



RETHINKING

**SECOND
EDITION**



SOCIAL INQUIRY



Diverse Tools, Shared Standards

EDITED BY HENRY E. BRADY AND DAVID COLLIER



Rethinking Social Inquiry

Rethinking Social Inquiry

Diverse Tools, Shared Standards

Second Edition

Edited by
Henry E. Brady and David Collier

ROWMAN & LITTLEFIELD PUBLISHERS, INC.
Lanham • Boulder • New York • Toronto • Plymouth, UK

Published by Rowman & Littlefield Publishers, Inc.
A wholly owned subsidiary of The Rowman & Littlefield Publishing Group, Inc.
4501 Forbes Boulevard, Suite 200, Lanham, Maryland 20706
<http://www.rowmanlittlefield.com>

Estover Road, Plymouth PL6 7PY, United Kingdom

Copyright © 2010 by Rowman & Littlefield Publishers, Inc.

Chapters 6–9 from the previous edition are available as downloadable files from the Rowman & Littlefield website. For more information about these chapters and to obtain the required username and password visit <http://www.rowmanlittlefield.com/RL/books/RSI2e/> or e-mail textbooks@rowman.com.

All rights reserved. No part of this book may be reproduced in any form or by any electronic or mechanical means, including information storage and retrieval systems, without written permission from the publisher, except by a reviewer who may quote passages in a review.

British Library Cataloguing in Publication Information Available

Library of Congress Cataloging-in-Publication Data

Rethinking social inquiry : diverse tools, shared standards / edited by Henry E. Brady and David Collier.—2nd ed.

p. cm.

Includes bibliographical references and index.

ISBN 978-1-4422-0343-3 (cloth : alk. paper)


ISBN 978-1-4422-0344-0 (pbk. : alk. paper)

1. Social sciences—Research. 2. Social sciences—Methodology. I. Brady, Henry E. II. Collier, David.

H62.R4646 2010

300.72—dc22

2010022477

™ The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials, ANSI/NISO Z39.48-1992.

Printed in the United States of America

To the memory of David A. Freedman

Brilliant statistician,

Guardian against precarious statistical models,

Champion of joining quantitative and qualitative analysis,

Friend of remarkable wit and generosity

And indeed, social science methodology

is now catching up with him

Contents

List of Figures and Tables	xi
Preface to the Second Edition	xiii
Preface to the First Edition	xv
Introduction to the Second Edition: A Sea Change in Political Methodology	1
<i>David Collier, Henry E. Brady, and Jason Seawright</i>	
Part I. A Debate on Methodology	11
A. Framing the Debate	13
1. Refocusing the Discussion of Methodology	15
<i>Henry E. Brady, David Collier, and Jason Seawright</i>	
2. The Quest for Standards: King, Keohane, and Verba's <i>Designing Social Inquiry</i>	33
<i>David Collier, Jason Seawright, and Gerardo L. Munck</i>	
B. Critiques of the Quantitative Template	65
3. Doing Good and Doing Better: How Far Does the Quantitative Template Get Us?	67
<i>Henry E. Brady</i>	
4. Some Unfulfilled Promises of Quantitative Imperialism	83
<i>Larry M. Bartels</i>	

5. How Inference in the Social (but Not the Physical) Sciences Neglects Theoretical Anomaly <i>Ronald Rogowski</i>	89
C. Linking the Quantitative and Qualitative Traditions	99
6. Bridging the Quantitative-Qualitative Divide <i>Sidney Tarrow</i>	101
7. The Importance of Research Design <i>Gary King, Robert O. Keohane, and Sidney Verba</i>	111
D. Diverse Tools, Shared Standards	123
8. Critiques, Responses, and Trade-Offs: Drawing Together the Debate <i>David Collier, Henry E. Brady, and Jason Seawright</i>	135
9. Sources of Leverage in Causal Inference: Toward an Alternative View of Methodology <i>David Collier, Henry E. Brady, and Jason Seawright</i>	161
Part II. Causal Inference: Old Dilemmas, New Tools	201
Introduction to Part II <i>David Collier, Henry E. Brady, and Jason Seawright</i>	
E. Qualitative Tools for Causal Inference	205
10. Process Tracing and Causal Inference <i>Andrew Bennett</i>	207
11. On Types of Scientific Inquiry: The Role of Qualitative Reasoning <i>David A. Freedman</i>	221
12. Data-Set Observations versus Causal-Process Observations: The 2000 U.S. Presidential Election <i>Henry E. Brady</i>	237
Addendum: Teaching Process Tracing <i>David Collier</i>	243
F. Quantitative Tools for Causal Inference	245
13. Regression-Based Inference: A Case Study in Failed Causal Assessment <i>Jason Seawright</i>	247

14. Design-Based Inference: Beyond the Pitfalls of Regression Analysis? <i>Thad Dunning</i>	273
Glossary <i>Jason Seawright and David Collier</i>	313
Bibliography	361
Acknowledgment of Permission to Reprint Copyrighted Material	387
Subject Index	389
Name Index	397
About the Contributors	405

Figures and Tables

FIGURES

2.1	Steps in the Research Cycle: A Framework for Summarizing <i>Designing Social Inquiry</i>	50
8.1	Evaluating Necessary and/or Sufficient Causes	147
13.1	Mean Growth Rates by Level of Democracy	252
13.2	Stable Empirical Results Are Compatible with Three Alternative Causal Models	257
14.1	Plausibility of <i>As-If</i> Random Assignment	292
14.2	Credibility of Statistical Models	296
14.3	Substantive Relevance of Intervention	301
14.4	Typology of Natural Experiments	305

TABLES

2.1	Quantitative Tools Employed in <i>Designing Social Inquiry</i>	39
2.2	Anticipating the Debate on <i>Designing Social Inquiry</i>	58
6.1	Tools for Bridging the Qualitative-Quantitative Divide	104
9.1	Four Approaches to the Qualitative-Quantitative Distinction	179
9.2	Data-Set Observation versus Causal-Process Observation	185
9.3	Adding Different Forms of Data: Consequences for Causal Inference	192
10.1	Process Tracing: Four Tests for Causation	210
13.1	Economic Growth Regressed on Democracy and Lagged Income Level	253

13.2	Economic Growth Regressed on Democracy, Lagged Income Level, and Primary Educational Attainment	255
13.3	Alternative Combinations of Control Variables	260
14.1	Examples of Natural Experiments, Including Regression Discontinuity (RD) and Instrumental Variable (IV) Designs	282

Preface to the Second Edition

Rethinking Social Inquiry seeks to redirect ongoing discussions of methodology in political and social science. This Preface presents our two goals in launching a second edition.

The first goal (a central focus of Part I) is to sustain the debate with King, Keohane, and Verba's (KKV)¹ *Designing Social Inquiry*. Nine chapters from the first edition are included here to continue this exchange. Although published more than 15 years ago, KKV remains a fundamental point of reference in political science methodology and in controversies on methods—as we discuss in the Introduction to the Second Edition. Through articulating the approach we call “mainstream quantitative methods,” KKV has wide importance in the political science discipline—and, correspondingly, in graduate student training. While we admire aspects of the book's contribution, our strong dissent from many of the arguments remains highly salient today. KKV has played a key role in narrowing attention to a particular set of quantitative tools, and the methodological horizon of political science has been shortened by the book's continuing influence. Sustaining this debate in 2010 therefore remains as necessary as it was when our first edition appeared in 2004.

The second goal is to open new avenues of discussion in methodology, both qualitative and quantitative. A number of chapters from the first edition—in particular chapters 8 and 9—explore these wider themes. In addition, a new set of chapters has been incorporated as Part II of the second edition. These chapters offer an innovative view of the crucial qualitative tools of process tracing and causal process observations, as well as an extended new discussion of the weaknesses and strengths of regression analysis and other quantitative tools.

1. To avoid personalizing the debate, we previously adopted the abbreviation *DSI* in referring to the book. However, the abbreviation KKV is now ubiquitous, and we have deferred to standard usage. In the present edition, *DSI* has been replaced by KKV.

A detailed overview of the new chapters is provided in the Introduction to Part II below. A central theme of these chapters is the importance of methodological pluralism and the value of multi-method research. Qualitative analysis is strengthened when used in conjunction with quantitative research; and quantitative analysis, in turn, contributes more if it is built on a foundation of qualitative analysis and insight.

Two distinctive features of the second edition must be underscored. The first is the online placement, on the Rowman & Littlefield website, of four chapters from the first edition that are not included here. The online chapters are part of the original debate with KKV, and they also extend the discussion well beyond that debate.² Thus, we are able to retain all the original chapters and accommodate the new chapters in Part II, with little change in the overall length of the printed book. These chapters are accessible by following the instructions on the copyright page of this volume.

Second, with the goal of advancing the understanding of process tracing and improving the teaching of this method, the online resources include a set of exercises. These challenge readers to push further in examining the case study evidence provided in the chapters by Bennett, Freedman, and Brady. The exercises also focus on additional readings, including the Sherlock Holmes story “The Adventure of Silver Blaze,” an excellent illustration of process tracing.

We are grateful for the extensive help we have received in preparing the second edition. It was our good fortune that the late David Freedman, prior to his untimely death in 2008, had already made many suggestions for this edition. Kimberly Twist—drawing on her long experience with professional editing and manuscript preparation—secured permissions from publishers and skillfully coordinated and assembled the book. Taylor Boas, Christopher Chambers-Ju, Fernando Daniel Hidalgo, Jody LaPorte, Simeon Nichter, and Neal Richardson drew on their strong methodological training to provide incisive comments on the new chapters. Alexis Dalke, Zoe Fishman, Maria Gould, Annette Konoske-Graf, and Miranda Yaver worked tirelessly in checking, correcting, and editing chapters, and as always, Jennifer Jennings provided astute advice. Niels Aaboe and Elisa Weeks of Rowman & Littlefield contributed both suggestions and great patience.

Henry E. Brady
David Collier
Berkeley, California
May 2010

2. These chapters are “Warnings About Selection Bias” by David Collier, James Mahoney, and Jason Seawright; “Tools for Qualitative Research” by Gerardo Munck; “Turning the Tables” by Charles Ragin; and “Case Studies” by Timothy McKeown.

Preface to the First Edition

Crafting good social science research requires diverse methodological tools. Such tools include a variety of qualitative and quantitative approaches: small-N and large-N analysis, case studies and structural equation modeling, ethnographic field research and quantitative natural experiments, close analysis of meaning and large-scale surveys. Yet diverse tools are not enough. Without shared standards, social science can lose its way. Shared standards help ensure that the application of these tools leads to meaningful conceptualization and measurement, interpretable causal inferences, and a better understanding of political and social life.

We come to the enterprise of editing this volume with different methodological starting points, yet with the joint conviction that our approaches converge in major respects. Henry E. Brady, who is primarily a quantitative survey researcher, repeatedly finds that he must come to grips with interpreting the meanings conveyed in survey responses and with comprehending the qualitative complexity of the political behavior he studies in various national contexts. David Collier, who is primarily a qualitative comparativist, recognizes that it is sometimes productive to quantify concepts such as corporatism and democracy, the historical emergence of labor movements, and the international diffusion of policy innovations. Our joint teaching and extensive discussions have reinforced our commitment to diverse tools, as well our conviction that we share basic standards for evaluating their use.

This concern with diverse tools and shared standards provides the framework for the present volume. Within that framework, a central focus is on a major scholarly statement about the relationship between quantitative and qualitative methods—Gary King, Robert O. Keohane, and Sidney Verba's book, *Designing Social Inquiry* (hereafter *DSI*). *DSI* is deservedly influential and widely read, in part because it offers an accessible statement of the ana-

lytic position that we call “mainstream quantitative methods.”¹ The book likewise makes the important claim that quantitative methods can solve many problems faced by qualitative researchers.

Notwithstanding *DSI*'s major contribution, we have misgivings about important parts of the book's argument. First of all, *DSI* does not adequately address basic weaknesses in the mainstream quantitative approach it advocates. The book does not face squarely the major obstacles to causal assessment routinely encountered in social science research, even when sophisticated quantitative techniques are employed. *DSI*'s treatment of concepts, operationalization, and measurement is also seriously incomplete.

Further, we disagree with the claim that *DSI* provides a general framework for “scientific inference in qualitative research,” as the authors put it in the book's subtitle. The book's failure to recognize the distinctive strengths of qualitative tools leads the authors to inappropriately view qualitative analysis almost exclusively through the optic of mainstream quantitative methods.

We are convinced that the perspective offered by ideas drawn from what we call “statistical theory”²—in contrast to *DSI*'s perspective of mainstream quantitative methods—provides a more realistic approach to evaluating qualitative tools. Statistical theory sometimes points to valuable justifications for practices of qualitative researchers that *DSI* devalues. We therefore consider not only how qualitative research can be justified in its own terms, but also the idea of statistical rationale for qualitative research.

Our project began with the idea of reprinting several insightful review essays focused on *DSI*, which we had intended to bring together as a small volume with some opening and concluding observations of our own. As sometimes happens with book projects, this one expanded greatly, and the newly written material constitutes well over half the text.³ The book includes an entire chapter that summarizes *DSI*'s recommendations (chap. 2), as well as two substantial concluding chapters (chaps. 8 and 9 in the second edition), an appendix, and a glossary.

Especially in a book with multiple authors, the reader may find it helpful to be able to locate quickly the overall summaries of the arguments. These

1. We define mainstream quantitative methods as an approach based on regression analysis, econometric refinements on regression, and the search for statistical alternatives to regression models in contexts where specific regression assumptions are not met.

2. We understand statistical theory as a broad, multidisciplinary enterprise concerned with reasoning about evidence and inference. Important scholars in the tradition of statistical theory have expressed considerable skepticism about the application to observational data of the regression-based methodology identified with mainstream quantitative methods.

3. Acknowledgment of permission to reprint copyrighted material is presented at the end of this book.

are found in the first part of chapter 1 (pp. 15–26); pp. 52–63 at the end of chapter 2; and pp. 196–199 at the end of chapter 9, as well as chapters 8 and 9 more broadly. The second part of chapter 1 provides a chapter-by-chapter overview of the volume. The glossary defines key concepts: the core definition is presented in the initial paragraph of each entry, and additional paragraphs are included for concepts that require more elaboration.

We wish to acknowledge our intellectual debt to the many people who have contributed to this project. It has been an enormous pleasure to work with Jason Seawright, whose immense contribution is reflected in the coauthorship of five chapters and the glossary. His mastery of methodological and statistical issues, combined with a remarkable command of substantive agendas, has made him an exceptional collaborator. David A. Freedman of the Berkeley Statistics Department has been a paragon of collegiality, again and again providing new ideas, specific suggestions about the text, and outstanding commentary on broader methodological issues. We also thank the other authors of the chapters within the book for their participation in the project.

David Collier's earlier book, *The New Authoritarianism in Latin America* (1979), which sought to systematically organize a substantive and methodological debate in comparative social science, provided a model for the structure of the present volume, and also for the spirit of constructive criticism that animates it. Correspondingly, renewed thanks are due to two colleagues who played a special role in shaping that earlier book: Louis W. Goodman and the late Benjamin A. Most.

We extend our gratitude to Christopher H. Achen and Larry M. Bartels, whose breadth of vision, elegant approach to methodological problems, and simple good sense have helped to stimulate our thinking about the importance of research design and the use of techniques appropriate to the task at hand. Neal Beck, Alexander L. George, Giovanni Sartori, J. Merrill Shanks, Paul Sniderman, and Laura Stoker have also been key colleagues in discussions of methodological and substantive issues.

Our work on this project convinces us again that institutional context matters. The strong commitment of the Berkeley Political Science Department to methodological and analytic pluralism encouraged us to write this book. At the national level, we have been inspired by the initiative and enterprise of a younger cohort of scholars who have reinvigorated efforts to bridge qualitative and quantitative methods, and some of whom have played a key role in forming the Consortium for Qualitative Research Methods (CQRM), and also the Organized Section on Qualitative Methods of the American Political Science Association. At the potential risk of omitting key names, we would especially mention, among these younger scholars, Andrew Bennett, Bear Braumoeller, Michael Coppedge, David Dessler, Colin Elman, John Gerring, Gary Goertz, Evan Lieberman, James Mahoney, Gerardo L. Munck, Andreas Schedler, and David Waldner.

Several people have made an unusually large contribution through providing either very extensive substantive suggestions or sustained assistance in coordinating the manuscript: Robert Adcock, Michelle Bonogofsky, Maiah Jaskoski, Diana Kapiszewski, Sebastián Mazzuca, Reilly O'Neal, Sara Poster, and Sally Roever.

We also received insightful comments from Michael Barzelay, Andrew Bennett, Mark Bevir, Taylor Boas, George Breslauer, Christopher Cardona, Jennifer Collier, Ruth Berins Collier, Stephen Collier, Michael Coppedge, Rubette Cowan, David Dessler, Jorge Dominguez, Paul Dosh, Ralph Espach, Sebastián Etchemendy, Andrew Gould, Kenneth Greene, Ernst Haas, Peter Houtzager, William Hurst, Simon Jackman, Jonathan Katz, Jee-won Kim, Peter Kingstone, Daniel Kreider, Lien Lay, James Mahoney, Scott Mainwaring, Walter Mebane, Geraldo L. Munck, Guillermo O'Donnell, Wagner Pralon, Charles Ragin, Jessica Rich, Eric Schickler, Carsten Schneider, Taryn Seawright, Jasjeet Sekhon, Wendy Sinek, Jeffrey Sluyter-Bultrao, Alfred Stepan, Laura Stoker, Tuong Vu, Michael Wallerstein, and Alexander Wendt.

Excellent feedback was likewise provided by colleagues who attended presentations on the project at the Kellogg Institute, University of Notre Dame; the Departments of Political Science at Columbia University and at the University of Minnesota; the Institute of Development Studies, London School of Economics; and meetings of the American Political Science Association, the Midwest Political Science Association, the Western Political Science Association, the Institute for Qualitative Research Methods at Arizona State University, the Political Methodology Society, and the Southern California Political Behavior Seminar.

Bruce Cain, Director of the Berkeley Institute of Governmental Studies, has been very supportive throughout the project. Gerald C. Lubenow and Maria A. Wolf of the Berkeley Public Policy Press, and also Jennifer Knerr of Rowman & Littlefield, provided untiring assistance with issues of manuscript preparation and editing. The project received financial support from the Survey Research Center, the Department of Political Science, the Institute of International Studies, and International and Area Studies, all at the University of California, Berkeley.

Henry Brady was supported during 2001–2002 as a Hewlett Fellow (98–2124) at the Center for Advanced Study in the Behavioral Sciences, as well as through a grant (2000–3633) from the William and Flora Hewlett Foundation. Jason Seawright's work on the project was funded by a National Science Foundation Graduate Research Fellowship.

Henry E. Brady
David Collier
Berkeley, California

Introduction to the Second Edition: A Sea Change in Political Methodology

David Collier, Henry E. Brady, and Jason Seawright

We begin with rival claims about the “science” in social science.¹ In our view, juxtaposing these claims brings into focus a sea change in political science methodology.

King, Keohane, and Verba’s (KKV) 1994 book, *Designing Social Inquiry*,² proposes a bold methodological agenda for researchers who work in the qualitative tradition. The book’s subtitle directly summarizes the agenda: “*scientific inference in qualitative research*” (italics added). To its credit, the book is explicit in its definition of science. It draws on what we and many others have viewed as a “quantitative template,” which serves as the foundation for the desired scientific form of qualitative methods. In KKV’s view, standard research procedures of qualitative analysis are routinely problematic, and ideas drawn from conventional quantitative methods are offered as guideposts to help qualitative researchers be scientific.

1. For our own work, we share Freedman’s view of plurality in scientific methods, and we recognize social versus natural science as partially different enterprises. Yet the two can and should strive for careful formulation of hypotheses, intersubjective agreement on the facts being analyzed, precise use of data, and good research design. With this big-tent understanding of science, we are happy to be included in the tent.

2. As explained above in the preface, in the second edition we use the abbreviation KKV to refer to the book, rather than *DSI*, as in the first edition.

A starkly different position has been emerging over a number of years, forcefully articulated by the statistician David A. Freedman in chapter 11 of the present volume. He reviews the central role of qualitative analysis in six major breakthroughs from the history of epidemiology—a field highly relevant to political science because it faces many of the same challenges of doing large-N analysis with observational data and because, as Freedman insists, one does indeed find interesting opportunities for qualitative insight. He argues, in fact, that in epidemiology as well as the social sciences, qualitative analysis is indeed a “type of *scientific inquiry*” (italics added), within the framework of recognizing multiple types. In characterizing this form of quantitative analysis, Freedman employs the expression “causal-process observation” (CPO—a term of central importance to the present volume).³ In his view, such strategically selected pieces of evidence play a critical role in disciplined causal inference. Freedman comments pointedly on the contributions of CPOs.

Progress depends on refuting conventional ideas if they are wrong, developing new ideas that are better, and testing the new ideas as well as the old ones. The examples show that qualitative methods can play a key role in all three tasks . . . (chap. 11, this volume)

Relatedly, Freedman underscores the fragility of the quantitative template.

Indeed, far-reaching claims have been made for the superiority of a quantitative template that depends on modeling—by those who manage to ignore the far-reaching assumptions behind the models. However, the assumptions often turn out to be unsupported by the data. . . . If so, the rigor of advanced quantitative methods is a matter of appearance rather than substance. (chap. 11, this volume)

In this Introduction, against the backdrop of these starkly contrasting views of appropriate methods, we examine new developments in methodology that have framed our approach to the second edition of *Rethinking Social Inquiry*. The discussion focuses on: (1) ongoing controversy regarding KKV’s legacy; (2) growing criticism of the standard quantitative template, including regression modeling, significance tests, and estimates of uncertainty; and (3) emerging arguments about both qualitative and quantitative methods that hold the promise of greatly strengthening tools for causal inference.

3. We define a causal-process observation as an insight or piece of data that provides information about context, process, or mechanism, and that contributes distinctively to causal inference. A data-set observation (DSO), by contrast, is the standard quantitative data found in a rectangular data set. See Glossary.

A further initial point should be underscored. The focus in both editions of *Rethinking Social Inquiry* is on the study of causes and consequences—and specifically on causal inference. Of course, this focus is just one facet of methodology. In our own work we have written extensively on conceptualization and measurement, and indeed, assessing causes and consequences emphatically calls for careful attention to concept formation and operationalization. Yet the central concern here is with causal inference.

ONGOING CONTROVERSY OVER KKV

The methodological positions adopted by KKV continue to be of great importance in political science and well beyond. The book has an exceptionally high level of citations, and year after year it has impressive sales rankings with online book sellers.

In the period since the publication of our first edition in 2004, quantitative and qualitative methodologists alike have underscored KKV's importance. Philip A. Schrodtt, a quantitative methodologist, argues that it has been the "canonical text of the orthodox camp" among political methodologists. In many graduate programs, it is considered "the complete and unquestionable truth from on high" (Schrodtt 2006: 335). On the qualitative side, James Mahoney notes the book's striking importance and remarkable impact in political science (2010: 120).

