hope that this book might serve as a guide to such students, helping to keep the big picture in clear view even as they work hard on the technical details.

Acknowledgments

As we mentioned, some of the materials in this book were developed as a joint effort for an undergraduate class of Anthony's and for an executive education class that Jake Shapiro, Liam Collins, Cathy Fetell, and Ethan designed together. Jake, Liam, and Cathy are due lots of credit and have our deep appreciation.

We'd like to thank Scott Ashworth, Chris Berry, Chris Blattman, Matt Brems, Bruce Bueno de Mesquita, Kerwin Charles, Devin Chesney, Lindsey Cormack, Andy Eggers, Nathan Favero, Alex Fouirnaies, Matt Gabel, Jeff Grogger, Andy Hall, Kosuke Imai, Renan Levine, Andrew Little, Jens Ludwig, Mordecai Magencey, Andrew Means, Pablo Montagnes, Emily Ritter, Steve Schwab, Mike Spagat, Dustin Tingley, Stephane Wolton, and Austin Wright for their incredibly helpful feedback.

Tom Budescu, Gautam Nair, Tom Naset, Jeff Ruff, Vanitha Virudachalam, Becky Wang, and Xingyu Yin provided terrific research assistance at the early stages of putting this project together. It was great fun to work with them and we are grateful for their contributions.

We'd like to thank our students for catching numerous mistakes and errors in previous drafts. AK Alilonu, Denise Azadeh, Ellie Rutkey, and Al Shah found a truly embarrassing number of them. Thank you!

The team at Princeton University Press was terrific. We are especially appreciative of Bridget Flannery-McCoy for believing in the project and for her guidance and to Alena Chekanov for overseeing so much of the process. And we owe a debt of gratitude to

Danna Lockwood for her incredible work helping us improve the writing, presentation, and organization of the book. We also very much appreciate Melody Negron's terrific oversight of production and David Luljak's always excellent indexing.

Ethan would like to thank his colleagues at the Harris School and his many collaborators, coauthors, and students, all of whom are an intellectual inspiration and have improved the clarity of his own thinking with years of heroic effort. He is deeply grateful to his wife, Rebecca, who put up with the grumpiness that inevitably accompanies finishing a book with the support, love, and patience with which she meets all the joys of life with Ethan. This one is dedicated to his kids, Hannah and Abe, who are a joy and a pleasure. His fondest hope is that this book will find its way into enough classrooms that, one day, they might have the mortifying experience of being assigned to read a book dedicated to themselves. That would be quantifiable success.

Anthony would like to thank his advisors, coauthors, and colleagues who make him a clearer thinker with every interaction. He is grateful to his parents, who have have encouraged and supported him throughout his life, even though he currently has no plans to go to law school or medical school. And most importantly, he thanks Gloria, his wife and best friend, who has read countless drafts of academic papers, endured far too many conversations about regressions, dabbled in data entry, vetted every idea, contributed a disproportionate share of her own, and enriched his life in ways that defy quantification.