Ironically, achieving "doctrinal status was not necessarily the intention of KKV's authors" (Schrodtt 2006: 336), and their perspectives have doubtless evolved in the intervening years. Yet notably, in 2002—eight years after the book's original publication—King published an extended, programmatic statement on methodology, nearly the length of a short book, entitled "The Rules of Inference" (Epstein and King 2002). This publication departs little from the arguments of KKV.⁴

KKV is controversial, as well as influential, and its continuing importance is of great concern to scholars disturbed by its narrow message. Our first edition already contained strong critiques, and new commentaries—some extremely skeptical—have continued to appear. These more recent arguments merit close examination.

Schrodtt presents a bruising critique:

4. We were grateful for King, Keohane, and Verba's willingness to contribute their article "The Importance of Research Design" to our first edition, and we are very pleased to include it in this new edition. It contributes important ideas to the debate among authors who have commented on their original book. However, we do not see it as a substantial departure from their book.

KKV establishes as the sole legitimate form of social science a set of rather idiosyncratic and at times downright counterintuitive frequentist statistical methodologies that came together . . . to solve problems quite distant from those encountered by most political scientists. . . . (2006: 336)

Schrodt views the book as promoting “a statistical monoculture” that is “not even logically consistent” (2006: 336). In his view, this raises the concern that

one of the reasons our students have so much difficulty making sense of [KKV] is that in fact it does not make sense. (2006: 336)

Mahoney (2010), in his comprehensive essay “After KKV: The New Methodology of Qualitative Research,” argues that KKV has “hindered progress in political science” by “controversially and perhaps unproductively promoting a singular quantitative approach” (2010: 121). Weyland, with obvious annoyance, suggests that the authors of KKV “offered to help out their inferentially challenged qualitative brethren,” proposing that their work should be “as similar as possible to quantitative studies.” The book in effect makes claims of “quantitative superiority” that “rest on problematic assumptions” (2005: 392), thereby reinforcing the mindset in which “qualitative research was often seen as lacking precision and rigor and therefore undeserving of the ‘methods’ label” (2005: 392).

These and other scholars have also noted the sharp contrast in views between KKV and our own book. For example, Benoît Rihoux sees a “polarized” discussion that reflects a “fierce methodological debate which cuts across the whole of empirical social science in North America” (2006: 333, 334).

In discussing our book, Schrodt suggests that in this polarized context, “adherents of the [methodological] orthodoxy consider the heresies proposed therein to be a distraction at best; a slippery slope . . . at worst” (2006: 335). To take one example, what we would view as one of the orthodox commentaries is found in Nathaniel Beck (2006, *passim*), who entitles his article “Is Causal-Process Observation an Oxymoron?”—thereby essentially dismissing a basic concept in our book. He repeatedly acknowledges that scholars should “understand their cases” (e.g., 350) and that qualitative evidence contributes to this background knowledge, but he questions the idea that causal-process observations meet acceptable standards for causal inference (352).

Schrodt views elements of the response to *Rethinking Social Inquiry* among mainstream quantitative methodologists as reflecting an unfortunate, defensive reaction. He argues that

many in the statistical community have taken criticism of any elements of the orthodox approach as a criticism of all elements and circled the wagons rather than considering seriously the need for some reform. (Schrodt 2006: 338)

He also notes that when the editor of the methodology journal *Political Analysis* announced at the 2005 summer methodology meetings that the journal planned a symposium on *Rethinking Social Inquiry*, the room responded as if to express concern that “there are traitors in our midst!” (2006: 338). Schrodt comments that this resistance reflects “a worrisome contentment with the status quo” among quantitative methodologists (2006: 338).

Based on this discussion, it seems clear that major controversies over methods stand behind these criticisms. We now explore two of these controversies.

CRITICISM OF THE STANDARD QUANTITATIVE TEMPLATE

Our discussion here focuses on two facets of current criticism of the standard quantitative template, concerning basic ideas about statistical modeling and regression analysis, and alternative approaches to the important task of estimating uncertainty.

Statistical Modeling and Regression Analysis

In the past few years, the standard quantitative template centered on regression analysis has come under even heavier criticism. This development has two implications here. First, given KKV’s reliance on this template, it further sharpens concern about the book’s influence. Second, looking ahead, this development greatly extends the horizon of methodological approaches that should be—and in fact are being—discussed and applied, both among methodologists and consumers of alternative methods.

Much of this discussion centers on the enterprise of statistical modeling that stands behind regression analysis. In important respects, the precariousness of work with regression derives from the extreme complexity of statistical models. A statistical model may be understood as “a set of equations that relate observable data to underlying parameters” (Collier, Sekhon, and Stark 2010: xi—see Glossary). The values of these parameters are intended to reflect descriptive and causal patterns in the real world.

Constructing a statistical model requires assumptions, which often are not only untested, but largely untestable. These assumptions come into play “in choosing which parameters to include, the functional relationship between the data and the parameters, and how chance enters the model” (Collier, Sekhon, and Stark 2010: xi). Thus, debates on the precariousness of regression analysis are also debates on the precariousness of statistical

models. It is unfortunate that more than a few quantitative researchers believe that when the model is estimated with quantitative data and results emerge that appear interpretable, it validates the model. This is not the case.

We agree instead with the political scientist Christopher H. Achen, who argues that with more than two or three independent variables, statistical models will “wrap themselves around any dataset, typically by distorting what is going on” (2002: 443). Thus, what we might call a “kitchen sink” approach—one that incorporates numerous variables—can routinely appear to explain a large part of the variance without yielding meaningful causal inference. Relatedly, Schrodtt states that with just small modifications in the statistical model, estimates of coefficients can

bounce around like a box of gerbils on methamphetamines. This is great for generating large bodies of statistical literature . . . but not so great at ever coming to a conclusion. (2006: 337)

The econometrician James J. Heckman emphasizes that “causality is a property of a model,” not of the data, and “many models may explain the same data” (2000: 89). He observes that “the information in any body of data is usually too weak to eliminate competing causal explanations of the same phenomenon” (91).⁵

Sociologists have expressed related concerns, and Richard A. Berk concisely presents key arguments:

Credible causal inferences cannot be made from a regression analysis alone. . . . A good overall fit does not demonstrate that a causal model is correct. . . . There are no regression diagnostics through which causal effects can be demonstrated. There are no specification tests through which causal effects can be demonstrated. (2004: 224)

Berk amusingly summarizes his views in section headings within the final chapter of his book on regression analysis: “Three Cheers for Description,” “Two Cheers for Statistical Inference,” and “One Cheer for Causal Inference” (2004: chap. 11).⁶

Mathematical statisticians have likewise confronted these issues. Freedman’s skepticism about regression and statistical modeling has already been noted above, and his incisive critiques of diverse quantitative methods have now been brought together in an integrated volume that ranges across a broad spectrum of methodological tools (Freedman 2010).

5. From the standpoint of econometrics, see also Leamer (1983, 36–38).

6. Related arguments of sociologists have been advanced by Morgan and Winship (2007: *passim*), Hedström (2008: 324), and many other authors who have developed these themes. Statements by psychometricians include Cliff (1983, 116–18) and Loehlin (2004, 230–34).

Also from the side of mathematical statistics, Persi Diaconis argues that “large statistical models seem to have reached epidemic proportions” (1998: 797), and he laments the harm they are causing. He states that “there is such a wealth of modeling in the theoretical and applied arenas that I feel a sense of alarm” (804). Given these problems, methodologists should take more responsibility for the epidemic of statistical models by advocating “defensive statistics” (1998: 805). Thus, it should be a professional obligation to proactively warn scholars about the host of methodological problems summarized here.

In sum, many authors are now expressing grave concern about methods that have long been a mainstay of political and social science, and that are foundational in KKV’s approach.

Estimating Uncertainty

Standard practices in mainstream quantitative methods for estimating the uncertainty of research findings have also been challenged. The quest to estimate uncertainty is quite properly a high priority, prized as a key feature of good research methods. KKV views understanding and estimating uncertainty as one of four fundamental features of scientific research (1994: 9). In its discussion of “defining scientific research in the social sciences,” the book states that “without a reasonable estimate of uncertainty, a description of the real world or an inference about a causal effect in the real world is uninterpretable” (9). The received wisdom on these issues is central to mainstream quantitative methods.

Unfortunately, KKV presumes too much about how readily uncertainty can be identified and measured. In conjunction with the original debate over KKV, for example, Larry M. Bartels (chap. 4, this volume: 86–87) argues that these authors greatly overestimate the value of the standard insight that random error on an independent variable biases findings in knowable ways, whereas such error on the dependent variable does not. Bartels demonstrates that this would-be insight is incorrect.

A more pervasive problem involves significance tests. Any scholar acquainted with conventional practice in reporting regression results is well aware of the standard regression table with “tabular asterisks” scattered throughout.⁷ The asterisks indicate levels of statistical significance, calculated on the basis of the standard errors of the coefficients in the table. Too often, when researchers report their causal inferences they simply identify the coefficients that reach a specified level of statistical significance. This is a dubious research practice.

A central problem here is that findings reported in regression tables are

7. Meehl (1978), cited in Freedman and Berk (2010: 24).

routinely culled from numerous alternative specifications of the regression model, which obviates the standard meaning and interpretation of the asterisks. Once again, Schrodtt states the objection with particular clarity:

The ubiquity of exploratory statistical research has rendered the traditional frequentist significance test all but meaningless. (2006: 337)

Freedman and Berk (2010: 24) underscore the dependence of significance tests on key assumptions. For descriptive inference (external validity), they assume a random sample, rather than the convenience sample common in political science. Even with a random sample, missing data—including the problem of non-respondents—can make it more like a convenience sample.⁸ Another assumption requires a well-defined—rather than ill-defined or somewhat arbitrarily defined—population. For causal inference (internal validity), avoiding data snooping is crucial if significance tests are to be meaningful. Here, the presumption is that the researcher has begun with a particular hypothesis and tested it only once against the data, rather than several times, adjusting the hypothesis and model specification in the search for results deemed interesting. This inductive approach is *definitely* a valuable component of creative research, but it muddies the meaning of significance tests.

Against this backdrop, Freedman, Pisani, and Purves (2007) are blunt and—as usual—entertaining in their warnings on significance tests.

1. “If a test of significance is based on a sample of convenience, watch out” (556).
2. “If a test of significance is based on data for the whole population, watch out” (556).
3. “Data-snooping makes P-values hard to interpret” (547).
4. “An ‘important’ difference may not be statistically significant if the N is small, and an unimportant difference can be significant if the N is large” (553).⁹

A key point should be added. In his various single-authored and co-authored critiques of significance tests, Freedman does not turn to the alternative of Bayesian analysis. Rather, as in his other writings on methodology (see, e.g. chap. 11, this volume), he advocates common sense, awareness

8. See Freedman (2008b: 15). Thus, starting with a random sample, in the face of problems such as resource constraints that limit tracking down respondents, the researcher can end up with what is in effect a type of convenience sample.

9. I.e., if assumptions are not met, “significance” level depends on the sample size, without reflecting the real meaning of statistical significance.

that statistical tools have major limitations, and substantive knowledge of cases as an essential foundation for causal inference.

WHERE DO WE GO FROM HERE?

The practical importance of these problems is quickly seen in the fact that, to a worrisome degree, a great deal of quantitative research in political science has proceeded as if regression-based analysis, including associated measures of uncertainty, yields reliable causal inference. A vast number of journal articles have sought to make causal inferences by estimating perhaps half a dozen related (though quite typically under-theorized) model specifications, picking and choosing among these specifications, and offering an ad hoc interpretation of a few selected coefficients—generally, quite inappropriately, on the basis of significance levels. These failings have been further exacerbated by the readily available statistical software that makes it easy for researchers with virtually no grasp of statistical theory to carry out complex quantitative analysis (Steiger 2001).

In the face of these grave problems, we explore two avenues of escape: first, new developments in quantitative analysis; and second, continuing innovation in qualitative methods, which offer a very different means of addressing these difficulties. In our own work, and in scholarship more broadly, quantitative methods are of course deemed to be of enormous importance in their own right, and this continuing innovation certainly contributes more broadly to strengthening these tools.

Quantitative Methods

One hope has been that solutions can be found in refinements on regression analysis. This aspiration has motivated the new chapters by Jason Seawright and Thad Dunning (chaps. 13 and 14), which explore both some disasters of causal inference in quantitative research, and also potential solutions. They consider, for example, matching designs and the family of techniques associated with natural experiments—including regression discontinuity designs and instrumental variables. In some substantive domains, as Seawright shows, these tools are of little help, especially in macro-comparative analysis. He urges scaling down to more modest frameworks of comparison that potentially incorporate a substantial use of qualitative evidence.

Dunning points to the potentially large contribution of natural experiments—which, in his examples, focus entirely on much smaller-scale comparisons. At the same time, Dunning underscores severe trade-offs that may arise in employing these research designs, and both he and Seawright make

clear that perhaps too often, these methodological tools do not escape the confines of regression analysis to the degree that many methodologists hope they will.

Qualitative Methods

Another avenue is opened by further refinements in qualitative tools. A familiar, traditional option here is typically called the small-N comparative method, a strategy common in research that entails both cross-national comparisons and comparison of political units within nations—whether they be regions, provinces or states, or metropolitan areas. Here, the analyst juxtaposes two, or four, or perhaps six cases, with a central idea often being to set up matching and contrasting cases in a way that is seen as “controlling” for extraneous factors and allowing a focus on the principal variables of concern. This approach is often identified with J. S. Mill’s (1974 [1843]) methods of agreement and difference, and with Przeworski and Teune’s (1970) most similar and most different systems designs.

In our view, this small-N comparative approach is truly invaluable in concept formation and in formulating explanatory ideas (see chap. 1 and online chapters of this book). It is much weaker as a basis for causal inference. It involves, after all, what is in effect a correlation analysis with such a small N that it is not an appropriate basis for evaluating causal claims. The matching and contrasting of cases employed probably cannot succeed, by itself, in controlling for variables that the researcher considers extraneous to the analysis.

Rather, as is well known, the key step is to juxtapose this comparative framing with carefully-executed analysis carried out within the cases. The challenge, therefore, is to find strong tools of within-case analysis.

Correspondingly, the objective of chapters 10, 11, and 12, by Andrew Bennett, David A. Freedman, and Henry E. Brady, is to systematize and refine the tools of process tracing and causal-process observations. Through a new typology of process tracing, along with many examples, both macro and micro, we seek to place these procedures of qualitative analysis on a more secure foundation, thereby strengthening their value and legitimacy as procedures for causal inference. To reiterate, these chapters are accompanied by exercises posted with the online materials for this book.

In sum, our objective in the second edition is to sustain a clear-eyed awareness of limitations inherent in standard inferential tools; and to push forward in strengthening these tools, both quantitative and qualitative.

I

A DEBATE ON METHODOLOGY

A. FRAMING THE DEBATE

1

Refocusing the Discussion of Methodology

Henry E. Brady, David Collier, and Jason Seawright

MAINSTREAM QUANTITATIVE METHODS, QUALITATIVE METHODS, AND STATISTICAL THEORY

The quest for shared standards of methodology and research design is an abiding concern in the social sciences. A recurring tension in this quest is the relationship between quantitative and qualitative methods. This book aims to rethink the contribution of these alternative approaches and to consider how scholars can most effectively draw on their respective strengths.

One view of the relation between quantitative and qualitative methodology is provided by what we call “mainstream quantitative methods,” an approach based on the use of regression analysis and related techniques for causal inference. Scholars who champion this approach often invoke norms identified with these tools to argue for the superiority of quantitative research, sometimes suggesting that qualitative research could be greatly improved by following such norms more closely. These scholars in effect propose a quantitative template for qualitative research. In doing so, they have made some valuable suggestions that qualitative researchers would do well to consider.

Qualitative methodologists,¹ for their part, have raised legitimate con-

1. We understand qualitative methods as encompassing partially overlapping approaches such as the case-study method, small-N analysis, the comparative method, concept analysis, the comparative-historical method, the ethnographic tra-

cerns about the limitations of the quantitative template. Some qualitative analysts are dubious that the quantitative approach provides the only appropriate model for qualitative analysis. Others consider the quantitative template entirely inappropriate. Still others argue that the qualitative approach has strengths often lacking in quantitative studies and that quantitative analysts have much to learn from the qualitative tradition.

Yet another perspective on quantitative and qualitative methods is provided by ideas drawn from what we call “statistical theory.” In contrast to mainstream quantitative methods, these ideas reflect a long history of skepticism about applying the assumptions behind regression analysis and related tools to real-world data in the social sciences.² This methodological approach sometimes advocates alternative techniques that allow researchers to draw more limited inferences based on fewer untested assumptions. According to this perspective, it is by no means evident that conventional quantitative tools are more powerful than qualitative tools.

Indeed, it is possible to draw on statistical theory to provide what may be thought of as a “statistical rationale” for many standard practices of qualitative research. This does *not* involve an admonition that qualitative analysts, in designing research, are expected to prove theorems in order to demonstrate that they have adopted the right methods. Rather, this rationale provides other kinds of insight into the analytic contribution of qualitative methods. A basic theme of this volume is that many qualitative research practices can be justified both on their own terms, and on the basis of this statistical rationale.

Overall, a meaningful discussion of methodology must be grounded in the premise that strengths *and* weaknesses are to be found in both the qual-

dition of field research, interpretivism, and constructivism. For many purposes, the quantitative-qualitative distinction may be disaggregated. In chapter 9 and the glossary, we propose four component dimensions: level of measurement, number of cases, whether explicit statistical tests are employed, and what we call thick versus thin analysis. Yet the simple quantitative-qualitative dichotomy offers a heuristic distinction that productively structures much of the current discussion.

2. The tradition to which we refer grows out of debates among statisticians on causal inference in experiments and observational studies. It may be dated to Karl Pearson’s 1896 critique of G. Udny Yule’s causal assessment, based on a regression analysis of observational data, of the relation between welfare policy and poverty in Britain (Stigler 1986: 351–53, 358). For a recent statement about this debate, see Freedman (1999). In addition to work within the discipline of statistics, we consider this tradition to encompass studies in the fields of econometrics, psychometrics, and measurement theory that, like Pearson’s critique, explore the foundations of inference. We would also include methodological contributions by some scholars in political science and sociology whose work stands outside of the basic regression framework.

itative and quantitative approaches. Regarding the weaknesses, as Brady (chap. 3, this volume) puts it, qualitative researchers are perhaps “handicapped by a lack of quantification and small numbers of observations,” whereas quantitative researchers may sometimes suffer from “procrustean quantification and a jumble of dissimilar cases.” The most productive way to reconcile these two approaches is not through the unilateral imposition of norms, but rather through mutual learning.

THE DEBATE ON *DESIGNING SOCIAL INQUIRY*

In the present volume, we explore the relationship between quantitative and qualitative methodology through an extended discussion of a book that exemplifies the approach of mainstream quantitative methods: *Designing Social Inquiry: Scientific Inference in Qualitative Research* (hereafter KKV), by Gary King, Robert O. Keohane, and Sidney Verba.

KKV'S CONTRIBUTION

KKV has emerged as one of the most influential statements ever published on the relationship between quantitative and qualitative methods. The book is based on the tacit assumption that quantitative, large-N researchers have superior tools for solving many problems of methodology and research design, compared to their qualitative counterparts. Accordingly, KKV seeks to make such tools accessible to qualitative analysts, so as to help them design better research. While the premise is, in effect, the superiority of quantitative methods, the goal is to build bridges. The authors take seriously the idea that we should seek a common language for framing issues that arise in all forms of inquiry, and their effort to articulate the shared concerns of quantitative and qualitative research is a valuable contribution.

KKV's wide influence also stems from the systematization of quantitative methods that it offers. Although framed as an extended set of recommendations for qualitative researchers, the book is based on ideas drawn from the mainstream quantitative framework. In the course of summarizing these ideas, KKV offers numerous specific recommendations about different steps in the research process: for example, defining the research problem, specifying the theory, selecting cases and observations, testing descriptive and causal arguments, and subsequently retesting and refining the theory. In sum, KKV's reach is broad and its practical advice abundant.

At the most general level, by focusing scholarly attention on problems of research design, KKV aims to improve the practice of social science, understood as a collective effort to describe and explain political and social phe-

nomena. KKV characterizes this collective effort as being concerned with descriptive and causal inference, a term which may seem alien to some qualitative researchers. However, as Charles Ragin emphasizes (chap. 3 online), “there is no necessary wedge separating the goal of ‘inference’—the key concern of quantitative approaches—from the goal of making sense of cases—a common concern of qualitative approaches.” The term “inference” can thus be seen as one specific label for a shared objective that spans diverse traditions of research.

KKV has had as great an impact, in terms of encouraging analysts to think about research design, as any book in the history of political science. The book is widely read in other fields as well, and it has exercised a salutary influence on many different branches of qualitative research. Even qualitative analysts who strongly disagree with KKV have adopted terms and distinctions introduced in the book. In addition, the concern of qualitative analysts with defending their own approach vis-à-vis KKV has pushed these scholars toward a more complete systematization of qualitative methods. In this and other ways, KKV has been strikingly successful in achieving its basic goal of encouraging researchers to think more carefully about methodological issues.

Finally, the authors of KKV deserve praise for their willingness to participate in an ongoing dialogue that is helping to advance this methodological discussion. In their response (reprinted as chapter 7 below) to a 1995 symposium on their book in the *American Political Science Review*, they observe that, “although our book may be the latest word on research design in political science [as of its publication in 1994], it is surely not the last” (111 this volume).

WHERE DO WE GO FROM HERE?

The present volume extends this methodological debate. We take as a point of departure a number of basic concerns about KKV’s framework.

In our view, KKV gives insufficient recognition to well-known limitations of mainstream quantitative methods. The book does present a useful discussion of assumptions that underlie regression analysis. Yet KKV does not devote adequate attention to a key statistical idea: Regression analysis depends on the model, and if the model is wrong, so is the analysis. For this reason, estimating a regression model with empirical data does not fully test the model. Relatedly, KKV places strong emphasis on evaluating uncertainty. Yet the book fails to acknowledge that significance tests are designed to evaluate *specific* kinds of uncertainty, and that the common practice of employing them as a *general-purpose* tool for estimating uncertainty extends these tests beyond the uses for which they were intended.

Against this backdrop, KKV goes too far in advocating the perspective of mainstream quantitative methods as a foundation for research design and qualitative inquiry. We are convinced that this perspective provides an excessively narrow understanding of the research process. More specifically, along with being too confident about the strengths of quantitative tools, the book gives insufficient recognition to the contributions of qualitative tools. KKV overemphasizes the strategy of increasing the number of observations, and it overlooks the different kinds of observations and the different ways that data are used in quantitative and qualitative research. The book is inattentive to the risk that increasing the N may push scholars toward an untenable level of generality and a loss of contextual knowledge. It overstates its warning against post hoc hypothesis formation and standard practices of disciplined inductive research. Relatedly, it neglects the fact that econometric writing on “specification searches” has sought to systematize inductive procedures. Finally, KKV occasionally refers to trade-offs, yet the book does not acknowledge that they must be a basic concern in designing research.

We want to be clear about what these criticisms do and do not amount to. They do not amount to a rejection of the basic enterprise of striving for a shared vocabulary and framework for both quantitative and qualitative research. Indeed, we are strongly committed to the quest for a common framework. While we have great respect for scholars who explore epistemological issues, we worry that such concerns may sometimes unnecessarily lead researchers and students to take sides and to engage in polemics. Thus, we share KKV’s (4–5) view that quantitative and qualitative methods are founded on essentially similar epistemologies.

Correspondingly, the present volume is certainly not meant to widen the gap between the qualitative and quantitative approaches by identifying profound and obdurate differences. Indeed, we would argue that the differences are less deep-seated than is sometimes believed. To the extent that differences do exist, however, we take the normative position that a basic goal in work on methodology is to overcome these differences. We should seek a shared framework allowing researchers using diverse analytic techniques to develop evidence that is convincing to analysts of differing methodological persuasions. This larger body of mutually accepted evidence can, in turn, contribute to finding better answers to the substantive questions that drive social research.

TOOLS AND STANDARDS

As we suggest in the subtitle of this book, while analysts have diverse tools for designing, executing, and evaluating research, it is meaningful to seek

shared standards for employing such tools. These shared standards can facilitate recognition of common criteria for good research among scholars who use different tools. Methodological pluralism and analytic rigor can be combined.

By tools we mean the specific research procedures and practices employed by quantitative and qualitative researchers. Some tools are highly systematized and have elaborate technical underpinnings. Examples of such tools are regression analysis, structural equation modeling, factor analysis, tests of statistical significance, and probability theory. Increasing the number of observations is a research tool repeatedly advocated by KKV. Other tools include qualitative research practices such as within-case analysis, process tracing, procedures for avoiding conceptual stretching, qualitative validity assessment, and strategies for the comparison of matching and contrasting cases. Methods of data collection are also tools: for example, public opinion research, focus groups, participant observation, event scoring, archival research, content analysis, the construction of “unobtrusive measures,” and the systematic compilation of secondary sources. At various points in the text, we have introduced summary tables that provide an overview of the different tools being discussed, and many tools are also discussed in the glossary.

The chapters in the present volume devote considerable attention to various methodological tools that KKV undervalues or overlooks. The following paragraphs enumerate four broad methodological literatures with which many of these tools are identified. Some correspond to standard practices of qualitative researchers; others are derived from statistical theory.

1. *Logical and Statistical Foundations of Causal Inference*. A large body of research on the logical and statistical foundations of causal inference expresses considerable skepticism about causal inference based on observational data. This literature points to the need for more robust approaches than those advocated in mainstream quantitative methodology.
2. *Concepts*. Research on concepts, concept formation, and the evolution of concepts in the course of research makes it clear that sustained attention to conceptual issues is an indispensable component of research design. The insights of this literature suggest that the limited advice that KKV does give on working with concepts in fact points in the wrong direction.
3. *Measurement*. A major literature located in the fields of mathematical measurement theory and psychometrics provides researchers with systematic guidance for measurement. This literature emphasizes, for example, the contextual specificity of measurement claims, reinforcing the conviction of many political scientists that knowledge of con-

- text and care in bounding the generality of research findings must be a central concern in research design. Such guidance is lacking in KKV.
4. *Causal Inference in Case Studies*. A long tradition of writing has explored tools and strategies of causal inference in case studies: for example, process tracing and other forms of within-case analysis; the deliberate selection of “most-likely,” “least-likely,” and “deviant” cases; and, in the comparative case-study tradition, the methods of agreement and difference. KKV seeks to subsume these tools within its own framework, based on the norms of large-N quantitative analysis. The case-study literature in effect turns KKV’s argument on its head, suggesting that (a) the practice of causal inference in qualitative research is viable on its own terms, and (b) inference in quantitative research can sometimes be improved through the use of tools strongly identified with the qualitative tradition.

Through focusing on tools drawn from these diverse areas of methodology, as well as on more conventional quantitative tools, we seek to lay a stronger foundation for an integrated approach to the design and execution of research.

All research tools, both qualitative and quantitative, must be subject to critical evaluation. Correspondingly, scholars should seek shared standards for assessing and applying these tools. Relevant standards must include attention to basic trade-offs that arise in conducting research. Once we acknowledge that not all analytic goals can be achieved simultaneously—Przeworski and Teune’s trade-offs among accuracy, generality, parsimony, and causality are a famous example (1970: 20–23)—then it is easier to move toward a recognition that alternative methodological tools are relevant and appropriate, depending on the goals and context of the research.

Neither qualitative nor quantitative analysts have a ready-made formula for producing good research. We are convinced that the wide influence exercised by KKV derives in part from the book’s implicit claim that, if scholars follow the recommendations in the book, it is relatively straightforward to do good quantitative research; as well as the explicit argument that qualitative researchers, to the degree possible, should apply the quantitative template.³

3. KKV does briefly note the limitations of quantitative research. The book states that “[i]n both quantitative and qualitative research, we engage in the imperfect application of theoretical standards of inference to inherently imperfect research designs and empirical data” (7; see also 8–9). However, in the eyes of many critics, KKV does not follow through on these words of caution, instead going too far in extending the norms of quantitative analysis to qualitative research. Further, KKV’s statements on the pages just cited are closely linked to its arguments about estimating error, and the authors are far more confident than we are about the viability of error estimates in quantitative research, not to mention in qualitative research. See,

In fact, it is difficult to make causal inferences from observational data, especially when research focuses on complex political processes. Behind the apparent precision of quantitative findings lie many potential problems concerning equivalence of cases, conceptualization and measurement, assumptions about the data, and choices about model specification such as which variables to include. The interpretability of quantitative findings is strongly constrained by the skill with which these problems are addressed. Thus, both qualitative and quantitative research are hard to do well. It is by recognizing the challenges faced in both research traditions that these two approaches can learn from one another.

Scholars who make particular choices about trade-offs that arise in the design of research should recognize the contributions of those who opt for different choices. For example, let us suppose that a scholar has decided, after careful consideration, to focus on a small N to carry out a fine-grained, contextually sensitive analysis that will facilitate operationalizing a difficult concept. A large-N researcher should, in principle, be willing to recognize this choice as legitimate.

At the same time, the small-N researcher should recognize that the advantages of focusing on few cases must be weighed against the costs. These costs include, for example, forgoing large-N tools for measurement validation and losing the generality that might be achieved if a wider range of cases is considered. In short, researchers should recognize the potential strengths and weaknesses of alternative approaches, and they should be prepared to justify the choices they have made.

TOWARD AN ALTERNATIVE VIEW OF METHODOLOGY

Building on these themes, the present volume develops alternative arguments about the appropriate balance between the quantitative and qualitative traditions, and about research design and methodology more broadly.⁴ Here are some key steps in these arguments.

1. *In the social sciences, qualitative research is hard to do well. Quantitative research is also hard to do well. Each tradition can and should learn from the other.* One version of conventional wisdom holds that achieving

for example, Bartels's discussion of assessing measurement error (chap. 4, this volume), as well as the discussion in chapter 9 focused on the misuse of significance tests.

4. While issues of descriptive inference are a recurring theme in the following chapters (see, e.g. 34–37, 132–40 this volume), the focus here is primarily on causal inference.

analytic rigor is more difficult in qualitative than in quantitative research. Yet in quantitative research, making valid inferences about complex political processes on the basis of observational data is likewise extremely difficult. There are no quick and easy recipes for either qualitative or quantitative analysis. In the face of these shared challenges, the two traditions have developed distinctive and complementary tools.

- a. *A central reason why both qualitative and quantitative research are hard to do well is that any study based on observational (i.e., nonexperimental) data faces the fundamental inferential challenge of eliminating rival explanations.* Scholars must recognize the great divide between experiments and observational studies. Experiments eliminate rival explanations by randomly assigning the values of the explanatory variable to the units being analyzed. By contrast, in all observational studies, eliminating rival explanations is a daunting challenge. The key point, and a central concern of this book, is that quantitative and qualitative observational studies generally address this shared challenge in different ways.
2. *Mainstream quantitative methodologists sometimes advocate the quantitative approach as a general template for conducting research. By contrast, some statistical theorists question the general applicability of the conventional quantitative approach.* Strong advocacy of the quantitative template is found in many disciplinary subfields. Yet it is essential that political scientists—and scholars in other fields as well—take a broader view and reflect more deeply on the contributions and limitations of both qualitative and quantitative methods. A valuable component of this broader view draws on ideas from statistical theory.
 - a. *One recurring issue regarding the tradition of advocacy based on the quantitative template concerns how much scholars can in fact learn from findings based on regression analysis, as well as their capacity to estimate the degree of uncertainty associated with these findings.* For regression results to be meaningful, analysts must assume, as noted earlier in this chapter, that they have begun with the correct statistical model. Empirical data analysis may provide some insight into the plausibility of this assumption, yet such analysis does not fully test the assumption. Another key idea identified with the quantitative template concerns the capacity to estimate uncertainty. Unfortunately, in some areas of research, standard practice in the use of significance tests extends their application to evaluating forms of uncertainty that they were not designed to assess.
 - b. *Another issue regarding the quantitative template is the recurring recommendation that researchers can gain inferential leverage in addressing rival explanations by increasing the number of observations—in the con-*

ventional sense of increasing the N . Yet this advice is not always helpful, in part because it may push scholars to compare cases that are not analytically equivalent. Although adding new observations is frequently useful, adding observations from a different spatial or temporal context or at a different level of analysis can extend the research beyond the setting for which the investigator can make valid inferences. While some scholars might be concerned that this focus on context leads researchers toward a posture of excessive particularism, concern with context is in fact a prerequisite for achieving descriptive and causal inference that is valid and rigorous.

3. *In making choices about increasing leverage in causal inference, and to address the concerns just noted, scholars should recognize the contributions of different kinds of observations.* It is productive to distinguish between two quite distinct uses of the term “observation,” one drawn from the quantitative tradition, the other from the qualitative tradition. Examples of these two types are presented in the appendix (see also 184–96 this volume).
 - a. *Data-set observations.* These observations are collected as an array of scores on specific variables for a designated sample of cases, involving what is sometimes called a rectangular data set. Missing data are an obstacle to causal inference based on data-set observations; it is therefore valuable that the data set be complete. Data-set observations play a central role not only in quantitative research, but also in qualitative research that is based on cross-case analysis.
 - b. *Causal-process observations.* These observations about context, process, or mechanism provide an alternative source of insight into the relationships among the explanatory variables, and between these variables and the dependent variable. Causal-process observations are sometimes less complete than data-set observations, in the sense that they routinely do not constitute a full set of scores across a given set of variables and cases. The strength of causal-process observations lies not in breadth of coverage, but depth of insight. Even one causal-process observation may be valuable in making inferences. Such observations are routinely used in qualitative research based on within-case analysis, and they can also be an important tool in quantitative analysis.
 - c. *These two types of observations have contrasting implications for maintaining an appropriate scope of comparison.* A focus on increasing the number of data-set observations, either at the same level of analysis or in subunits at a lower level of analysis, can yield major analytic gains, but it can also push scholars toward shifts in the domain of analysis that may be counterproductive. By contrast,

the search for additional causal-process observations may occur within the original domain.

4. *Methodological discussions could benefit from stronger advocacy from the side of the qualitative template, and all researchers should consider carefully some long-standing methodological priorities that derive from the qualitative perspective.* The qualitative template can make important contributions to broader methodological agendas. For example:
 - a. *Knowledge of cases and context contributes to achieving valid inference.* To expand on the earlier argument (2b and 3c), analytic leverage can derive from a close knowledge of cases and context, which can directly contribute to more valid descriptive and causal inference. This knowledge sensitizes researchers to the impact of cultural, economic, and historical settings, and to the fact that subunits of a given case may be very different from the overall case. In other words, knowledge of context provides insight into potentially significant factors that are not among the variables being formally considered. In this sense, it helps us to know what is hidden behind the assumption “other things being equal,” which is in turn crucial for the causal homogeneity assumption that is a requisite for valid causal inference. As discussed in this volume, such contextual knowledge is also crucial for measurement validity. Leverage derived from detailed knowledge of cases and context is closely connected to the idea of causal-process observations just discussed. Such knowledge is invaluable in both quantitative and qualitative research.
 - b. *Inductive analysis can play a major role in achieving valid inference and generating new ideas. Induction is important in both qualitative and quantitative research.* Mainstream quantitative researchers are sometimes too quick in dismissing the contribution to scholarly knowledge of inductive analysis and of the retesting of hypotheses against the same set of cases, on occasion invoking the traditional mandate to avoid “post hoc” hypothesis reformulation and theory testing. Yet even in technically advanced forms of statistical estimation, quantitative researchers routinely test alternative specifications against a given set of data (i.e., specification searches) and on this basis seek to make complex judgments about which specification is best. This iterated refinement of models and hypotheses constitutes a point of similarity to the inductive practices that are perhaps more widely recognized in qualitative research. Inductive procedures play a role in both traditions, and developing norms that guide, systematize, and make explicit these procedures for causal inference should be a basic concern of methodology.
 - c. *These arguments add up to a view of methodology in which qualitative*

research has a major role. The norms and practices of qualitative research deserve, in their own terms, serious attention in broader discussions of methodology. Further, ideas drawn from qualitative methodology can improve quantitative practices by addressing weaknesses in the quantitative approach.

5. *The contribution of qualitative methods can be justified both from within the qualitative tradition itself, and from the perspective of statistical theory.* Greater attention to qualitative methods can be justified, first of all, by the lessons that qualitative analysts learn from their own research. Many qualitative practices can also be justified on the basis of arguments drawn from statistical theory. Among the goals of this volume are to develop what may be thought of as a statistical rationale for qualitative research and to explore specific ways in which statistical theory can improve both qualitative and quantitative analysis. This perspective is very different from that of much writing in the tradition of mainstream quantitative methods, which seeks to subordinate qualitative research to the quantitative template.
6. *If both qualitative and quantitative methods are to play important roles as sources of norms and practices for good research, scholars must face the challenge of adjudicating between potentially conflicting methodological norms.* Such adjudication requires recognition of a basic fact and a basic priority.
 - a. *Research design involves fundamental trade-offs.* Methodological advice needs to be framed in light of basic trade-offs among: (a) alternative goals of research, (b) the types of observations researchers utilize, and (c) the diverse tools they employ for descriptive and causal inference. A methodological framework that does not centrally consider trade-offs is incomplete.
 - b. *Scholars should develop shared standards.* A basic goal of methodology should be to establish shared standards for managing these trade-offs. Shared standards can become the basis for combining the strengths of qualitative and quantitative tools.

These arguments form the basis for the ideas presented throughout this volume. The remainder of this introduction provides an overview of the chapters that follow.

OVERVIEW OF THE CHAPTERS

Part I of this book seeks to advance this methodological debate by building on the discussion stimulated by King, Keohane, and Verba's *Designing Social Inquiry*. We bring together a number of previously published statements in

this discussion—some presented basically in their original form, others extensively revised⁵—along with two introductory chapters, two concluding chapters that draw together different strands in this debate, and an appendix. The glossary defines basic terms, with a core definition presented in the first paragraph of each entry; for certain terms, subsequent paragraphs elaborate on the definition. Part I is divided into four sections: an Introduction (chaps. 1–2), Critiques of the Quantitative Template (chaps. 3–5), Linking the Quantitative and Qualitative Traditions (chaps. 6–7), and Diverse Tools, Shared Standards (chaps. 8–9).

INTRODUCTION

Following the present introductory chapter, David Collier, Jason Seawright, and Gerardo L. Munck (chap. 2) provide a detailed summary of the methodological recommendations offered by KKV, thereby framing the discussion developed later in the book. Chapter 2 focuses on the definition of scientific research, the treatment of descriptive and causal inference, and the assumptions that underlie causal inference. The chapter then synthesizes KKV's recommendations by formulating a series of guidelines for the design and execution of research. Although KKV does not present most of its methodological advice in terms of explicit rules, much of its argument can productively be summarized in this manner. Chapter 2 concludes by offering an initial assessment of KKV's framework.

CRITIQUES OF THE QUANTITATIVE TEMPLATE

How useful is the quantitative template as a guide for qualitative research? This question is addressed in chapters 3–5. It merits emphasis that these chapters praise KKV for presenting mainstream ideas of quantitative inference in a minimally technical manner; for offering many useful didactic arguments about how qualitative analysts can improve their research by applying simple lessons from statistics and econometrics; and for making genuine contributions to the field of methodology. At the same time, however, these chapters reconsider and challenge some of KKV's basic arguments.

"Doing Good and Doing Better: How Far Does the Quantitative Template Get Us?" by Henry E. Brady (chap. 3) argues that KKV does not adequately

5. The relationship of each chapter to previously published material is explained in the acknowledgment of permission to reprint copyrighted material at the end of this volume.

consider the foundations of causal inference in quantitative research, and that the book does not properly attend to conceptualization and measurement. Regarding causal inference, Brady suggests that KKV pays insufficient attention to the challenges faced in research based on observational, as opposed to experimental, data. Specifically, the book fails to discuss how theory and preexisting knowledge can justify a key assumption that underlies causal assessment with observational data, that is, the assumption that conclusions are not distorted by missing variables. Concerning the second theme, Brady finds that KKV ignores major issues of concept formation and basic ideas from the literature on measurement. This latter body of work shows that quantitative measurement is ultimately based on qualitative comparisons, suggesting a very different relation between quantitative and qualitative work than is advocated by KKV.

"Some Unfulfilled Promises of Quantitative Imperialism" by Larry M. Bartels (chap. 4) suggests that KKV's recommendations for qualitative researchers exaggerate the degree to which quantitative methodology offers a coherent, unified approach to problems of scientific inference. KKV classifies research activities that do not fit within its framework as prescientific, leading the authors to a false separation between (a) producing unstructured knowledge and "understanding," and (b) making scientific inferences. Bartels is convinced that unstructured knowledge and understanding are a necessary part of inference. Likewise, in Bartels's view, KKV claims to have solutions to several methodological problems that neither its authors nor anyone else can currently solve. These include the challenge of estimating the uncertainty of conclusions in qualitative (and even quantitative) research; distinguishing between the contribution made by qualitative evidence and quantitative evidence in analyses that employ both; assessing the impact of measurement error in multivariate analysis; and multiplying observations without violating the causal homogeneity assumption. According to Bartels, the fact that leading practitioners in political science cannot adequately address these problems suggests that they may be the most important issues currently pending for further research on methodology.

"How Inference in the Social (but Not the Physical) Sciences Neglects Theoretical Anomaly" by Ronald Rogowski (chap. 5) argues that KKV underestimates the importance of theory in the practice of research. KKV's rules about case selection and the number of cases needed to support or challenge a theory reflect this inattention. In fact, following KKV's rules would lead scholars to reject as bad science some of the most influential works in the recent history of comparative politics. Single-case studies are particularly useful in challenging already-existing theories, if these theories are precisely formulated; yet KKV claims that a single case cannot discredit a scientific theory. Rogowski suggests that if the analyst employs theory that

is both powerful and precise, carefully constructed studies that examine anomalous cases can be invaluable, notwithstanding KKV's warnings about selection bias.

QUALITATIVE TOOLS

The basic analytic tools of quantitative researchers are reasonably well understood. By contrast, qualitative tools are less well codified and recognized. What are these tools? This question was addressed in Chapters 7 to 9 of the first edition (as well as in Chapter 6), and for the second edition these chapters are now available on the Rowman & Littlefield website (as discussed in the Preface).

LINKING THE QUANTITATIVE AND QUALITATIVE TRADITIONS

Given that the qualitative and quantitative traditions have distinctive strengths, how can they best be combined? The third section offers two perspectives on this challenge. "Bridging the Quantitative-Qualitative Divide" by Sidney Tarrow (chap. 6) offers valuable suggestions for linking quantitative and qualitative research. Qualitative analysis is better suited than quantitative research for process tracing, for exploring the tipping points that play a critical role in shaping long-term processes of change, and for providing more nuanced insight into findings derived from quantitative investigation. Quantitative analysis, in turn, can frame and generalize the findings of qualitative studies. In Tarrow's view, the most valuable interaction between the two research traditions occurs when scholars "triangulate" among alternative methods and data sources in addressing a given research problem.

"The Importance of Research Design" (chap. 7), reprinted here with the kind permission of Gary King, Robert O. Keohane, and Sidney Verba, is from the 1995 symposium on *Designing Social Inquiry*, published in the *American Political Science Review*. This chapter should be understood as the authors' interim response to the ongoing debate about linking the quantitative and qualitative traditions. Because it was written in 1995, it obviously does not take into account all the arguments in the present volume, though it does make reference to ideas presented here by Rogowski and Tarrow (and also Collier, Mahoney, and Seawright, from the online posting), as well as to arguments advanced in some other chapters.

King, Keohane, and Verba underscore central themes in KKV and clarify certain key ideas. The authors argue that the fundamental challenge for

both quantitative and qualitative analysis is good research design. King, Keohane, and Verba agree with Rogowski on the importance of theory, although they emphasize that telling people how to theorize is not their goal. Perhaps most significantly, they argue that “much of the best social science research can combine quantitative and qualitative data, precisely because there is no contradiction between the fundamental processes of inference involved in each” (chap. 7). All researchers, whether quantitative or qualitative, need to understand and utilize the same logic of inference.

King, Keohane, and Verba go on to explore and illustrate two related themes: the idea of science as a collective enterprise, which they discuss in relation to well-known books of Arend Lijphart and William Sheridan Allen; and problems of addressing selection bias, which they illustrate by reference to books by Peter Katzenstein and Robert Bates. Finally, the chapter proposes that Tarrow’s arguments about “triangular conclusions” provide a valuable unifying idea that brings together the diverse perspectives on methodology under discussion.

DIVERSE TOOLS, SHARED STANDARDS

The final part of the book synthesizes and extends the debate on quantitative and qualitative methods. We argue that, precisely because researchers have a diverse set of methodological tools at their disposal, it is essential to seek shared standards for the application of these tools.

“Critiques, Responses, and Trade-Offs: Drawing Together the Debate,” by David Collier, Henry E. Brady, and Jason Seawright (chap. 8), integrates and evaluates this methodological discussion. In a further effort to bridge the quantitative-qualitative divide, chapter 8 reviews the critiques of KKV offered in chapters 3–6 of the present volume and in the online chapters and formulates responses that draw on ideas derived from statistical theory. Two of the critiques concern the challenge of doing research that is important and the issue of probabilistic versus deterministic models of causation. For these topics, the statistical response calls for a synthesis that combines elements of KKV’s position and the critique. For other parts of the debate—on conceptualization and measurement, and on selection bias—statistical arguments emerge that more strongly reinforce the critique of KKV. The final part of this chapter explores the idea that trade-offs are inherent in research design and develops the argument that the search for shared standards necessarily poses the challenge of managing these trade-offs.

The final chapter of Part I offers some broader conclusions about tools for causal inference. “Sources of Leverage in Causal Inference: Toward an Alternative View of Methodology,” by David Collier, Henry E. Brady, and

Jason Seawright (chap. 9), focuses on the fundamental challenge of eliminating rival explanations and making good causal inferences. This chapter formulates several methodological distinctions that help bring into sharper focus the relationship between the quantitative and qualitative traditions and, more specifically, the contrasts in how they deal with causal inference. A further goal of this discussion is to explore the implications of the distinction between data-set observations and causal-process observations. The chapter argues that this distinction offers a more realistic picture of the contributions to causal inference of both quantitative and qualitative tools—and of how these differing contributions can be integrated.

Taken together, the arguments developed in this volume lead us to reflect on the expanding influence in social science of increasingly technical approaches to method and theory. We advocate an eclectic position in response to this trend. While it is essential to recognize the powerful contribution of statistically and mathematically complex forms of method and theory, simpler tools are sometimes more economical and elegant, and potentially more rigorous. Scholars should carefully evaluate the strengths and weaknesses of these diverse tools in light of existing knowledge about the topic under study, and with reference to broader shared standards for descriptive and causal inference and for refining theory. This eclectic approach is the most promising avenue for productive decisions about research design.

2

The Quest for Standards: King, Keohane, and Verba's *Designing Social Inquiry*

David Collier, Jason Seawright, and Gerardo L. Munck

Scholars turn to methodology for guidance in conducting research that is systematic, rigorous, and cumulative. *Designing Social Inquiry: Scientific Inference in Qualitative Research*, by Gary King, Robert O. Keohane, and Sidney Verba (hereafter KKV), has commanded wide attention because it forcefully and articulately provides such guidance. With clarity of exposition and many examples, the book presents an extended set of practical recommendations for the design and execution of research. In conjunction with KKV's goal of providing a new framework for qualitative research, the book offers an important synthesis of what we will call mainstream quantitative methods. KKV therefore constitutes a general statement about methodology, and this fact helps account for the wide attention it has deservedly received.

The present chapter provides an overview of KKV. We first introduce three fundamental ideas in KKV's view of methodology: (1) the criteria for scientific research; (2) the concept of inference—a term used in the title of the book and central to KKV's exposition; and (3) the assumptions that justify causal inference.

The second part of this chapter adopts a different approach to summarizing KKV's framework by presenting it in terms of a set of guidelines for conducting research. KKV does not explicitly synthesize its recommendations

as an over-arching set of rules,¹ yet we believe these guidelines provide a summary that plays a constructive role in focusing the discussion.

Finally, the conclusion to the chapter anticipates the debate in the remainder of the present volume, noting both points of convergence and areas of substantial divergence vis-à-vis the perspective presented by KKV (see table 2.2 toward the end of this chapter).

In this summary of KKV's arguments, we occasionally provide examples of our own. At certain points, as with the discussion of conditional independence, we offer a somewhat more elaborate presentation than KKV, given that these are topics to which we return later in the present volume. Nevertheless, the intent of the chapter, except for the conclusion, is to present KKV's framework.

SCIENTIFIC RESEARCH, INFERENCE, AND ASSUMPTIONS

Three central components of KKV are its treatment of scientific research, inference, and assumptions. In relation to prior discussions of these topics, KKV's goal is not primarily to present new ideas. However, as a set of recommendations designed specifically for qualitative researchers, KKV's treatment of these topics is innovative and deserves careful attention.

Scientific Research

KKV argues that social science ought to be good *science*. To that end, the book presents a careful definition of what makes research scientific. Some readers may find KKV's insistence on the idea of science jarring and this framing of goals too narrow. Yet these goals are in fact of broad relevance. How, then, does KKV define scientific research? First of all, such research always seeks to make *inferences*, "attempting to infer beyond the immediate data to something broader that is not directly observed" (8). The idea of inference is of such importance in KKV's methodological approach that it is explored in detail in the next section of this chapter.

Next, scientific research makes its procedures *public*. Researchers should report how they select cases, gather data, and perform analysis. This is nec-

1. Munck's (1998) review essay on KKV was the first effort to summarize the book in terms of a complete set of rules. Subsequently, Epstein and King (2002) adopted this approach in their long essay, "The Rules for Inference." The recommendations in their essay are quite similar to those in KKV, except that they give more attention to the tasks of defining the universe of cases and building a tradition of publicly available data sets.

essary if the scholarly community is to judge the quality of the research and the plausibility of its conclusions. If analysts do not report how they conduct their research, then “[w]e cannot evaluate the principles of selection that were used to record observations, the ways in which observations were processed, and the logic by which conclusions were drawn” (8).

Moreover, researchers must view their conclusions as inherently *uncertain*. “A researcher who fails to face the issue of uncertainty directly is either asserting that he or she knows everything perfectly or that he or she has no idea how certain or uncertain the results are” (KKV 9). Neither measurement nor theory in the social sciences is ever perfect and complete. According to KKV, scientific research requires scholars to acknowledge this fact and to estimate the degree of uncertainty in their inferences.

The final characteristic of scientific research is that findings are judged in light of the *method* employed, because, as KKV (9) argues, the content of science is the method. In other words, scientific findings should not be accepted or rejected according to the authority of the researcher, or in light of whether they correspond to the particular results preferred by a given investigator. Rather, the credibility of the methods employed should be a central criterion in evaluating research findings.

These criteria present a simple, reasonably straightforward basis for distinguishing scientific research from other kinds of intellectual pursuits.

Inference

The idea of inference is a major component of KKV’s methodological framework. Indeed, KKV views “inference”—in the sense of drawing larger conclusions on the basis of specific observations—as a foundation of social science. The book treats inference in broad terms, stating that “[i]nference, whether descriptive or causal, quantitative or qualitative, is the ultimate goal of all good social science” (34). KKV develops this idea in extended discussions of descriptive inference (chap. 2) and causal inference (chaps. 3–6).²

2. The relation between description and explanation is complex, as is clear in the discussion below of the contrast between the systematic and random components of phenomena. Even so, description versus explanation remains a fundamental heuristic distinction, both in KKV and in the present volume. At the simplest level, description addresses the question of “what?” and explanation addresses the question of “why?” Also, as noted in chapter 1 above (15–16 this volume), although the ideas of descriptive and causal “inference” may seem nonstandard to some readers, they can be viewed as convenient labels for the ubiquitous research task of moving from specific observations to more general ideas.

Descriptive Inference

In KKV's view, descriptive inference entails three tasks. **First**, it encompasses the idea of generalizing from a sample to a universe of cases, as routinely occurs in public opinion research. The researcher establishes the universe and the sample, analyzes the cases included in the sample, and makes inferences about the universe on the basis of the sample (e.g., KKV 70–71).

Second, descriptive inference encompasses inferences from observations to concepts. Analysts are rarely interested in reporting raw facts. Rather, they seek to describe political institutions, social structures, ideologies, and other complex phenomena. As conceptualized by social scientists, these phenomena are never directly observable: no one has ever *seen* an entire "social structure." Scholars observe certain facts, often at only one point in time, that are relevant to the complex idea of a social structure, that presumably persists over time. They must therefore make inferences from these particular facts to the broader idea of a social structure. Hence, "[d]escriptive inference is the process of understanding an unobserved phenomenon on the basis of a set of observations" (KKV 55).

A **third** aspect of descriptive inference, which is strongly emphasized by KKV, is the more complex issue of separating the "systematic" and the "random" components of any phenomenon. KKV (43) argues that descriptive inference inherently involves simplification, and one productive form of simplification can be to focus description on the systematic component of the phenomenon that the researcher seeks to explain.

Although in practice the separation of the systematic and random components may be difficult to achieve, it is important to see why this can be a useful idea. The rationale for this distinction depends on making a link between descriptive inference and causal inference. The systematic component of a phenomenon is understood as that which is explained by an accepted causal model; the random component is that which is not (60, 63).³

KKV points to alternative views of this random component. In one view, the world is inherently probabilistic. Thus, "[r]andom variation exists in nature and [in] the social and political worlds and can never be eliminated" (59). Another view rejects the idea that the world is inherently probabilistic, contending instead that what appears to be random "is only that portion of the world for which we have no explanation" (59). In other words,

3. KKV presents this idea by taking as a point of departure the supposition that the researcher lacks any prior knowledge of causal patterns: "[W]e begin any analysis with all observations being the result of 'nonsystematic' forces. Our job is then to provide evidence that particular events or processes are the result of systematic forces" (60).

causation is deterministic, and what appears to be random is simply the facet of reality that is explained by variables not yet included in the relevant model, or is due to measurement error.

KKV illustrates this distinction with the example of fluctuations in the vote for a given party within a particular electoral district (55). The vote for this party may vary over time in part due to factors that are truly random. Alternatively, it might vary due to specific events that are outside the conventional explanatory concerns of political scientists—for example, variations in the weather, or some accidental occurrence such as the use of ballots that voters find confusing. In either case, an analyst may wish to generate a description of the party's vote share from which these fluctuations are removed. A common way of accomplishing this is to take an average of the party's vote share across several elections, on the assumption that the random fluctuations will cancel one another out (58).

Of course, variation that falls outside the focus of one explanatory framework or theory may be a central concern for another theory. Correspondingly, a description based on a careful separation of systematic and random components that is well suited to one theory may be less appropriate to another theory. Notwithstanding this limitation, the possibility of such separation raises the important idea that analytically productive description may isolate that part of a phenomenon that we really seek to explain. More broadly, it serves as a useful reminder to researchers that the facts do not "speak for themselves." Rather, they are interpreted from some theoretical perspective.

KKV considers description a fundamental part of the social scientific enterprise, and the book warns that in research contexts where causal inference is unusually difficult, analysts should sometimes be satisfied with careful descriptive inference (44–45; also 34, 75 n. 1). Nonetheless, KKV pays greater attention to causal inference, arguing that the best description is organized as a collection of evidence that evaluates a causal claim (46–49). It is therefore hardly surprising that the larger part of KKV's focus is on research designed to test causal hypotheses.

Causal Inference

KKV's treatment of causation follows in the tradition of Neyman (1990 [1923]), Hodges and Lehmann (1964), Rubin (1974, 1978), and Holland (1986), who developed a counterfactual understanding of causation.⁴ According to this account, the idea that "X causes Y" in any given unit of analysis raises the hypothetical question of how the outcome on Y would have differed if X had not occurred in that unit. Given that it is impossible

4. This approach is reviewed in more detail on 44–49 below, in the discussion of conditional independence.

to observe both the occurrence and nonoccurrence of X for any given unit at one point in time, causal inference involves comparing something that did occur with something that did not occur. This is the source of what Holland and KKV (79, 82) call the “fundamental problem of causal inference,” that is, the problem that causal inference implicitly depends on a comparison with something that did not occur.

Using this counterfactual view of causation, KKV (76–82) hypothetically posits the existence of two parallel universes, exactly alike in every way except for one. Taking the example of a dichotomous independent variable, we might find that in one of these two universes, the unit being studied has a positive score on the hypothesized cause and thus receives the “treatment.” In the other universe, the hypothesized cause does not occur in the unit being studied: it is a “control.” The causal effect of the explanatory variable is the difference in the outcome between the two parallel universes.

This definition helps researchers in reasoning about causation as an abstract concept. It serves to clarify why scholars do indeed face a fundamental problem of causal inference: out of the two observations of a given case needed to directly assess a causal effect, researchers can, in the real world, only make one. Either a case gets the treatment, or it does not. In observational studies, analysts cannot even choose which of these two universes to observe, because they cannot manipulate the independent variable. Some kind of inference is necessary to overcome this fundamental problem; hence, causal inference is the only way to appraise causation. When this understanding of causation is applied in observational studies, analysts seek to approximate these hypothetical comparisons through real-world comparisons among observed cases. A central component of KKV’s advice focuses on how to carry out these real-world comparisons.

Making Inferences: Quantitative Tools and Analytic Goals

KKV’s recommendations can usefully be summarized in terms of the tools the book proposes, and in light of the goals it seeks to pursue with these tools. KKV draws heavily on regression analysis, econometrics, and other standard techniques of quantitative methodology (table 2.1). These include basic methods for describing quantitative data, such as means and variances, and, very crucially, the use of regression analysis for causal assessment. Regression analysis in the social sciences relies on quantitative tools of parameter estimation (i.e., estimating the coefficients associated with each independent variable), and generally also on significance tests (which address uncertainty due to sampling error or other forms of randomness in the model). In discussing causal inference from a regression perspective, KKV implicitly draws on these statistical techniques. Increasing the number of observations is frequently recommended as a basic tool for

Table 2.1. Quantitative Tools Employed in *Designing Social Inquiry*

<i>Tools</i>	<i>Comments</i>
Means and Variances	Means and variances are the basis for other tools discussed below.
Regression Analysis	Regression analysis is KKV's basic tool for causal inference from empirical data (e.g., 95–97, 121–22, 130–32, 168–72). Parameter estimation and significance tests, as used in regression analysis, provide a major part of the statistical basis for KKV's discussion of causal inference.
Increasing the N	KKV repeatedly advocates increasing the number of observations as the best way to enhance the inferential leverage of empirical tests (e.g., 19, 23–24, 29–31, 46–49, 52, 67, 99, 117–18, 120–21, 123, chap. 6).
Probability Theory	Many of KKV's "Formal Analysis" text boxes (e.g., 97–99, 166–68, 184–85) evaluate the variance and bias of different estimators by applying tools of probability theory.

enhancing inferential leverage in empirical tests (i.e., achieving higher levels of statistical significance). Finally, KKV employs tools of probability theory, such as expected value and variance of the estimator. KKV's tools are designed for use with quantitative data, and the book's fundamental advice to qualitative analysts is to use procedures in their own research that make a parallel contribution to valid inference. Although the chapters below debate whether it is in fact possible to implement this recommendation, there is not the slightest question that this advice has extended the analytic horizon of qualitative researchers.

With regard to KKV's broader analytic agenda, within the framework of what we will call the book's "overarching goals" of achieving valid descriptive and causal inference, a central focus is on "intermediate goals," which provide a justification for the use of these quantitative tools in pursuit of the overarching goals. Two major intermediate goals are avoiding bias and minimizing the variance of estimators in order to achieve higher levels of statistical significance.⁵ Analysts should seek to avoid bias, potential sources of which include systematic measurement error (155–57), selection procedures that are correlated with the dependent variable—including proce-

5. KKV uses the term "efficiency" to refer to the goal of minimizing estimator variance. However, the technical definition of efficiency in statistics is somewhat different, so we have used this more general phrase in the text. KKV does not explicitly defend its preference for lower-variance estimators in terms of statistical significance, but this is the most obvious interpretation.

dures that may cause selection bias (128–37), missing explanatory variables (168–76), and endogeneity, that is, the problem that the outcome variable or the error term influences the explanatory variables (185–96). Researchers should also minimize the variance of their estimators by excluding irrelevant explanatory variables (182–85) and by reducing non-systematic measurement error (157–68). In addition to reducing variance, which maximizes the precision of the inferences that can be drawn from a given data set, KKV recommends increasing leverage by creating data sets that have greater inferential power. Additional intermediate goals are summarized in the guidelines below. KKV thus builds on the tools of mainstream quantitative methods to propose a series of procedures for achieving valid inference in qualitative research.

KKV does not simply present these tools and goals in a mechanical fashion, but at various points considers how some of them intersect with concerns that derive from the qualitative tradition. For example, although researchers can avoid some types of selection bias through random sampling, the book recognizes that in small-N research, random sampling may create as many problems as it solves (124–28). Within the framework of nonrandom sampling, KKV is careful to avoid a piece of clichéd advice that is often invoked in discussions of selection bias—that is, “do *not* select on the dependent variable.” Instead, KKV argues that scholars who, for good reason, avoid random sampling and *do* select on the dependent variable should choose cases to reflect the full range of variation on that variable (141).⁶

Assumptions

KKV discusses the assumptions routinely employed to justify causal inference. Some scholars may think of these as “quantitative” or “statistical” assumptions. However, KKV (93) argues that these assumptions should not be understood narrowly as relevant only for quantitative analysis. Rather, assumptions are important for any study, whether quantitative or qualitative, that seeks to make the kind of inferences discussed in the previous section.

KKV urges researchers to “make the substantive implications of [their assumptions] extremely clear and visible to readers” (91). This advice is valuable because inferences depend on the assumptions that produce them, and a somewhat different set of assumptions can generate radically divergent inferences. This is one of the reasons why—as noted in chapter 1 above—it is hard to do really good quantitative research, just as it is hard

6. This corresponds to the second meaning of “selecting on the dependent variable” discussed in the glossary.

to do really good qualitative research. KKV consequently advises researchers to justify their assumptions with theory and empirical evidence to the greatest extent possible (91). Yet KKV recognizes that it is often difficult to establish such justifications (93, 95).

Causal homogeneity, independence of observations,⁷ and conditional independence are three major assumptions that KKV's authors view as essential for causal inference.⁸ These assumptions focus researchers' attention on three interrelated tasks: analyzing an appropriate set of cases; considering how cases and observations can influence each other in a way that may affect causal inference; and selecting variables appropriately and modeling the relations among them.

Causal Homogeneity

The assumption of causal homogeneity⁹ states that "all units with the same value of the explanatory variables have the same expected value of the dependent variable" (KKV 91). In other words, the outcomes for all the cases in the analysis must be produced by one causal model; after controlling for the values of the included independent variables, every case must have the same expected value on the dependent variable.¹⁰

Discussions of causal homogeneity are motivated by the concern that a given form of a causal model may only be appropriate to a particular domain of cases. If the model is extended to further cases, the researcher may have to make it more complex to accommodate distinctive causal features of those cases. Hence, this assumption is concerned with the relation between our causal ideas and the cases on which we focus.

In the statistical literature on causation (e.g., Rubin 1974; Holland 1986), a stronger version of the causal homogeneity assumption is presented, which Rubin and Holland call "unit homogeneity." According to this version of the assumption, different units are presumed to be *fully iden-*

7. This assumption is not treated in the same pages as the other two (KKV 91–97), yet it is likewise important (222–23).

8. We would add that somewhat modified versions of these assumptions do also permit causal inference. For example, independence of observations can be weakened, as in time-series analysis, where autocorrelation often arises. However, even the modified assumptions must, in fact, have the same basic properties as the assumptions discussed here.

9. KKV refers to this assumption as "unit homogeneity," as we explain below.

10. Two points should be made here. First, the "expected value" refers not to the value that one should anticipate for every case being analyzed, but rather to the average value across many hypothetical replications of each case. Second, KKV notes that one way to meet the causal homogeneity assumption is through the related assumption of "constant causal effects" (92–93).

tical to each other in all relevant respects except for the values of the main independent variable. This strong version is sufficient to allow causal inference without the assumption of conditional independence discussed below, but it is extremely unlikely that this strong homogeneity assumption will ever hold in the social sciences.

However, the weaker version of causal homogeneity that we discuss in this section, which allows units to differ from each other but requires that the causal parameters in the analyst's model be constant across all units, is more plausible and plays an important role in causal inference.

Though KKV occasionally makes reference to the stronger version of this assumption,¹¹ much of its discussion invokes the weaker version.¹² KKV refers to both versions of this assumption as "unit homogeneity." However, in labeling the weaker version of the assumption, which is much more central to KKV's overall framework, we find the term "causal homogeneity" more useful, both because it distinguishes this concept from the more rigorous standard of unit homogeneity and because it calls more explicit attention to the need for all cases to share the same causal model.

Specifically, if the causal homogeneity assumption is not met, and a researcher analyzes the data as if it were, the inference will be a misleading average that lumps together differences among subgroups of cases. This average may not adequately represent the pattern of causation in any given case. For example, it has been argued that among advanced industrial countries, in some national contexts the more highly paid workers are more class conscious, whereas in other national contexts they are less class conscious

11. KKV (91) defines unit homogeneity as being met if "the expected values of the dependent variables from each unit are the same when our explanatory variable takes on a particular value" (*italics omitted*). In this quote, the reference to multiple dependent variables for each unit invokes the Rubin-Holland framework for causality, and this clearly should be read as a reference to the strong version of unit homogeneity.

12. KKV (91) alternatively defines unit homogeneity as "the assumption that all units with the same value of the explanatory variables have the same expected value of the dependent variable." This statement, which refers only to the observed value of the dependent variable for each unit, does not invoke more complex statistical ideas of causation. Therefore, it would seem that it should be read as referring to the weaker version of unit homogeneity, involving constancy of causal parameters. This weaker version is also more compatible with KKV's (93) claim that "[t]he notion of unit homogeneity . . . lies at the base of all scientific research." In the Rubin-Holland framework, much scientific research specifically does not employ the unit homogeneity assumption, turning instead to alternatives such as randomization, conditional independence, and "ignorable treatment assignment." Hence, KKV's statement should be read as referring to the weaker assumption, and we therefore use the label "causal homogeneity" in discussing their arguments.

(Przeworski and Teune 1970: 26). If researchers simply average these two findings, they may find no relationship, resulting in a misleading conclusion. The appropriate solution would be to analyze the two groups of countries separately. Researchers would thus address causal heterogeneity by recognizing that causal processes are different between the two groups of countries, and by assuming that they are similar within each group. In regression analysis, this can sometimes be accomplished by introducing an interaction term that includes a dummy variable. In qualitative comparison, separate comparisons can be employed for the two groups. The fact that causal heterogeneity can thus be overcome by using a more complex model underscores a key point: causal homogeneity is not simply a property of the data, but of the data in relation to a particular causal model.

Independence of Observations

Another assumption concerns the independence of observations, that is, the idea that for each observation, the value of a particular variable is not influenced by its value in other observations and therefore provides new information about the phenomenon in question (222–23).¹³ If independence of observations is not met, this does not necessarily bias the causal inference. However, it does reduce the amount of new evidence gained from each additional observation, thereby increasing the variance associated with an inference.

For some readers, a familiar alternative label for this assumption, which is appropriate for discussing cross-sectional analysis, is “independence of cases.” However, this same assumption plays a major role in time-series analysis, in which the researcher analyzes multiple observations over time for each “case.” Hence, the broader idea of independence of multiple observations for the same case becomes a central issue, and it is therefore useful to employ this more general label.

An example of this problem in time-series analysis is found in the literature on advanced industrial countries that explores the impact of corporatism and partisan control of government on economic growth. Scholars who had been working with an *N* of twelve to fifteen countries sought to achieve a major increase in the *N* by combining cross-sectional and time-series analysis, focusing on the period 1967–1984 (Alvarez, Garrett, and Lange 1991). However, subsequent research argued that prior results had been based on an incorrect assumption about the independence of observations. Consequently, the estimates of standard errors were too low, yielding excessive confidence in the conclusions. Revised estimates, based on a recognition of interdependence among observations—both among coun-

13. Unlike the other two assumptions discussed in this chapter, the assumption of independence of observations is also important for descriptive inference.

tries and within countries over time—supported some of the findings of the 1991 study, but cast doubt on others (Beck et al. 1993; Beck and Katz 1995; Kittel 1999).

Of course, the nonindependence of observations can also be viewed *not* as a methodological problem, but as a substantive topic—that is, as causation that occurs through processes of diffusion. However, within the framework of most work in regression analysis, it is indeed a methodological problem.

Conditional Independence

KKV's final major prerequisite for causal inference with observational data is the assumption of conditional independence, or, to give it a more complete name, conditional independence of assignment and outcome. We present this assumption by first returning to the counterfactual definition of causation noted above, from which the idea of conditional independence emerges, and then by offering two examples to make clear the importance of this assumption. Our presentation here will be more detailed than for the other two assumptions, given that this third assumption is particularly important to the discussion later in the present volume (172–177).

According to the counterfactual understanding of causation, causal inference consists of comparing (a) the value of the outcome variable (Y_t with “*t*” for treatment) for a particular case when that case is exposed to a treatment, with (b) the value of the outcome variable (Y_c with “*c*” for control)¹⁴ for the same case when that case is not exposed to the treatment. Y_t and Y_c are thus two different variables that reflect the outcomes a case will experience on the dependent variable, according to whether the independent variable, conceptualized as an experimental treatment, is present or absent.¹⁵

The causal effect of the treatment for a given case is the difference between the two variables for the case: $Y_t - Y_c$. However, to restate the fundamental problem of causal inference discussed above, it is impossible to simultaneously observe Y_t and Y_c for any particular case. The value of one variable may be observed, but the value of the other is necessarily hypothetical. Consequently, it is impossible to compute $Y_t - Y_c$. Hence, in practice, causal inference seeks to replicate this hypothetical comparison by

14. We follow here the Rubin-Holland notation of “*t*” and “*c*,” which is also employed in chapter 13 below. In chapter 3 below, where Brady presents his direct commentary on KKV, he follows the book's notation, which is based on KKV's running example: “*i*” for “incumbent” and “*n*” for “nonincumbent.”

15. In this discussion, the independent variable may be dichotomous; alternatively, the treatment and control may reflect two different values on a continuous variable.

making real-world comparisons across (hopefully) similar units, some of which are exposed to the treatment and some of which are not.

When a real-world comparison is employed, the quality of the resulting causal inference depends on how cases are “assigned” to the treatment group and to the control group. Two issues are important here. First, a question of terminology: In observational studies, researchers do not actually assign cases to treatment and control groups. However, what we refer to as assignment does take place; it is carried out by social and political processes over which the researcher usually has no control.

The second issue, which is vital to the quality of causal inference, concerns the relationship between the assignment process and the outcome variables, Y_t and Y_c . The key question here is whether the cases are assigned in such a way that those in the treatment category have the same average values on both Y_t and Y_c as the cases in the control category. In other words, is the average of Y_t across the cases exposed to the treatment equal to the average of Y_t across the cases in the control group? Is this also true for Y_c ?

If the answers to these questions are “yes,” then the standard of independence has been met,¹⁶ and the researcher will be able to make a good inference about the causal effect of the treatment by comparing the observed Y_t among the cases given the treatment with the observed Y_c among the cases assigned to the control. The underlying logic here is that, if independence of assignment holds, any difference between the treatment group and the control group must be due to the treatment—because all other relevant factors are balanced between the two groups. If, on the other hand, cases are assigned in such a way that those in the treatment group tend to have a different Y_t or Y_c than the cases in the control group, then causal inference will be biased. For example, if cases with a high value of Y_t are more likely to enter the treatment group than cases with a lower value of Y_t , the researcher will probably overestimate the causal effect of the treatment.

Independence of assignment is a strong condition, and it is rarely plausible in an observational study. Observational studies often employ an assumption of *conditional* independence, which serves to justify causal inference even though the treatment and control groups initially do not

16. To be more precise, what is discussed here as independence is *mean* independence. Likewise, conditional independence as discussed here is actually *mean* conditional independence. For a discussion of these distinctions, see Stone (1993). Finally, the text above neglects two important, although somewhat narrow, technical issues: (a) whether there is a broader population from which the cases under investigation are a sample; and (b) whether the *expected* means of Y_t and Y_c , rather than the observed means, are in fact equal. The equality of the expected means is actually the key condition for mean independence and, if control variables have been introduced, for mean conditional independence.

have the same hypothetical average values on Y_t and Y_c . Suppose a variable (which we shall call Z) identifies subgroups of cases, within which independence of assignment does hold, and among which it does not hold. Then controlling for Z by comparing Y_t and Y_c within subgroups allows researchers to make unbiased inferences from the observational data. By stratifying in this manner, the standard of conditional independence is met.¹⁷ In fact, because Y_t and Y_c cannot both be directly observed, the researcher never knows with certainty that their average values are equal. But in principle, the introduction of the appropriate control can make them equal, and hence yield conditional independence. In practice, achieving appropriate statistical control may involve more than one control variable (Z_i to Z_n), and multivariate techniques are needed to introduce these multiple controls. For convenience, we will use the label Z to refer to one or more controls.

Given the importance of introducing control variables, the two words in this label, “conditional independence,” thus bring together two essential ideas. (a) It is best for inference that assignment to the treatment and control groups be independent of the two outcome variables Y_t and Y_c . Correspondingly, the full name of the assumption is “conditional independence of assignment and outcome.” (b) When independence does not hold, researchers can, in principle if not in practice, make inferences *as if* assignment were independent of Y_t and Y_c by statistically controlling for, or “conditioning” on, Z .

Conditional independence can be established if the appropriate statistical controls are introduced, removing the effect of an assignment process that does not meet the standard of independence. The assumption of conditional independence is thus addressed by employing with observational data the procedure of *statistical* control, as a substitute for the *experimental* control that is achieved through random assignment.

The effort by scholars to satisfy conditional independence by introducing the appropriate control can be illustrated with a well-known example of spurious correlation. In the United States, political participation is lower for African Americans and Latinos than for whites. In other words, if we hypothetically think of “nonwhite” as the treatment condition, and “white” as the control condition, individuals “assigned” to be African American and Latino have an average rate of participation, or average Y_t , that is lower than

17. Regression analysis depends on related assumptions about causation, such as the specification assumption discussed in chapter 9. For most purposes, these assumptions may be seen as similar, in that they both focus attention on the potential problem of missing variable bias. However, it is important to remember that alternative analytic tools (e.g., regression versus stratification) depend on assumptions that sometimes differ in important ways.

the average rate of participation, or average Y_c , among people “assigned” to be white. The lower participation rate of the first two groups provides an appropriate basis for descriptive inference (i.e., describing their levels of participation), but it is problematic as a basis for causal inference. It does not necessarily follow that being African American or Latino causes citizens to participate less. Rather, membership in these two groups is correlated with other factors, such as education and income, that could explain lower participation rates. These other factors serve the role of identifying salient subgroups among the cases; hence these other factors may be equivalent to the variable Z in the discussion above. When these other factors are controlled for, thus making it more plausible that conditional independence is satisfied, “neither being African American nor being Latino has a direct impact” on participation (Verba, Schlozman, and Brady 1995: 442).

In other words, after conditioning on—that is, controlling for— Z , these authors conclude that the average value of Y_i is in fact about the same as the average value of Y_c . It is not being African American or Latino that reduces the political activity of individuals within these groups. That apparent causal relation is spurious, and other factors such as low education or low income account for the lower rate of participation. Once the effect of these other factors is removed statistically, the underlying causal relationship emerges.

A second example illustrates the point that the conditional independence assumption is hard to meet when analysts cannot identify, or cannot measure, the variable or set of variables that must be controlled for. Consider the question of whether the *size* of revolutionary movements (independent variable) affects their *success* in overthrowing an existing regime (dependent variable). As Goldstone (1991: 137) emphasizes, because the personal cost of participating in an unsuccessful revolutionary movement can be high, many individuals will only join revolutionary movements that are seen as having at least some probability of defeating the regime. This evaluation obviously depends on the perceived strength of both the revolutionary movement and the regime. Specifically, the probability that a revolutionary movement will grow in size (which corresponds to the treatment) depends in part on the particular characteristics of the national regime that individuals evaluate in judging the relative strength of that regime. Yet the strength of the regime *also* plays a key, direct role in influencing the likelihood that the regime will fall, which is the outcome being explained.

Thus, due to these regime characteristics, those countries most susceptible to revolution may be most likely to face large revolutionary movements, and are in effect assigned to the treatment group. In this discussion, characteristics of the national regime are an instance of the variable Z above. Contrasts in these characteristics group together regimes that differ in the degree to which they are perceived as weak. Perceptions of weakness

are, in turn, correlated: (a) with the likelihood of regime collapse, given a strong insurgent movement, or Y_i ; and (b) with potential insurgents' decisions to rebel, which, when aggregated, constitutes the treatment. Unless these regime characteristics are included in the analysis and controlled for, researchers will overestimate the importance of popular participation in revolutionary opposition movements for causing regime collapse—given that greater popular participation is more likely when the chance of regime collapse is high.¹⁸

To meet the assumption of conditional independence, the researcher would need to collect data on these characteristics that adequately capture their role in influencing both the size of revolutionary movements and the likelihood of regime collapse. Yet collecting these variables and adequately controlling for them is doubtless more difficult than it is for the education variable in the prior example. The researcher would have to collect enough information about regime characteristics to arrive at the same evaluations and judgments that potential revolutionaries make about the strength of the regime. Hence, the idea of conditional independence is crucial here, but it is difficult to meet this assumption.

Overall, the idea of conditional independence uses the counterfactual definition of causation to provide a logical framework for reasoning about the critical task of controlling for rival explanations in causal inference.

To summarize the discussion of these three assumptions, KKV's goal is to underscore the idea that they are important to all researchers, and not just quantitative analysts. In all observational studies, causal inference never relies exclusively on the actual data, but also on assumptions about the political and social processes we are studying. It is evident that not only

18. This problem can arise regardless of whether the researcher takes a more structural or a more actor-centered view of revolution. One interpretation of this causal pattern could be that the perception of these revolutionary actors is an intervening variable that links these regime characteristics to the revolutionary outcome, involving an actor-centered and potentially "agental" explanatory perspective. Another interpretation views regime characteristics as direct, structural causes of revolution. For example, according to Chehabi and Linz (1998), under sultanistic regimes a poorly institutionalized, personalistic military is a critical structural factor in regime breakdown. Although the perception of the military on the part of revolutionary and regime actors may have some importance, this weakness of the military is seen, in its own right, as a critical causal factor. The point here is not to adjudicate between a structural and an actor-centered perspective, but rather to show that, from either perspective, failure to satisfy conditional independence may interfere with causal inference. Whether the structural weakness in the military causes revolution directly, or primarily through the perceptions of state and popular actors, varying degrees of regime strength can still confound our attempts to estimate the impact of popular participation on revolution.

KKV's discussion of these assumptions, but also the book's treatment of inference and the definition of scientific research, involve a perspective that is far more familiar to quantitative than to qualitative researchers. However, KKV is strongly committed to the idea that these issues are of equal relevance to both traditions. Even a scholar who disagrees with KKV must recognize that the book makes a fundamental contribution by pushing a broader range of researchers to grapple with these questions.

GUIDELINES: SUMMARIZING KKV'S FRAMEWORK

This section adopts a different approach to synthesizing KKV by presenting many of the book's more specific methodological recommendations as a set of guidelines. These guidelines are largely concerned with what we refer to in chapter 1 as intermediate goals, focusing on procedures for linking specific quantitative tools to the overarching goals of valid descriptive and causal inference. The guidelines help to make clear how KKV's broad ideas, summarized in the present chapter, inform the book's treatment of specific decisions about research design.

We organize the guidelines in terms of a research cycle (figure 2.1): defining the problem, specifying the theory, selecting cases and observations, carrying out descriptive and causal inference, and retesting and reformulating the theory. The final step completes this cycle by bringing the researcher back to the step of theory specification, and potentially also to redefining the research problem (see dashed arrow in the figure). Although research routinely moves through a series of ordered steps such as this, what is learned at each step certainly may lead to revisiting prior steps or jumping forward to subsequent steps. Hence, one could in fact place many more arrows in the diagram.

These guidelines are, of course, our summary of KKV's arguments. KKV makes periodic reference to "rules" for research (e.g., 6–7, 9), and the book presents five specific rules for constructing causal theories (99–114). However, the book does not synthesize its recommendations in terms of an overall set of rules or guidelines.¹⁹ Each of the guidelines presented below is introduced as a brief, self-explanatory phrase. For some of the guidelines, we spell out the idea in greater detail, often drawing on quotations from KKV. In all cases, specific page references are provided.

KKV states that "[a]ny meaningful rules admit of exceptions. . . . We seek not dogma, but disciplined thought" (7). Correspondingly, we do not want to give the impression that KKV's framework consists of rigid rules. Rather,

19. See note 1 above.

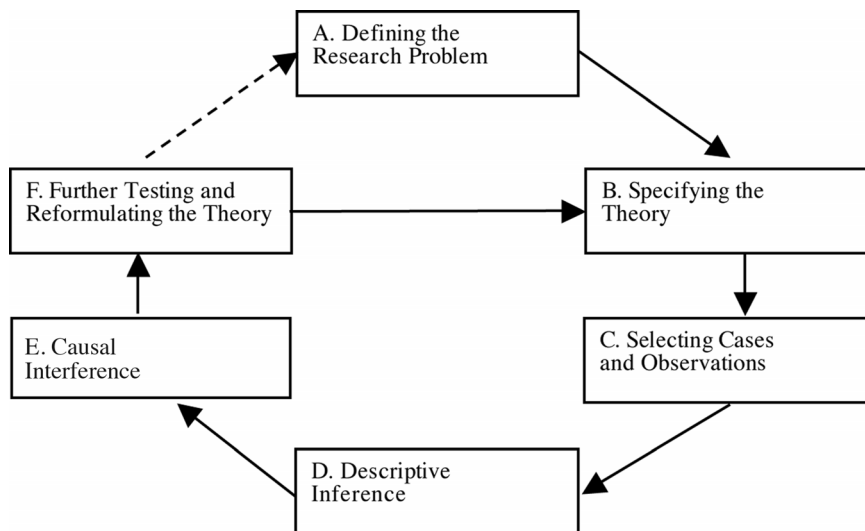


Figure 2.1 Steps in the Research Cycle: A Framework for Summarizing *Designing Social Inquiry*

Note: Solid arrows show the main links among steps in the cycle. Choices made at any one step can, of course, potentially affect any other step. This is reflected, for example, by the placement of a dashed line from F to A, in addition to the solid line from F to B.

we seek to bring together systematically the large number of specific recommendations offered by the book, as a means of demonstrating both the scope of these recommendations, and KKV's relative emphasis on different methodological issues.

A. Defining the Research Problem

1. *Address a problem that is important in the real world* (15).
2. *Contribute to a scholarly literature.* Contribute to "an identifiable scholarly literature by increasing the collective ability to construct verified scientific explanations of some aspect of the world" (15, 16–17).²⁰
3. *Modify or abandon a topic that cannot be refined into a research project that permits valid inference* (18).

B. Specifying the Theory

4. *Construct falsifiable theories.* "[C]hoose theories that could be wrong" (19; also 100).
 - a. *Strengthen falsifiability by choosing a theory that maximizes observable implications* (19).

20. The italics in many quotations have been omitted.

- b. *Strengthen falsifiability by being concrete.* “Theories that are stated precisely and make specific predictions can be shown more easily to be wrong and are therefore better” (20, 109–12).
- 5. *Build theories that are logically consistent.* “[I]f two or more parts of a theory generate hypotheses that contradict one another, then no evidence from the empirical world can uphold the theory” (105).
- 6. *Increase leverage by explaining more with less.* Explain “as much as possible with as little as possible” (29).
 - a. *Increase leverage through parsimony.* “[M]aximize leverage by limiting the number of explanatory variables” (123).
 - b. *Increase leverage by explaining more observable outcomes.* “State theories in as encompassing [a way] as feasible” (113), and “list all possible observable implications of [the main] hypothesis that might be observed in [the] data or in other data” (30).

C. Selecting Cases and Observations

- 7. *Distinguish between cases and observations.* “Cases” are understood as the broader units, that is, the broader research settings or sites within which analysis is conducted; “observations” are pieces of data, drawn from those research sites, that form the direct basis for descriptive and causal inference (52–53, 117–18, 217–18).
- 8. *Focus on the range of variation relevant to the theory.* Select cases among which the dependent variable in fact exhibits “the variation [researchers] wish to explain” (108). It is thus important not merely to have variation on the dependent variable, but that this variation capture the contrasts addressed by the theory.
- 9. *Construct a determinate, rather than an indeterminate, research design by including a sufficient number of observations.*²¹ Avoid an indeterminate research design from which “virtually nothing can be learned about the causal hypotheses” because the researcher has “more inferences to make than implications observed” (118, 119; also 116, 120, 178–79, 213–17, 228). In the face of an insufficient number of observations, scholars can:
 - a. *Address indeterminacy by increasing the number of observations—either through changing the dependent variable, or through focusing on subunits* (24, 47, 120, 217–28).
 - b. *Address indeterminacy by gaining leverage from strong theory.* If the number of observations is insufficient, “limited progress in understanding causal issues is nevertheless possible, if the theo-

21. A determinate research design also requires the absence of perfect multicollinearity. This likewise involves the issue of having enough observations, in that a sufficiently large N can help overcome multicollinearity. See no. 30 below.

retical issues with which [researchers] are concerned are posed with sufficient clarity and linked to appropriate observable implications" (179).

- c. *Address indeterminacy by situating observations within a larger research program.* Even "a single observation can be useful for evaluating causal explanations if it is part of a research program. If there are other single observations, perhaps gathered by other researchers, against which it can be compared, it is no longer a single observation" (211, 129 n. 6).
10. *Seek causal homogeneity.* Causal homogeneity²² is "the assumption that all units with the same value of the explanatory variables have the same expected value of the dependent variable" (91, 116).
11. *Avoid selection bias.* Selection bias poses important "dangers" (116), in that it can invalidate both causal inference (129–32) and descriptive inference (135). One important source of such bias is the failure of the sample to reflect the full range of variation on the dependent variable. The random selection of cases is a standard means for avoiding important forms of selection bias, yet in small-N research this may not be appropriate (126).
12. *Select cases nonrandomly in small-N analysis.* Random selection in small-N research can too easily fail to capture the full range of variation on the variables of interest. "Usually, selection must be done in an *intentional* fashion, consistent with . . . research objectives and strategy" (139). This recommendation is relevant both for descriptive (135) and causal (129–32) inference. With reference to causal inference, KKV suggests the following standards for nonrandom selection:
 - a. *Avoid selecting a set of observations in which either the independent or dependent variable is constant.* "[T]he causal effect of an explanatory variable that does not vary cannot be assessed . . ." (146). Researchers "can also learn nothing about a causal effect from a study which selects observations so that the dependent variable does not vary" (147; also 108–9, 129, 148–49). "The cases of extreme selection bias—where there is by design no variation on the dependent variable—are easy to deal with: avoid them!" (130).
 - i. *In selecting observations on either the independent or dependent variable, ensure that these observations encompass sufficient variation on this variable.* For example, when selecting on the dependent variable, "select observations with particularly high and particularly low values . . ." (129, 141, 147–49).

22. Regarding definitions of causal homogeneity versus unit homogeneity, see the glossary.

- ii. *To address the problem of a no-variance design, seek variance by situating observations within a larger research program* (146–47).
 - b. *Selecting simultaneously on both the independent and dependent variables can pose a grave problem. “The most egregious error is to select observations in which the explanatory and dependent variables vary together in ways that are known to be consistent with the hypothesis that the research purports to test”* (142).
13. *If observations are not independent from one another, recognize that this reduces the certainty of the findings; researchers may also address the causes of this interdependence. When observations are not fully independent of each other, “each new [observation] does not bring as much new information to bear on the problem as it would if the observations were independent of one another. . . . [W]hen dealing with partially dependent observations . . . be careful not to overstate the certainty of the conclusions. . . . [C]arefully analyze the reasons for the dependence among the observations”* (222).

D. Descriptive Inference

14. *Description requires inference. Description in social science research must be understood not as the process of collecting unmediated facts, but rather as involving inferences from observations to the broader ideas and comparisons around which the research is organized* (chap. 2).
15. *Recognize the similarity between quantitative or formal work and “interpretation,” as compared to the full complexity of reality. “[T]he difference between the amount of complexity in the world and that in the thickest of descriptions is still vastly larger than the difference between this thickest of descriptions and the most abstract quantitative or formal analysis”* (43).
16. *Extract analytically relevant features from the uniqueness of cases* (42). “All phenomena, all events, are in some sense unique. . . . The real question . . . [is] whether the key features of social reality that we want to understand can be abstracted from a mass of facts” (42).
17. *Know the context. “Where possible, analysts should simplify their descriptions only after they attain an understanding of the richness of history and culture. . . . [R]ich, unstructured knowledge of the historical and cultural context of the phenomena with which they want to deal in a simplified and scientific way is usually a requisite for avoiding simplifications that are simply wrong”* (43).
18. *Good description is better than bad explanation. In research contexts in which good causal inference is difficult, it may be preferable to stick*

- to carefully executed descriptive inference (44; also 34, 45, 75 n. 1, 178–79).
19. *Study observable concepts.* “[C]hoose observable, rather than unobservable, concepts wherever possible” (109). “Attempting to find empirical evidence of abstract, unmeasurable, and unobservable concepts will necessarily prove more difficult and less successful than for many imperfectly conceived specific and concrete concepts” (110).
 20. *In general, avoid typologies and classifications, except as preliminary heuristic devices.* “[C]onstructs such as typologies, frameworks, and all manner of classifications, are useful as temporary devices [for] collecting data. . . . However, in general, we encourage researchers *not* to organize their data in this way” (48).
 21. *Use valid indicators.* “Validity refers to measuring what we think we are measuring” (25). Among the issues that arise in striving for validity is the need to “use the measure that is most appropriate to [the researcher’s] theoretical purposes” (153).
 22. *Use reliable data-collection procedures that, if applied again, would produce the same data* (25).
 23. *Estimate measurement error.* “Since all observation and measurement . . . is imprecise,” researchers should “estimate the amount of [measurement] error . . . ” (151); “qualitative researchers should offer uncertainty estimates in the form of carefully worded judgments about their observations” (152).
 24. *Separate the systematic and random components of phenomena.* “[O]ne of the fundamental goals of [descriptive] inference is to distinguish the systematic component from the nonsystematic component of the phenomena” being studied (56). Thus, analytically productive description may seek to isolate the systematic component, as it is this component that researchers really seek to explain.

E. Causal Inference

25. *Causal assessment requires inference.* Causation is not observed directly. Rather, causation is inferred on the basis of data and assumptions (chap. 3).
26. *Demonstrate, to the extent possible, that the assumptions underlying causal inference are met in a given context of research.* Assumptions such as causal homogeneity, conditional independence, and the independence of observations “can and should be justified” to the greatest extent possible on the basis of insights derived from prior research and knowledge of the research setting (91).
27. *Use theory to select appropriate explanatory variables and avoid “data min-*

- ing." "Without a theoretical model, [researchers] cannot decide which potential explanatory variables should be included in [the] analysis." "[W]ork toward a theoretically motivated model rather than 'data mining' . . ." In other words, researchers should not simply run "regressions or qualitative analyses with whatever explanatory variables [they] can think of" (174).
28. *Avoid missing variable bias by including all relevant explanatory variables.* "[S]ystematically look for omitted control variables and consider whether they should be included in the analysis" (172). If a given variable is correlated with both the dependent variable and an explanatory variable, then failure to include it will bias the causal inference (170). The following three steps can help avoid missing variable bias:
 - a. *First, list potentially relevant explanatory variables* (174).
 - b. *Second, control for relevant explanatory variables* (174).
 - c. *Third, in estimating the main causal effect, do not control for intervening variables.* "[I]n general, [researchers] should not control for an explanatory variable that is in part a consequence of [the] key explanatory variable" (174).
 29. *Minimize the variance of estimators by excluding irrelevant variables.* Do not "collect information on every possible causal influence . . ." (182, italics omitted) because "[t]he inclusion of irrelevant variables can be very costly" (183). While the best solution to the problem of "many variables, small N" is to collect more observations, "if this is not possible, researchers are well-advised to identify irrelevant variables" (184) and exclude them from the analysis.
 30. *Avoid an indeterminate research design due to multicollinearity.*²³ Avoid a research design in which two or more of the explanatory variables are so highly correlated that it is impossible to separate their causal effects (119). The proposed solution to this problem is to:
 - a. *Address multicollinearity by collecting additional observations.* "[S]earch for observable implications at some other level of analysis" (123), which can give more leverage in differentiating the causal effects of highly correlated explanatory variables.
 31. *Avoid endogeneity.* "A very common mistake is to choose a dependent variable which in fact causes changes in [the] explanatory variables. . . . [T]he easiest way to avoid [this mistake] is to choose explanatory variables that are clearly exogenous and dependent variables that are endogenous" (107–8; also 94, 185). Five solutions to endogeneity are:
 - a. *Address endogeneity by careful selection of observations.* "[W]e can

23. A determinate research design also requires a sufficient number of observations. See guideline 9 above.

- first translate a general concern about endogeneity into [a concern about] specific potential sources of omitted variable bias and then search for a subset of observations in which these sources of bias could not apply" (193).
- b. *Address endogeneity by transforming it into an omitted variable problem.* "By transforming [a research] problem in this way, scholars [can] get a better handle on the problem since they [can] explicitly measure this omitted variable and control for it . . ." (190).
 - c. *Address endogeneity by disaggregating the dependent variable.* "[R]econceptualize the dependent variable as itself containing a dependent and an explanatory component. . . . The goal of this method of avoiding endogeneity bias is to identify and measure only the dependent component of [the] dependent variable" (188–89).
 - d. *Address endogeneity by disaggregating the explanatory variable.* "[D]ivide a potentially endogenous explanatory variable into two components: one that is clearly exogenous and one that is at least partly endogenous. . . ." Then use "only the exogenous portion of the explanatory variable in a causal analysis" (193).
 - e. *Address endogeneity by correcting the biased inference.* "[E]ven if [researchers] cannot avoid endogeneity bias, [they] can sometimes improve . . . inferences after the fact by estimating the degree of bias. At a minimum, this enables [them] to determine the direction of bias, perhaps providing an upper or lower bound on the correct estimate" (188).
32. *Estimate and, if possible, correct for selection bias.* "[I]f selection bias is unavoidable, [researchers] should analyze the problem and ascertain the direction and, if possible, the magnitude of the bias, then use this information to adjust [their] original estimates in the right direction" (133). If they "know there is bias but cannot determine its direction or magnitude . . . [researchers should] at least increase the level of uncertainty [they] use in describing [their] results" (199; also 128–37, 168–82).

F. Further Testing and Reformulating the Theory

33. *Report research procedures, thereby allowing other analysts to evaluate and replicate the findings.* "Only by reporting the study in sufficient detail so that it can be replicated is it possible to evaluate the procedures followed and methods used" (26; also 8, 23, 51).
34. *Test the theory with data other than that used to generate the theory* (46). The original data can be used to test a new implication of a theory,

“as long as the implication does not ‘come out of’ the data but is a hypothesis independently suggested by the theory or a different data set” (30).

35. *The theory should generally not be reformulated after analyzing the data.* “Ad hoc adjustments in a theory that does not fit existing data must be used rarely . . . ” (21).
 - a. *If the theory is reformulated by making it more restrictive, retest it with new data.* If a theory is modified after analyzing the data, researchers “can make the theory less restrictive (so that it covers a broader range of phenomena and is exposed to more opportunities for falsification), but [they] should not make it more restrictive without collecting new data to test the new version of the theory” (22, italics omitted).

ANTICIPATING THE DISCUSSION OF KKV'S FRAMEWORK

Subsequent chapters in the present volume provide alternative perspectives on quantitative and qualitative methods, making central reference to the framework offered by KKV. This final section of chapter 2 anticipates the assessment presented in the following chapters.²⁴ As can be seen in table 2.2, we organize the discussion with reference to specific guidelines. Some aspects of KKV's framework evoke agreement, whereas for others there is disagreement.

I. Areas of Convergence

a. *Broad Convergence.* The chapters in this volume strongly endorse the overall goal of developing shared standards for descriptive and causal inference. This convergence once again calls attention to the contribution made by KKV in focusing scholarly attention on such standards.

b. *Specific Points of Convergence.* Many of KKV's suggestions are not challenged or reevaluated. The recommendation to move beyond the uniqueness of cases by extracting analytically relevant features (guideline no. 16 above) articulates a fundamental priority in social science research. KKV's suggestion to distinguish between cases and observations (no. 7) and the discussion of descriptive and causal inference (nos. 14, 25) have given some qualitative researchers a useful new vocabulary. As noted earlier in

24. Whereas the last section in chapter 1 above summarizes the arguments chapter by chapter, the organization here is thematic.

Table 2.2. Anticipating the Debate on *Designing Social Inquiry*

<i>Evaluation of KKV's Contribution and Selected Examples Drawn from Guidelines Presented in Chapter 2</i>	<i>Relevant Chapters in RSI (includes online chapters)</i>
I. Areas of Convergence	
a. Broad Convergence. Consensus on importance of standards for good descriptive and causal inference.	All Chapters
b. Specific Points of Convergence. Consensus that KKV offers much valuable advice with direct practical application in social science research (2, 3, 4, 5, 7, 12, 12b, 14, 16, 25, 31b/c/d, 33).	All Chapters
II. Areas of Divergence	
a. Extensive Treatment of Causal Inference, but Insufficient Attention to Its Logical Foundations. Greater attention needed to adequately address the obstacles faced in causal inference based on observational data (10, 26, 28, 29, 31, 31a/b/c/d/e, 32).	Brady (chap. 3); Bartels; Collier, Brady, and Seawright (chap. 9); Ragin (online); McKeown (online)
b. Important Issues Are Noted, but Seriously Neglected. Valuable advice is discussed briefly, but this advice must play a far more central role in research design (8, 9b/c, 12a–ii, 17, 21, 22).	Brady (chap. 3); Rogowski; Collier, Mahoney, and Seawright (online); Ragin (online); McKeown (online)
c. Regarding Key Advice, Practical Application May Not Be Feasible. Some advice may be hard to apply, not only in qualitative, but even in quantitative, research (13, 18, 23, 26, 28c, 31).	Brady (chap. 3); Bartels; Collier, Brady, and Seawright (chap. 9); Munck (online); McKeown (online)
d. Idea of Trade-Offs Is Mentioned, but Not Recognized as a Central Issue. Trade-offs among methodological goals must be a central concern in designing research (4a, 6b, 9, 9a, 11, 12a, 19, 27, 30a, 31, 34, 35, 35a).	Brady (chap. 3); Bartels; Rogowski; Tarrow; Collier, Brady, and Seawright (chaps. 8 and 9); Collier, Mahoney, and Seawright (online), Munck (online), Ragin (online)
e. Independent Contribution of Qualitative Tools Is Undervalued. Qualitative analysts have developed valuable tools that must to a greater degree be taken seriously on their own terms (1, 10, 13, 15, 17, 21, 22, 24, 30, 31).	Rogowski; Tarrow; Collier, Brady, and Seawright (chap. 9); Brady (chap. 12); Collier, Mahoney, and Seawright (online); Munck (online); Ragin (online); McKeown (online)

this chapter, part of the advice about selection bias is quite nuanced, in that KKV recognizes the importance of nonrandom sampling in the context of small-N research. Rather than offering the excessively limiting recommendation that scholars should not select on the dependent variable, the book suggests how sampling on the dependent variable is best carried out (no. 12). Replicability (no. 33) is certainly a widely held goal in the social sciences,²⁵ and other areas of agreement likewise emerge, as indicated in the table.

II. Areas of Divergence

In a number of other areas, the authors in the present volume raise questions about KKV's recommendations.

a. *Extensive Treatment of Causal Inference, but Insufficient Attention to Its Logical Foundations.* KKV is on the right track in pushing analysts to consider the assumptions that constitute the logical foundations of inference. However, the book's presentation of methodological norms falls short in helping scholars include the right variables, exclude the wrong ones, and more generally design their research and specify their models appropriately.

KKV's suggestion that researchers systematically search for and include relevant omitted variables (no. 28) usefully raises the issue of confounding variables, but does not say enough about which kinds of omitted variables ought to be included and which should be excluded. The recommendation that researchers exclude irrelevant explanatory variables (no. 29) leaves the same kinds of questions unanswered: How, exactly, should analysts distinguish between relevant and irrelevant explanatory variables before making a causal inference? Likewise, the advice that analysts should avoid endogeneity (no. 31) does too little to help researchers understand the substantive and theoretical reasons that endogeneity might or might not be a problem in a particular context. The specific techniques for addressing problems of endogeneity (nos. 31a–e) are valuable in pushing analysts to seek solutions to these problems, but much more needs to be said about the rather stringent assumptions behind these techniques.

Overall, KKV appears to embrace the proposition that these key problems of causal inference have been largely solved in mainstream quantitative research, and that, by extension, qualitative researchers should come as close as they can to adopting these solutions. By contrast, as argued by Brady, Bartels, and Seawright (chaps. 3, 4, 13, this volume), we are convinced that causal inference—not only in qualitative but also in quantitative research—is often problematic. Related issues of the logical founda-

25. Gary King has played a central role in subsequent debate on this issue. See *PS: Political Science and Politics* (1995) and *APSA-CP* (1996).

tions of inference are addressed by Ragin and McKeown in their online chapters.

KKV simply does not confront these difficulties squarely. The book does not give adequate recognition to problems of causal inference created by omitted variables and endogeneity. These issues are not easily resolved, even with advanced quantitative techniques. Consequently, causal inference, even with a large N , is often problematic. Hence, the applicability of KKV's methodological framework for causal inference in qualitative research remains doubtful.

b. *Important Issues Are Noted, but Seriously Neglected.* KKV mentions some key issues once or perhaps twice, yet some authors in the present volume consider them to be fundamental problems in the design of research that require far more attention. For example, KKV does cite Lieberman's (1985: chap. 5) incisive discussion of the need to focus empirical analysis on the range of variation relevant to the theory (no. 8); KKV also refers to using strong theory to address the problem of indeterminacy (no. 9b). Likewise, KKV notes that situating observations within a larger research program can help address the small- N problem (indeterminacy) and the problem of no-variance designs (nos. 9c, 12a-ii). Further, the book does mention the importance of knowing the context of research and of seeking validity and reliability in measurement (nos. 17, 21, 22). However, although these topics are noted briefly, they require much greater attention, given that KKV aims to provide a balanced set of recommendations for research design. These themes are explored below in the chapters by Brady and Rogowski. See also the online chapters by Collier, Mahoney, and Seawright; Ragin; and McKeown.

c. *Regarding Key Advice, Practical Application May Not Be Feasible.* Many of KKV's guidelines offer potentially useful methodological recommendations, yet authors in the present volume are concerned that it sometimes may not be feasible to apply this advice. For example, KKV usefully suggests that researchers pay close attention to the implications of measurement error for causal inference (no. 23). However, as Bartels argues, current statistical knowledge suggests that it can be difficult to know what those consequences are, even in quantitative research. Likewise, it is probably good advice to suggest that, in contexts where good causal inference is difficult, it is preferable to stick to good descriptive inference (no. 18). Yet this advice runs against the prevailing intellectual orientation within political science (and in KKV), where causal inference is strongly privileged over descriptive inference. As Brady (chapter 3) and McKeown (online chapter 4) argue, more reflection is needed on the proper relation between descriptive and causal inference.

Returning to the topic of endogeneity (no. 31), we find it useful to raise this issue, but it is also valuable to be candid about the fact that it can be

exceedingly hard to address this problem, in either qualitative or quantitative research. Finally, the priority of demonstrating that the assumptions underlying causal inference are met in a given context of research (no. 26) is obviously important—as discussed in chapter 9 and in the online chapter by Munck—but little attention is devoted to exploring how this is to be done. In many contexts, it is simply not possible to demonstrate that these assumptions are met.

d. *Idea of Trade-Offs Is Mentioned, but Not Recognized as a Central Issue.* KKV pays insufficient attention to trade-offs, failing to recognize that they are an overarching issue in research design. Trade-offs are a central theme in the chapters below. As discussed in this volume by Brady (chap. 3) and Bartels, and in chapters 8 and 9, the mandate to increase the number of observations—for the purpose of strengthening falsifiability, increasing leverage, and addressing indeterminacy and multicollinearity (nos. 4a, 6b, 9a, 30a)—may make it harder to achieve other important goals, such as maintaining independence of observations, measurement validity, and causal homogeneity.

Next, as emphasized by Brady (chap. 3) and in chapter 8 of this volume, while working with concrete and observable concepts (no. 19) certainly makes measurement easier, many theories depend on abstract concepts that are well worth measuring, even if it is not easy to do so. An obvious example is the concept of causation. KKV (76, 79) in fact recognizes it as an abstract, theoretical concept, and much of the book is devoted to discussing how best to measure it. Many other indispensable concepts are likewise hard to measure.

Additionally, the idea of a determinate versus indeterminate research design (no. 9) raises the important issue of having a sufficient number of observations to adjudicate among rival explanations; yet, as chapter 9 in the present volume argues, this distinction creates the misleading impression that research designs based on observational, as opposed to experimental, data can really be determinate—which is not the case. Indeed, causation can generally only be inferred in observational studies if the researcher imposes several restrictive assumptions, which may be difficult to test or even to defend.

Finally, as argued by Rogowski, and by Collier, Mahoney, and Seawright, the warning against designs that lack variance on the dependent variable (no. 12a) must be weighed against the analytic gains that can derive from closely analyzing positive cases of a given phenomenon, especially if little is known about it.

Other recommendations made by KKV also involve trade-offs. These recommendations involve issues of inductive analysis, endogeneity, and complexity. From one point of view, the injunctions against the post hoc reformulation and testing of hypotheses (nos. 34, 35, 35a) make good

sense, in that it weakens the power of statistical tests. However, as Ragin (online chapter 3), Munck (online chapter 2), and Tarrow argue, for qualitative researchers the refinement of theory and hypotheses through the iterated analysis of a given set of data is an essential research tool, and researchers lose other aspects of analytic leverage by not employing it.²⁶ Indeed, quantitative studies regularly follow a similar path. When quantitative researchers analyze observational data, they almost never conduct one test of the initially hypothesized statistical model and then stop. Rather, they routinely carry out elaborate specification searches, involving iterated attempts to find an appropriate fit between models and data. For this reason, a major literature within econometrics has discussed procedures and tools that help quantitative researchers conduct their specification searches in a disciplined manner. This literature recognizes that the quantitative analysis of observational data routinely involves an iterated, partly inductive, mode of research.

A closely related point concerns data mining. Indiscriminate data mining is a bad idea, and the statement that selecting relevant explanatory variables requires theory is uncontroversial (no. 27). However, as just noted, all research has an inductive component, and we should not foreclose the possibility of accidental discoveries. The challenge is to be open to such discoveries that are not anticipated by our theory; yet at the same time to avoid the atheoretical, indiscriminate pursuit of new hypotheses, which may lead to findings that are not analytically meaningful.

Finally, returning to the issue of endogeneity (no. 31), selecting cases so as to avoid this problem makes sense in that it facilitates causal inference. Yet this priority absolutely should not preclude, for example, looking at processes of change over time, where endogeneity is commonly present. Given the larger intellectual movement in recent decades toward the historicization of the social sciences, scholars who study causal processes over a long time horizon must routinely treat endogeneity as a problem to be confronted, rather than avoided.

e. *Independent Contribution of Qualitative Tools Is Undervalued.* KKV pays insufficient attention to the independent contributions of qualitative tools, sometimes too quickly subordinating them to a quantitative template. KKV makes an interesting argument that quantitative/formal work and interpretation are *similar* in an important respect: both simplify drastically, compared to the full complexity of reality (no. 15). While this is true, for the researcher trying to learn about the distinctive strengths of alternative methodological approaches, the dissimilarity of interpretation and quanti-

26. KKV does discuss the interaction between theory and data, but within the framework of arguing that any further test of the theory should be undertaken with *new* data (KKV 21, 46).

tative/formal analysis is a far more central concern, a theme that arises in chapter 13 below. KKV's framing inappropriately deemphasizes the contributions of interpretive work, and of other qualitative approaches, to goals that a regression-oriented framework addresses much less successfully—including concept formation and fine-grained description.

Qualitative researchers also have distinctive perspectives on causal heterogeneity (no. 10). It is a central component within Ragin's framework, and Tarrow shows how qualitative methods provide valuable tools for explaining transitions and nonlinearity that have been discovered through quantitative analysis. With reference to separating the systematic and the random components of phenomena (no. 24), Munck suggests that qualitative researchers may approach this issue by employing insights about causal mechanisms and the larger research context. Isolating the systematic components can, in turn, provide a substitute for statistical control by eliminating the variance on the dependent variable caused by factors outside the focus of the analysis.

Finally, and most importantly, KKV's arguments about strengthening causal inference through increasing the number of observations can be refined by recognizing the importance of different kinds of observations: that is, data-set observations and causal-process observations, a distinction introduced in chapter 1 above and explored at length in chapter 13 and in the appendix. Utilizing this distinction makes it easier to recognize the valuable leverage in causal inference that derives from within-case analysis—which has been a long-standing focus in discussions of qualitative methods and is an important concern in the chapter below by Rogowski, the online chapters by Collier, Mahoney, and Seawright, Munck, and McKeown, as well as in Tarrow's discussion of triangulation. KKV notes these procedures, but the book prematurely seeks to subordinate them to the standard tools of quantitative inference (KKV 85–87, 226–28).

To conclude, KKV articulates a clear summary of the mainstream quantitative framework in social science. At the same time, the book seeks to impose this framework on other kinds of research. In the process, KKV loses sight both of major weaknesses in the quantitative template and of many strengths that have made other tools worth developing in the first place. KKV's arguments have stimulated scholars to rethink both the quantitative and qualitative traditions. Based on this rethinking, the chapters below seek to present a more balanced view of methodology and research design.

B. CRITIQUES OF THE QUANTITATIVE TEMPLATE

3

Doing Good and Doing Better: How Far Does the Quantitative Template Get Us?

Henry E. Brady

What kind of contribution is *Designing Social Inquiry* (hereafter KKV) by Gary King, Robert O. Keohane, and Sidney Verba? Consider the traditional distinction between theology and homiletics.

THEOLOGY VERSUS HOMILETICS

Theological seminaries distinguish between theology, or the systematic study of religious beliefs, and homiletics, the art of preaching the gospel convincingly. Theologians ask hard questions, develop new systems of theology, and often espouse opinions that would shock and horrify the practicing and devout members of the religion's congregations. Homiletics is about homilies; it is about sermons that are practical, down to earth, simple, and above all, reliable interpretations of the faith. Religions understand, as the social sciences may not, that the goal is to save souls and not simply to increase our knowledge or understanding of the world. For this reason, both theology and homiletics have pride of place in seminaries.

The social sciences have a great deal of theology, but very little homiletics. Perhaps this is why we have saved so few souls. And it may also be why we do such a bad job of training students. A little homiletics might go a long way toward improving our discipline.

KKV is a homily, not theology. There is art in a good homily. Like all good homiletic literature, KKV puts aside doubt and complexity. After all, who would want to burden the average graduate student with the tedious complexity of St. Thomas Aquinas in *Summa Theologica* or Paul Tillich in *Systematic Theology*? And who would recommend the self-doubt of St. Augustine's *Confessions* or Kierkegaard's *Fear and Trembling* or *The Sickness unto Death*? Better to give them Norman Vincent Peale's *The Power of Positive Thinking*.

KKV, however, is not just about positive thinking. It is closer to Moses Maimonides' *Guide for the Perplexed* or Luther's *A Catechism for the People, Pastor and Preacher*. It has a powerful message about the need for reform, self-sacrifice, and discipline on the part of all political scientists—especially qualitative researchers.¹ It puts forth a simple, straightforward faith. It tries very hard to treat qualitative researchers as souls worthy of salvation. And it envisions a unified social science in which there are “Two Styles of Research, One Logic of Inference” (3).² To practice this one logic of inference, KKV presents a simple, unified series of steps, a faith to live by, based upon insights from conventional quantitative methods and econometrics. In chapter 3, for example, we are told to:

- Construct falsifiable theories.
- Build theories that are internally consistent.
- Select dependent variables carefully.
- Maximize concreteness.
- State theories in as encompassing a way as possible.

1. *Designing Social Inquiry*, subtitled *Scientific Inference in Qualitative Research*, begins by discussing the relationship between quantitative and qualitative research, but another dichotomy also runs through the book. Quite often the authors are more concerned with juxtaposing “small-N” versus “large-N” research than with the qualitative-quantitative distinction. These are not the same things. Small-N research is often qualitative, but it need not be, and large-N research can be qualitative. Roughly speaking, the qualitative-quantitative distinction revolves around issues of concept formation and measurement whereas the small-N versus large-N distinction brings up problems of defining the relevant populations, sampling from them, and dealing with statistical variability. I argue later in this chapter that these statistical issues are dealt with much more clearly in KKV than are those regarding concept formation and measurement. We return to these issues in chapter 12 below.

2. This phrase resonates especially well with someone like myself who was brought up as a Catholic where the faithful must deal with the mystery of three manifestations of God (in the Father, Son, and Holy Spirit) in a monotheistic religion. By childhood training, I am quite receptive to a message of monomethodism, even in those circumstances where it requires a leap of faith.

In homiletic literature, exhortations such as these should be simple, and they need not always be completely consistent (witness the last two rules listed above). A good sermon should have clear points; it should avoid doubt; it should provide plenty of examples. The goal should be to convert the heathen qualitative researcher to the true faith.

This book—to its credit—does these things. It is an extraordinarily good piece of homiletic literature and it should be used in the classroom. It is very nicely written. It is generally lucid and well organized. No one can fail to hear its message.

And indeed, we should all hear the message that is preached. I, for one, have great sympathy with this enterprise, having spent far too many hours listening to talks on comparative politics in which dependent variables or independent variables (or both) did not vary, in which selection bias seemed insurmountable, in which explanations seemed more like good stories than hard-won insights gained from ruling out alternative possibilities. In my introductory statistics classes, I, too, have tried to point out to comparativists that they could do so much better if they avoided omitted variable bias, stopped selecting on the dependent variable, and so forth. I have used some of the same diagrams displayed in the text of KKV (e.g., figures 4.1, 5.1, and 5.2) to make didactic points about good research.

Why, then, do I find myself worried about what this book tries to do? Perhaps I am worried because, despite the authors' desire for a unified approach to social science, there may be something wrong with quantitative researchers³—who luxuriate in large numbers of observations and even the possibility under some circumstances of doing experiments—trying to impose a code of conduct, a morality, taken from their own experiences. Certainly the authors, three of the most distinguished and intelligent political scientists in our discipline, mean well, think well, and write well. But I worry that, in the end, they are a little like the Reverend Ike who, when asked how he reconciled living in luxury while he preached to the poor, responded that he believed that the best thing you could do for the poor was not to be one of them. The book ends, in fact, with a chapter on "Increasing the Number of Observations."⁴ Is this the best thing we can do for qualitative researchers: to recommend that they not be "small-N" researchers?

Qualitative researchers may indeed profit by increasing the number of

3. Keohane is not a quantitative researcher, but two of the authors, King and Verba, certainly are, and the book's approach is so rooted in quantitative research that it seems fair to make this assertion.

4. This chapter means more and does more than just suggest that qualitative researchers get more data, although that is one of the recommendations. I make more comments about this interesting chapter later in the review.

observations, and one of the great strengths of KKV is that it tries to indicate how the poor in observations can become richer in their understanding. At the same time, the book's unspoken presumption that qualitative researchers are inevitably handicapped by lack of quantification and small numbers of observations is bothersome. It ignores the possibility that quantitative researchers may sometimes be handicapped by procrustean quantification and a jumble of dissimilar cases.

DESCENDING FROM THE RHETORICAL HEIGHTS

I have a number of specific concerns about KKV. Here I will focus on two: my belief that KKV is handicapped by a view of causality too closely tied to the experimental method, and my desire to see more discussion of measurement problems.

Before addressing these concerns, I wish to establish a fair standard for evaluating KKV. Given that I consider KKV to be a homily, and not a work of theology, it may be worth remarking that the value of the *Baltimore Catechism* in which I was drilled as a child should *not* be measured by its logic and argument. Rather it should be evaluated in terms of how many children it saved from perdition. In the end, I think that is how KKV should be judged. Does it work in a classroom? Does it make us better social scientists? By opening up a dialogue with qualitative researchers, the book does make us better, but in its treatment of causation and measurement, KKV may not help us very much.

Explanation and Causality

After a useful discussion of descriptive inference or "establishing facts" in chapter 2, KKV goes on in chapter 3 to discuss "Causality and Causal Inference." As far as I can tell, they equate explanation with causal thinking.⁵ Yet

5. It is not exactly clear how "explanation" fits into KKV's categories of descriptive and causal inference, but one reasonable interpretation is that the authors consider explanation to be identical with causal inference. In the first three paragraphs of chapter 2, they repeatedly refer to the "dual goals of describing and explaining" (34). They also note that "description and explanation both depend upon rules of scientific inference. In this chapter we focus on description and descriptive inference" (34). This suggests that chapter 3, on "Causal Inference," is about explanation. Yet, things cannot be quite so simple, because they go on to say that "as should be clear, we disagree with those who denigrate 'mere' description. Even if explanation—connecting causes and effects—is the ultimate goal, description has a central role in all explanation, and it is fundamentally important in and of itself." The first part of the sentence seems to define explanation as "connecting causes and effects," but the second part seems to say that description is also a form of explanation. In

philosophers of science are not so sure that the only kind of explanation involves causality. Take, for example, “classification” explanations such as the observation that iron has certain properties because it appears in a certain column of the periodic table. This does not appear to be a causal explanation.⁶ It could be argued that Bohr’s atomic theory and its extensions in modern quantum mechanics provide a causal explanation, but this only amounts to saying that there may be causal explanations as well as classification explanations. Moreover, there was a substantial period of time when the classification explanation was all we had. Should we discard these explanations, even when they are all we have, because they do not appear to be causal? We are not so rich with explanations in the social sciences that we can afford to do this without good reason. Qualitative social scientists, in fact, seem especially fond of typologies and classification systems. Do these tools contribute to the explanatory enterprise? I do not personally have an answer to my question, so perhaps I should not fault KKV for failing to include a discussion of this difficult issue. But it is perplexing and thought provoking.

The approach to causality advanced in KKV is based upon an interesting framework developed by the statisticians Donald Rubin (1974, 1978) and Paul Holland (1986). The great strength of this approach, to my mind, is that it emphasizes that a definition of causality requires (a) the careful description of a counterfactual condition (what would have happened if the cause had been absent?) and (b) a comparison of what did happen with what would have happened had the cause been absent. These are two powerful points, and KKV is to be commended for bringing them to the forefront of our discussion. Researchers of all stripes should spend more time describing the counterfactual world that underlies their “because.” What does it mean, for example, to say that “turnout is lower in that district because it has a high proportion of minorities”? What is the counterfactual

the sentence after this one, KKV retains the duality of description and explanation and seems to equate explanation with causal inference, but the book argues for the primacy of inference over either one: “It is not description versus explanation that distinguishes scientific research from other research; it is whether systematic inference is conducted according to valid procedures. Inference, whether descriptive or causal, qualitative or quantitative, is the ultimate goal of all good social science” (34).

6. For more discussion of this example and whether there are noncausal explanations, see Achinstein (1983: chap. 7). Brody and Grandy (1989) provide an excellent set of readings on these topics. Gary King has suggested (personal communication) that classification is a form of descriptive inference, but this seems to stretch KKV’s concept of descriptive inference beyond distinguishing “the systematic component from the nonsystematic component of the phenomena we study” (56). It also adds to the confusion noted in the preceding footnote.

world in which turnout would be higher? Is it simply one with a lower proportion of minorities? Would these nonminorities be like minorities in every other respect except race? How could this happen? What would it mean to have it happen?⁷ These are not easy questions.

I have already argued that there might be explanation without causality. I think there might also be causal effects without (much) explanation. Suppose we find, to use KKV's example, that incumbent legislators do better in elections than nonincumbent legislators. Suppose, in fact, we are as certain as we can be about this because we have done an experiment (random term limits, for example) with a large N to test it out. This finding immediately leads to other questions about what aspects of incumbency create this advantage (see, for example, Cain, Ferejohn, and Fiorina 1987). These questions amount to a desire to further specify the causal mechanism. KKV is not averse to specifying causal mechanisms, and the authors say that "any coherent account of causality needs to specify how the effects are exerted," but they believe that "our definition of causality is logically prior to the identification of causal mechanisms" (85–86). This claim of logical priority may or may not be true (I am not sure it is very important), but what is true is that a discussion of causality is inevitably tied up with a discussion of explanation, theories, and causal mechanisms, and KKV does not pay enough attention to this relationship. There is no discussion of Hempel's (1965) covering laws, of Wesley Salmon's (1984) model of statistical explanation, of Scriven's (1975) "Causation as Explanation," and many other important works on this topic. This is surprising because the philosophical literature, at least, cannot seem to separate the discussion of these issues.⁸

The statistics literature, in fact, is exceptional in defining causality without discussing explanation. Perhaps this is because statisticians want a method of inference that relies only upon the research design and the data, and not at all upon the substance of the research. Yet the net result of the Rubin-Holland papers is a definition that seems surprisingly distant from

7. I have deliberately chosen an example in which the putative cause is a characteristic that might be thought unchangeable. Holland, for example, argues that it is impermissible to call race or gender a cause because "for causal inference, it is critical that each unit be potentially exposable to any one of the causes. As an example, the schooling a student receives can be a cause, in our sense, of the student's performance on a test, whereas the student's race or gender cannot" (Holland 1986: 946). This point is not much in evidence in KKV, and I think the authors were wise to minimize its importance because it certainly seems possible to imagine a world in which gender or race changes, but nothing else.

8. Brody and Grandy (1989), for example, link them in part 2 of their reader entitled "Explanation and Causality." Scriven (1975) joins the two concepts in his famous article on causation as explanation, and every philosophical writer of whom I am aware deals with explanation and causation together.

the problems of theory building and explanation as it exists in the sciences. Most importantly, this approach provides no guidance on what constitutes a “good” explanation beyond what constitutes a good causal inference. Yet an analysis of the impact of incumbency may be an excellent causal inference while being a bad explanation.⁹

After defining causality, KKV goes on to describe a method for causal inference. In this, as in its definition, KKV is guided by the work of Rubin and Holland. The major strength and weakness of this approach is its reliance upon the metaphor of the controlled experiment for solving the problem of causal inference. Holland tells us that:

because experimentation is such a powerful scientific and statistical tool and one that often introduces clarity into discussions of specific cases of causation, I unabashedly draw on the language and framework of experiments for the model for causal inference. It is not that I believe an experiment is the only proper setting for discussing causality, but I do feel that an experiment is the simplest such setting. (1986: 946)

Fair enough. But it is worrisome that Holland finds it “beyond the scope of this article to apply the model for causal inference to nonrandomized studies” (949). Holland cites other literature (Rubin 1978) that essentially concludes that nonrandomized studies are exceptionally difficult to analyze. It is telling that Rubin’s extension of the basic framework requires modeling “(1) the prior distribution of the potentially observable data, (2) the mechanism that selects experimental units for exposure to treatments and assigns treatments, and (3) the mechanism that chooses values to record for data analysis” (Rubin 1978: 35). This is a lot of modeling, and it only seems possible if we have strong theories to draw upon.

KKV provides a simplified version¹⁰ of the Rubin and Holland framework, and in the process ignores some of its subtleties. The crucial part of

9. If the incumbency example does not persuade, consider a doctor called upon to explain the incidence of psychedelic experiences in a remote culture. In an experiment, the doctor shows that a treated group eating a plant diet consisting of peyote, hemp, beans, carrots, and other plants has a statistically significant increase in their incidence of psychedelic experiences. Thus, eating plants causes psychedelic experiences. This is clearly an incomplete explanation. I wish KKV had discussed by what method I might improve it. I think a discussion of a “good explanation” that went beyond methods for finding causal impacts would have gone a long way toward solving this problem.

10. The authors do add one complexity by making a useful distinction between “realized causal effect” and “random causal effect,” but they suppress so much notation and philosophical discussion in their presentation that many of the nuances in Holland’s (1986) presentation are lost and none of the extensions in Rubin (1978) are discussed.

KKV's argument is its discussion of "Conditional Independence" (KKV 94–96). In the Rubin-Holland setup there are as many dependent random variables as there are variations in the treatment condition or the explanatory variable(s). In the simplest case with two levels of the treatment, this implies two random variables. One describes the values on the dependent variable Y for the situation where all cases in the population¹¹ get one level of the treatment (call this Y^I to match KKV's terminology) and the other is for the values on the dependent variable for the situation where all cases in the population get the other level of the treatment (call this Y^N and assume for simplicity that it is no treatment at all). In the real world and for any feasible design, at least one of these values must be censored for each case. That is, we cannot give a case some treatment and no treatment at the same time. But Y^I and Y^N are not the censored variables; they include the unobserved (and unobservable) values as well as the observed ones. A reasonable definition of the causal effect of the treatment is the average of Y^I minus the average of Y^N , but this quantity cannot be calculated because of the unobserved values in these two random variables.

In the Rubin-Holland framework, a necessary assumption for estimating a causal effect is independence between the assignment of treatments and the random variables Y^I and Y^N . This ensures, for example, that people who are high on Y^I are not more likely to get a high level of treatment than those who are low on Y^I . Consequently, we can be sure, for a large enough sample, that the size of the causal effect is the difference between (a) the average of the dependent variable for those who did get the treatment (this quantity can be calculated) and (b) the average of the dependent variable for those who did not get the treatment (another calculable quantity). One way to achieve this kind of independence is to have correctly carried out randomized experiments.

KKV's discussion of this is a bit opaque, and the authors seem to conflate the independence assumption with conditional independence.¹² Condi-

11. In this exposition I ignore sampling problems by assuming observations are available on all members of the population. If the entire population cannot be observed, then some assumption has to be made about random sampling.

12. This accounts for the confusing set of sentences at the beginning of section 3.3.2 where KKV first says that "conditional independence is the assumption that values are assigned to the explanatory variables independently of the values taken by the dependent variables" and then goes on to say, "that is, after taking into account the explanatory variables (or controlling for them), the process of assigning values to the explanatory variable is independent of both (or, in general two or more) dependent variables, Y_i^N and Y_i^I " (94). The first quoted sentence must refer to the independence assumption (because conditional independence does not assume that the values assigned to the explanatory or control variables are independent of the values of the dependent variables) whereas the second quoted sentence appears to be about conditional independence.

tional independence is the assumption that the values of Y^I and Y^N conditional on “pre-exposure” or “control” variables are independent of the assignment of treatments. This is implied by independence but it is a much less stringent assumption. It is the assumption that is usually required for the analysis of quasi-experiments (Achen 1986). The conflation of these two different assumptions creates difficulties in the exposition because, whereas we have a method of random assignment to treatment for attaining independence, we have no comparable method for ensuring that the conditional independence assumption holds outside of a randomized design. The best we have is the checklist of “threats to internal validity” developed by Donald Campbell with Julian Stanley and Thomas Cook (Campbell and Stanley 1963; Cook and Campbell 1979).¹³ The rest of KKV can be considered another approach to developing a checklist of threats to validity.

Unfortunately, KKV does not allow itself enough pages in this short section to make this very important transition from a discussion of causal inference for experiments to causal inference with “quasi-experiments.” I wish the authors had taken more time to explain the independence assumption in detail and to show how randomized experiments might provide us with an operational procedure that would make this assumption plausible. In doing this, they would no doubt have come to the conclusion presented by Cook and Campbell (and updated and expanded recently by Heckman 1992) that there are many reasons to worry about the efficacy of randomization when humans are involved. There are numerous ways in which human beings can make the treatment endogenous by changing their behaviors. There are additional problems when dropouts (and hence censoring of observations) vary by treatment. And there are the difficulties of truly randomizing units when they are people or groups. Once these problems are recognized for randomized designs, it becomes easier to understand how difficult it is to ensure conditional independence for non-randomized designs.

This transition section might also benefit from a more careful discussion of how theories provide the fundamental basis for making a claim of conditional independence. This is an extraordinarily important step, and knowing how to do it can help researchers avoid the inferential nihilism that has crept into some statisticians’ discussion of causal thinking in the social sciences (e.g., Freedman 1991). According to this line of thinking, randomized experiments are practically the only reliable way to be confident that

13. I was surprised to find that none of Campbell’s publications were referenced in KKV. Besides the books referenced in the text of the present chapter, Campbell’s selected papers on *Methodology and Epistemology for Social Science* (1988) make excellent reading.

the conditions for reasonable inferences are met. Conditional independence is considered a chimera—seldom justifiable and usually accepted by the researcher as a matter of pure faith and nothing more. Indeed, if I accepted a notion of inference as bare of theories and the logic of explanation as that proposed by Rubin and Holland, I might also be skeptical of conditional independence. But I believe it is possible to use our prior knowledge, our theories, to carry out the three modeling steps laid out by Rubin (and cited above). Hence, I am more sanguine about the possibilities for cautiously asserting conditional independence.

It might be argued that I brood unnecessarily over technical points. But the section on “Conditional Independence” is the linchpin of KKV. The book wants to show us that concepts from conventional quantitative methods and econometrics will improve our ability to do qualitative research. It argues that the essence of good social research is establishing causal effects. This, in turn, requires making an assumption about conditional independence. This assumption, the authors believe, can be made plausible by avoiding clear-cut violations of it described in the statistics literature. Yet at the crucial transitional moment the argument seems muddy to me. Exactly how can we rule out the violations identified by quantitative researchers? Do quantitative researchers do a good job in this regard? How sure can I be that conditional independence holds after I have followed the instructions in KKV?

The authors of KKV go on to make many useful observations about causal assessment (although, to be honest, I think that Donald Campbell and his collaborators have more useful lists of threats to validity and more trenchant comments about the problems of doing quasi-experimental research). However, in KKV the crucial argument about assessing causation seems to be missing.

Measurement

KKV devotes eighteen pages to measurement (151–68). About five pages cover the “nominal, ordinal, interval” distinction found in the classic papers by S. S. Stevens (1946, 1951), and the remaining thirteen are about systematic and nonsystematic measurement error. The major results on measurement error are the classic ones dating from at least Tintner (1952) on how error in the dependent variable does not bias regression results whereas error in the independent variable produces bias in regression coefficients—in fact, biases them unambiguously downward in the bivariate case. These are well-known results, often repeated in one form or another in classic primers on research design such as Kerlinger (1979), but I do not think they get at the heart of what can be learned from the extensive literature on measurement.

KKV probably gives such short shrift to measurement because the authors believe that causal inference, roughly what Cook and Campbell call "internal validity," is the central problem of doing good social science. I trace this belief to their decision to equate explanation with causal thinking, and to define causal thinking in terms of a narrow analogy to the experimental model. Through this progression, the problems of theory construction, concepts, and measurement recede into the distance. Yet it seems to me that concept formation, measurement, and measurement validity are important in almost all research and possibly of paramount importance in qualitative research. Certainly notions such as "civil society," "deterrence," "democracy," "nationalism," "material capacity," "corporatism," "group-think," and "credibility" pose extraordinary conceptual problems just as "heat," "motion," and "matter" did for the ancients. It may be comforting for the qualitative researcher to know that the true effects of these error-laden variables are even larger in magnitude than what we would estimate using a standard regression equation, but most qualitative researchers are struggling with much more basic problems such as figuring out what it means to measure their fundamental concepts. These problems are certainly not solved by telling us to decide whether the concept is nominal, ordinal, or interval and by admonishing us to "use the measure that is most appropriate to our theoretical purposes" (KKV 153).

I will not pretend to have the answers to the problems of measurement validity in qualitative research, but I think that the debates on these problems would have been advanced by citing some of the more recent literature in this area. Among the notions that come to mind, let me mention three topics that might have been included. Something might have been said about the conceptualizations of measurement developed by Krantz et al. in their magisterial three-volume work on *Foundations of Measurement* (1971–1990), the related notions put forth by Georg Rasch in his quirky but very influential work on *Probabilistic Models for Some Intelligence and Attainment Tests* (1980 [1960]), and the fascinating *Notes on Social Measurement* (1984a) penned by Otis Dudley Duncan, who followed up this broadside on the limitations of social measurement with a brief for using Rasch models in the social sciences (1984b). These works show that *qualitative* comparisons are the basic building blocks of any approach to measurement, thus bridging the "quantitative-qualitative" divide by showing that the two approaches are intimately related to one another. This discussion would have easily led to a second topic: the dimensionality of concepts, the nature of similarity judgments that often underlie concept formation, and the role of taxonomies and classifications in science. Finally, there might have been a survey of how the LISREL framework (Bollen and Lennox 1991), especially when it is combined with the "multitrait-multimethod

approach" of Campbell and Fiske (1959), sheds light on the practical problems of measurement.

Let me discuss each of these literatures. Duncan's observations on Stevens's scale types are probably the best starting place:

I conclude that the Stevens theory of scale types, pruned of its terribly misleading confusion of classifications and binary variables with N scales, augmented to take more explicit account of the scales used in measuring numerosness and probability, and specified more clearly so that the examples could be properly understood and assessed, has utility in suggesting the appropriate mathematical and numerical treatment of numbers arising from different kinds of measurement. *Still, a theory of scale types is not a theory of measurement.* And I, for one, am doubtful that any amount of study devoted to either of those topics can teach you how to measure social phenomena, though it can conceivably be helpful in understanding exactly what is achieved by a proposed method of measurement or measuring instrument. (1984a: 154, italics added)

Lest anyone miss Duncan's point, his next chapter is entitled "Measurement: The Real Thing." What is "the real thing"? Krantz et al. (1971–1990) provide the fullest answer to this question, but Duncan provides a more accessible treatment. Measurement, Duncan argues, is not the same as quantification, and it must be guided by theories that emphasize the relationships of one measure to another. Take, for example, that favorite illustration of introductory methods classes, the measurement of temperature. Although the development of thermometry involves a complicated interplay between theory and invention, one of the important milestones was the discovery of the gas law for which temperature is proportional to pressure times volume. Thermometry only began to progress beyond crude ordinal distinctions such as cold, warm, or hot to true interval scales once laws like the gas law made it clear that temperature could be measured by the change in volume of some material under constant pressure.

One of the distinctive features of this way of measuring temperature is that it relies upon a simple multiplicative law, which relates temperature to two quantities that can be "extensively" measured. Extensive measurement refers to the use of the standard millimeter, gram, second, or some other quantity that can be duplicated so that a number of them can be added together ("concatenated") and compared with some object or phenomenon whose length, weight, duration, or other feature is unknown. There is no such standard for temperature, but it can still be measured because it is related to two quantities that can be measured extensively (i.e., volume as length times width times height and pressure as mass times length per time and area squared).

A fundamental difficulty facing empirical social science is the apparent impossibility of developing extensive measurements of many important theoretical quantities. Consider, for example, the notion of utility that is basic to both economics and public choice theory. Utility cannot be measured extensively, but economists avoid this difficulty through an ingenious ploy: They throw utility out of their empirical models by deriving demand curves from the maximization of utility with respect to a budget constraint that consists of the sum of prices times quantities. This produces a demand curve—an equation in prices and quantities—both of which can be measured extensively. This ploy, unfortunately, does not appear to be readily available to political scientists.

The contribution of Rasch (1980 [1960]; see also Andrich 1988) and of Krantz et al. (1971–1990) in their method of “conjoint measurement” has been to show how measurement can be carried out without an extensive measure that can be duplicated and combined: all that is needed is the ability to make *qualitative* distinctions about the amount of each of several variables that are thought to be multiplicatively related to one another. Rasch’s method, designed for scoring achievement tests, has the great virtue that it scores both test-takers and the items on the test simultaneously.

All this fancy talk does not provide us with a straightforward way to measure the basic concepts in qualitative social science, yet it does provide us with some clues about how we might go about measuring these concepts. First, it suggests that we have two basic strategies for measurement. We can either try to define a concept extensively (as with length, weight, prices, or quantities) or conjointly (as with achievement tests and subjective probability). Thus we can measure democracy extensively by the fraction of the population enfranchised or by the number of parties, or we can measure it conjointly by using ratings from knowledgeable observers. If we use the second method, as qualitative researchers might be inclined to do, then we might want to think about whether we should scale the raters as well as the countries that are rated. Maybe raters differ in their willingness to call a country a democracy; maybe they even have biases of some sort or another.

Second, this discussion suggests that theories must help to guide the measurement process. In their impressive series of papers on bias in electoral systems, Gelman and King (1994) follow just this strategy with a simple framework for thinking about representation. Steven Fish (1995) also does this (more implicitly than explicitly) in his discussion of the development of civil society in Russia. One of his indicators of civil society is the aggregation of interests by groups, which he describes as the group’s “identification of ‘cleavage issues’ and the formulation of specific goals and agendas [and] . . . the formation of a collective identity, which includes the identification of a membership” (53–54). Although Fish does not provide

a mathematical description of his measure, it could be conceptualized as the degree to which participation or membership in a group is highly correlated with some politically relevant characteristic or cleavage. This amounts to defining this component of civil society as the product of group participation and a politically relevant characteristic—a multiplicative relationship of the sort described by measurement theorists as indicative of true measurement. Fish's approach makes sense partly because it has exactly this form. Hence, measurement theory provides a clear-cut check on when we can say that we have the framework for measuring something.¹⁴

This approach leads immediately into the next topic I mentioned above. There is a very rich literature on the "topology" of measurement that indicates what is required for single or multidimensional measures; what is required for dimensionality itself; what is required before something is considered the same as something else; and under what conditions objects can be better taxonomized using "trees" or Euclidian space. These methods are now widely used in biology to inform studies of evolution. I suspect that they would be quite useful for the qualitative researcher who wants to trace the evolution of the concept of democracy over time, or the similarities and differences among contemporary democracies.¹⁵ After all, qualitative researchers often spend a great deal of time and effort developing typologies and taxonomies.

Finally, although I often worry about the wholesale use of LISREL in survey research, I think the marriage of factor analysis to simultaneous equation modeling in LISREL has made many researchers more aware of measurement problems. Kenneth Bollen (1993) presents an exemplary use of this technique in his analysis of ratings, developed by three different scholars, of political liberties and democratic rule in countries around the world. By having two concepts in mind, Bollen is able to search for "discriminant" as well as "convergent" validity as Campbell and Fiske (1959) tell us we should do. Bollen allows for the possibility that raters may have

14. Gary King (personal communication) suggests that these are points for quantitative researchers and not qualitative researchers because they deal with quantitative measures. Putting aside the fact that a discussion of measurement error or Stevens's scale types assumes the same thing (and the entirety of KKV is based upon the premise that quantitative methods provide lessons for qualitative researchers), it is worth noting that qualitative researchers also engage in comparisons that amount to a form of measurement. Qualitative researchers should know that quantitative research relies upon just the kinds of comparative statements that are at the core of qualitative research. In fact, a discussion of this sort would lead to a conclusion that qualitative and quantitative research are not really different at all.

15. Those interested in these topics should peruse the pages of *Psychometrika* or the *Journal of Classification*. Krantz et al. (1971–1990) also explore many of these issues.

biases, and he finds, for example, that one rater “tends to favor countries in Central America and South America, western industrial nations, and, to a lesser extent, countries in the Oceania region” while providing lower scores for sub-Saharan Africa, Eastern Europe, and Asia. One can imagine extending Bollen’s work by adding other methods for rating democracy and by examining (as he does in a preliminary way) how the characteristics of the raters affect their ratings. Bollen’s work suggests that qualitative researchers might improve their understanding of concepts by considering various definitions of them, by considering concepts closely related to them, and by considering concepts that are different from them. This strategy, for example, is followed by Hanna Pitkin in her classic work on representation (1967).¹⁶

An exploration of measurement issues along the lines sketched above would benefit both quantitative and qualitative researchers. Indeed, a discussion of these matters is worthwhile even if it only shows qualitative researchers how quantitative work must also grapple with complex measurement problems. Because its authors want to be constructive and want to instruct, KKV invariably tries to show how quantitative notions can improve qualitative research. This is laudable, but it leads the authors to neglect the multitude of problems that confront quantitative researchers, and it ignores the extent to which quantification is based upon qualitative judgments. Both qualitative and quantitative researchers might benefit from a less didactic approach that revealed problems as well as putative solutions. This might lead to a common effort to solve problems of concept formation and measurement that vex both quantitative and qualitative researchers.

CONCLUSION

KKV is an excellent sermon, without much condescension, on what qualitative researchers can learn from quantitative researchers. As a work on methodology it has some substantial defects, such as equating explanation with causal inference, proposing a narrow definition of causality, and drawing far too little sustenance from a strong literature on measurement and concept formation. But it also has substantial strengths. First and foremost, it opens a conversation between qualitative and quantitative researchers, and that is very good. Second, its presentation of causal thinking in terms of counterfactual reasoning forces researchers to consider more carefully the

16. Pitkin, of course, describes her methodology as “linguistic” analysis, and quantitative researchers might improve themselves by becoming more familiar with her methods.

counterfactuals behind their putative causal models. Third, it has an interesting discussion of selection bias that should be useful to many researchers.¹⁷ Fourth, the final chapter on "Increasing the Number of Observations" is one of the most important notions in the book. I wish KKV had given more concrete examples of how to do this, and I wish the authors had warned of the dangers of spatial and temporal autocorrelation that can thwart innovative attempts to increase observations, but the basic concept is a very important one.

Students will definitely profit from reading this book. The discipline has already benefited from the discussions it has kicked off. I look forward to seeing a generation of graduate students uplifted and improved by reciting its useful and informative homilies.

17. I wish, however, that they had not used the term "selection bias" (KKV 126) in an example that clearly involves sampling error. The example is presented in a section entitled "The Limits of Random Selection" so the authors may have not meant to use the term "selection bias" except in a colloquial fashion, but it is disconcerting, and certainly confusing, nevertheless.

4

Some Unfulfilled Promises of Quantitative Imperialism

Larry M. Bartels

King, Keohane, and Verba's *Designing Social Inquiry: Scientific Inference in Qualitative Research* (hereafter KKV) is an important addition to the literature on research methodology in political science and throughout the social sciences. It represents a systematic effort by three of the most eminent figures in our discipline to codify the basic precepts of quantitative inference and apply them with uncommon consistency and self-consciousness to the seemingly distinct style of qualitative research that has produced most of the science in most of the social sciences over most of their history. The book seems to me to be remarkably interesting and useful both for its successes, which are considerable, and for its failures, which are also, in my view, considerable.

Here I shall touch only briefly upon one obvious and very important contribution of the book, and upon one respect in which the authors' argument seems to me to be misguided. The rest of my discussion will be devoted to identifying some of the authors' more notable unfulfilled promises—not because they are somehow characteristic of the book as a whole, but because they are among the more important unfulfilled promises of our entire discipline. If KKV stimulates progress on some of these fronts, as I hope and believe it will, the book will turn out to represent a very significant contribution to qualitative methodology.

THE CONTRIBUTION AND A SHORTCOMING

Anyone who thinks about social research primarily in terms of quantitative and statistical inference, as I do, has probably thought—and perhaps even

said out loud—that the world would be a happier place if only qualitative researchers would learn and respect the basic rudiments of quantitative reasoning. By presenting those rudiments clearly, engagingly, and with a minimum of technical apparatus, KKV has helped shine the light of basic methodological knowledge into many rather dark corners of the social sciences. For that we owe its authors profound thanks.

At another level KKV's argument seems to be misguided, although in a way that seems unlikely to have significant practical consequences. It is hard to doubt that "all qualitative and quantitative researchers would benefit by more explicit attention to this logic [i.e., the logic "explicated and formalized clearly in discussions of quantitative research methods"] in the course of designing research" (3). However, it simply does not seem to follow that "all good research can be understood—indeed, is best understood—to derive from the same underlying logic of inference" (4). Even if we set aside theorizing of every sort, from Arrow's (1951) theorem on the incoherence of liberal preference aggregation to Collier and Levitsky's (1997) conceptual analysis of scores of distinct types and subtypes of "democracy," it seems pointless to attempt to force "all good [empirical] research" into the procrustean bed of "scientific inference" set forth by KKV. Would it be fruitful—or even feasible—to recast such diverse works as Michels's *Political Parties* (1915), Polanyi's *The Great Transformation* (1944), Lane's *Political Ideology* (1962), Thompson's *The Making of the English Working Class* (1963), and Fenno's *Home Style* (1978) in the concepts and language of quantitative inference? Or are these not examples of "good research"?

KKV attempts to skirt the limitations of their focus by conceding that "analysts should simplify their descriptions only after they attain an understanding of the richness of history and culture. . . . [R]ich, unstructured knowledge of the historical and cultural context of the phenomena with which they want to deal in a simplified and scientific way is usually a requisite for avoiding simplifications that are simply wrong" (43). But since they provide no scientific criteria for recognizing "understanding" and "unstructured knowledge" when we have it, the system of inference they offer is either too narrow or radically incomplete. Perhaps it doesn't really matter whether we speak of the process of "attain[ing] an understanding" as a poorly understood but indispensable requirement for doing science or as a poorly understood but indispensable part of the scientific process itself. I prefer the latter formulation, but the authors' apparent insistence upon the former will not keep anyone from relying upon—or aspiring to produce—"understanding" and "unstructured knowledge."

OMISSIONS AND AN AGENDA FOR RESEARCH

Most importantly, I am struck by what KKV leaves out of its codification of good inferential practice. I emphasize these limitations because they seem

to suggest (though apparently unintentionally) an excellent agenda for the future development of qualitative and quantitative methodology. As is often the case in scientific work, the silences and failures of the best practitioners may point the way toward a discipline's subsequent successes. Here I shall provide four examples drawn from KKV's discussions of uncertainty, qualitative evidence, measurement error, and multiplying observations.

Uncertainty

One of KKV's most insistent themes concerns the importance of uncertainty in scientific inference. Its authors proclaim that "inferences without uncertainty estimates are not science as we define it" (9), and implore qualitative researchers to get on the scientific bandwagon by including estimates of uncertainty in their research reports (9 and elsewhere). But how, exactly, should well-meaning qualitative researchers implement that advice? Should they simply attempt to report their own subjective uncertainty about their conclusions? How should they attempt to reason from uncertainty about various separate aspects of their research to uncertainty about the end results of that research, if not by the standard quantitative calculus of probability? What sorts of checks on subjective reports of uncertainty about qualitative inferences might be feasible, when even the systematic policing mechanism enshrined in the quantitative approach to inference is routinely abused to the point of absurdity (Leamer 1978, 1983; Freedman 1983)? Since KKV offers so little in the way of concrete guidance, its emphasis on uncertainty can do little more than sensitize researchers to the general limitations of inference in the qualitative mode without providing the tools to overcome those limitations. As far as I know, such tools do not presently exist; but their development should be high on the research agenda of qualitative methodologists.

Qualitative Evidence

KKV's discussion of the respective roles and merits of quantitative and qualitative evidence is equally sketchy. While its authors rightly laud Lisa Martin's (1992) *Coercive Cooperation* and Robert Putnam's (1993) *Making Democracy Work* for combining quantitative and qualitative evidence in especially fruitful ways (5), their discussion provides no clear account of *how*, exactly, Martin's or Putnam's juxtaposition of quantitative and qualitative evidence bolsters the force of their conclusions. Martin's work is rushed precipitously off the stage (as most of KKV's concrete examples are), while Putnam's work only reappears—other than in an unrelated discussion of using alternative quantitative indicators of a single underlying theoretical concept (223–24)—in a discussion of qualitative immersion as a

source of *hypotheses* rather than *evidence*. This in turn leads to the rather patronizing conclusion that “any definition of science that does not include room for ideas regarding the generation of hypotheses is as foolish as an interpretive account that does not care about discovering truth” (38).

There is more going on here than a simple-minded distinction between (qualitative) hypothesis generation and (quantitative) hypothesis testing, or a simple-minded faith that two kinds of evidence are better than one. Qualitative evidence does more than suggest hypotheses, and analyses combining quantitative and qualitative evidence can and sometimes do amount to more than the sum of their parts. The authors of KKV do little to illuminate those facts. But the larger and more important point is that nobody else does very well either. Just as the “persuasive force” of such classic works of social science as V. O. Key’s *Southern Politics in State and Nation* (1984 [1949]), Stouffer et al.’s *The American Soldier* (1949), and Berelson, Lazarsfeld, and McPhee’s *Voting* (1954) “is not easily explained in conventional statistical theory even today” (Achen 1982: 12), neither is the persuasive force of these and other compelling works convincingly accounted for by partisans of interpretive, ethnographic, historical, or any other brand of qualitative inquiry.

With reference to both uncertainty and qualitative evidence, the limitations of KKV’s analysis faithfully reflect the limitations of the existing methodological literature on qualitative inference. Other gaps in KKV’s account are attributable to the limitations of the theory of quantitative inference it offers as a model for qualitative research. As a quantitative methodologist—and the coauthor of a rather optimistic survey of the recent literature in quantitative political methodology (Bartels and Brady 1993)—I am chagrined to notice how wobbly and incomplete are some of the inferential foundations that KKV claims are “explicated and formalized clearly in discussions of quantitative research methods” (3). Again, two examples will suffice to illustrate the point.

Measurement Error

The first example of the weak foundations of inferential claims is KKV’s treatment of measurement error, which—like much of the elementary textbook wisdom on that subject—is both incomplete and unrealistically optimistic. The authors assert that unsystematic (random) measurement error in explanatory variables “unfailingly [biases] inferences in predictable ways. Understanding the nature of these biases will help ameliorate or possibly avoid them” (155). Later, they assert more specifically that the resulting bias “takes a particular form: it results in the estimation of a weaker causal relationship than is the case” (158). At the end of their discussion the authors acknowledge that their analysis is based upon a model with a

single explanatory variable. However, they assert that it “applies just the same if a researcher has many explanatory variables, but only one with substantial random measurement error,” or if researchers “study the effect of each variable sequentially rather than simultaneously” (166). Their only suggestion of potential complications is a claim that “if one has multiple explanatory variables and is simultaneously analyzing their effects, and if each has different kinds of measurement error, we can only ascertain the kinds of biases likely to arise by extending the formal analysis” (166).

KKV’s assertion about the case of several explanatory variables, where only one is measured with substantial error, is quite misleading in failing to note that the bias in the parameter estimate associated with the one variable measured with substantial error will be propagated in complicated ways to all of the other parameter estimates in the analysis. This will bias them upward or downward depending on the pattern of correlations among the various explanatory variables. The book’s assertion about sequential rather than simultaneous analysis of several explanatory variables is also misleading, at least in the sense that the resulting omitted variable bias may mitigate, exacerbate, or reverse the bias attributable to measurement error. And the promise of “ascertain[ing] the kinds of biases likely to arise” in more complicated situations “by extending the formal analysis” (KKV 166) can in general be redeemed only if we have a good deal of prior information about the nature and magnitudes of the various errors—information virtually impossible to come by in all but the most well-understood and data-rich research settings (Achen 1983; Cowden and Hartley 1993). Thus, while it seems useful to have alerted qualitative researchers to the fact that measurement error in explanatory variables may lead to serious biases in parameter estimates, it seems disingenuous to suggest that quantitative tools offer reliable ways to “ameliorate or possibly avoid” (155) those biases in real qualitative research.

Multiplying Observations

The second example is KKV’s chapter on “Increasing the Number of Observations,” which seems equally disingenuous in asserting that “almost any qualitative research design can be reformulated into one with many observations, and that this can often be done without additional costly data collection if the researcher appropriately conceptualizes the observable implications that have already been gathered” (208). While it is right to emphasize the importance of “maximizing leverage” by using the available data to test many implications of a given theory (or even better, of several competing theories), KKV’s discussion obscures the fact that having many *implications* is not the same thing as having many *observations*. In order for our inferences to be valid, each of our many implications must itself be

verified using a research design that avoids the pitfall of “indeterminacy” inherent in having more explanatory variables than relevant observations.

What, then, is a “relevant observation”? KKV provides the answer in its earlier, clear, and careful discussion of causal homogeneity.¹ Relevant observations are those for which “all units with the same value of the explanatory variables have the same expected value of the dependent variable” (91). But the more we succeed in identifying diverse empirical implications of our theories, the less likely it will be that those diverse implications can simply be accumulated as homogeneous observations in a single quantitative model. Having a richly detailed case study touching upon many implications of the same theory or theories is no substitute for “seek[ing] homogenous units across time or across space” (93), as KKV points out in the subsequent discussion of “process tracing” (226–28).

KKV allows that “attaining [causal] homogeneity is often impossible,” but goes on to assert in the next sentence that “understanding the degree of heterogeneity in our units of analysis will help us to estimate the degree of uncertainty or likely biases to be attributed to our inferences” (93–94). How is that? Again, the authors do not explain. But once again, the more important point is that nobody else does either—a point I am compelled to acknowledge despite my own efforts in that direction (Bartels 1996). If we accept KKV’s assertion that the “generally untestable” assumption of causal homogeneity (or the related assumption of “constant causal effects”) “lies at the base of all scientific research” (93), this is a loud and embarrassing silence.

CONCLUSION

In the end, KKV’s optimistic-sounding unification of quantitative and qualitative research seems to me to promise a good deal more than it delivers, and a good deal more than it could possibly deliver given the current state of political methodology in both its qualitative and quantitative modes. But perhaps that is the genius of the book. By presenting a bold and beguiling vision of a seamless, scientific methodology of social inquiry, KKV may successfully challenge all of us to make some serious progress toward implementing that vision.

1. KKV (91) uses the label “unit homogeneity” for this assumption.

5

How Inference in the Social (but Not the Physical) Sciences Neglects Theoretical Anomaly

Ronald Rogowski

Designing Social Inquiry, by King, Keohane, and Verba (hereafter KKV), deserves praise for many reasons. It attempts, seriously and without condescension, to bridge the gap between qualitative and quantitative political science. It reminds a new generation of students, in both traditions, of some main characteristics of good theory (testability, operationalizability, and “leverage” or deductive fertility). It clarifies, even for the profoundly mathematically challenged, some of the central strictures of quantitative inference (why one cannot have more variables than cases or select on the dependent variable, or why it biases results if measurement of the independent variable is faulty). It abounds with practical wisdom on research design, case selection, and complementary methodologies. Perhaps most importantly, it opens a dialogue between previously isolated practitioners of these two forms of analysis and provokes worthwhile discussion.

For all of these reasons and more, the book should be, will be, and—indeed even in its samizdat forms—already has been widely assigned and read. It is, quite simply, the best work of its kind now available; indeed, it is very likely the best yet to have appeared.¹ At the same time, I think, KKV falters in its aim of evangelizing qualitative social scientists; and it does so,

1. The only competition, long out of print and aimed more at the advanced undergraduate level, is probably Lave and March (1975).

paradoxically, because it attends insufficiently to the importance of problemation and deductive theorizing in the scientific enterprise.

PROBLEMATION AND DEDUCTIVE THEORIZING

As natural scientists have long understood (see Hempel 1966), inference proceeds most efficiently by three complementary routes: (1) making clear the essential model, or process, that one hypothesizes to be at work; (2) teasing out the deductive implications of that model, focusing particularly on the implications that seem *a priori* least plausible; and (3) rigorously testing those least plausible implications against empirical reality.² The Nobel physicist and polymath Richard Feynman may have put it best:³

Experimenters search most diligently, and with the greatest effort, in exactly those places where it seems most likely that we can prove our theories wrong. In other words we are trying to prove ourselves wrong as quickly as possible, because only in that way can we find progress. (1965: 158)

The classical example is Einstein's Theory of Relativity, which: (1) uniquely provided an overarching model that could explain both the anomalies and the enduring validities of classical Newtonian mechanics, indeed could subsume it as a special case; (2) had, among its many other implications, a quite specific, rather implausible, and previously untested one about how light reflected from the planet Mercury would be deflected by the sun's gravitation; and (3) appeared at the time to be precisely accurate in this specific and implausible implication.⁴ To test, however rigorously, hypotheses that challenge no deeper theory or that themselves lack deductive implications is an inefficient route of scientific inference, while theories that are precise and deductively fertile enough can often be sustained or refuted by surprisingly unelaborate tests, including ones that involve few observations or that violate normally sacrosanct principles of selection.

KKV, I contend, emphasizes the third part of scientific inquiry, the rigorous testing of hypotheses, almost to the exclusion of the first two—the elab-

2. Eckstein characterized this as the strategy of the "least-likely" case (1975: 118–19). See also Hempel (1966: 37–38).

3. I owe this citation to Mark Lichbach.

4. To quote a famous statement on this prediction in a letter of J. E. Littlewood to Bertrand Russell, written in 1919: "Dear Russell: Einstein's theory is completely confirmed. The predicted displacement was 1".72 and the observed 1".75 + .06. Yours, J. E. L." Quoted in Russell (1969: 149).

oration of precise models and the deduction of their (ideally, many) logical implications—and thus points us to a pure, but needlessly inefficient, path of social-scientific inquiry.

THEORY AND ANOMALY: SOME EXAMPLES

I can best illustrate these points by applying KKV's strictures to some landmark works in comparative politics, often cited as worthy of emulation. Each work, it seems to me, would fail KKV's tests and would be dismissed as insufficiently scientific. Yet in each case, the dismissal would be incorrect: the works illustrate—indeed epitomize—valid and efficient social-scientific inquiry; and the ways in which they do so illuminate the shortcomings in KKV's analysis.

Three of the classical works that I have in mind are single-observation studies; one involves three cases, but all within a single region; one selects chiefly on the independent—but also on the dependent—variable, in ways deprecated by KKV; and one selects on the dependent variable. I propose: (1) to sketch each briefly; (2) to argue that the conventional wisdom is right, and KKV is wrong, with regard to these works' worth; and (3) to reflect on the deficiencies that these works reveal in KKV's analysis.

The single-observation studies are Arend Lijphart's (1975 [1968]) study of the Netherlands, *The Politics of Accommodation*; William Sheridan Allen's single-city examination, *The Nazi Seizure of Power* (1965); and Peter Alexis Gourevitch's 1978 critique of Immanuel Wallerstein's *Modern World-System*. Each involves disconfirmation of a prevailing theory, by what Eckstein called the strategy of the "most likely" case (1975: 119).

Lijphart rightly saw in the Netherlands a serious empirical challenge to David Truman's (1951) then widely accepted theory of "cross-cutting cleavages." Truman had argued, plausibly enough, that mutually reinforcing social cleavages (class coterminous with religious practice, or religion with language) impeded social agreement and made conflict more likely. Only where each deep cleavage was orthogonal to another (e.g., Switzerland, where many Catholics are German-speaking, many Francophones Protestant) was social peace likely to endure. About the Netherlands, however, two things were abundantly clear: (1) it had virtually no cross-cutting cleavages; and (2) it had about as stable and amicable a democracy as one could find. Lijphart's study was taken at the time, I believe correctly, as having refuted Truman's theory.⁵

5. Lijphart went on to conjecture, on the basis of the Dutch case, about the precise circumstances in which non-cross-cutting cleavages were compatible with civic peace; but that is secondary to the point I am arguing here.

In attempting to explain popular support for such totalitarian movements as Fascism, many social scientists had, by the 1950s, accepted a theory whose roots went back to Montesquieu and Tocqueville but whose modern version had been shaped chiefly by Lederer (1940), Arendt (1958), and—the great synthesizer of this genre—Kornhauser (1959). Again simplifying it to the point of caricature, this theory held that societies were opened to totalitarianism's Manichean zealotries by the waning (e.g., through rapid modernization) of associational life—the disappearance of those “natural” groups that afforded meaning, balance, and a sense of efficacy. Totalitarian followers were “atomized” or “mass” individuals.

Tracing the growth of the National Socialist cause in a single midsized German town where it had prospered earlier and better than the average, however, Allen (1965) found, if anything, a superabundance of associational life: singing and shooting societies, card clubs, fraternal orders, religious associations, drinking groups, and *Stammtische* of long standing, to the point that one could hardly imagine a free evening in these proto-Fascists' lives. Neither could he observe any waning of this associational activity before or during the Nazi expansion, nor were Nazis drawn disproportionately from the less active (if anything, the contrary).⁶ Only *after* Hitler came to power, with the Nazi *Gleichschaltung* of all associations, did activity decline. Allen's results were read (again, I think, rightly) as having strongly impugned an otherwise plausible theory.

A central assertion of Immanuel Wallerstein's *Modern World-System*, vol. 1 (1974), was that the “core” states of the world economy, from the sixteenth century onward, had been likeliest to develop strong states (in order to guarantee capitalist property rights and to protect trade routes) and to pursue linguistic and cultural homogeneity (in order to lower administrative and transaction costs). Yet as Gourevitch and others quickly observed, it was, in fact, a central European state of what Wallerstein had called the “semiperiphery” (i.e., Prussia) that developed arguably the strongest state in the early modern world and that came earliest to mass education and the pursuit of linguistic homogeneity (1978: esp. 423–27). The case seriously undermined this aspect of Wallerstein's theory; but Gourevitch went on to speculate—and Charles Tilly (1990) has subsequently advanced considerable argument and evidence to show—that in fact, the correlation was the

6. To be sure, KKV distinguishes between *cases* and *observations*; and Allen's study could be read as a single case that encompasses many observations, given that Allen examines a variety of groups and individuals. Such a reading, in my view, would fundamentally misunderstand the underlying theory, whose central independent variable is the level of association that individuals encounter. Given the theory, the town (or, at most, the class within the town) is the relevant observation; and Allen's study is therefore a single case *and* a single observation.

reverse: The economically most advanced early modern states were often the least powerful, and vice versa.

Against the record amassed by these and other single-observation studies, KKV contends that “[I]n general . . . the single observation is not a useful technique for testing hypotheses or theories” (211), chiefly because measurement error may yield a false negative, omitted variables may yield an unpredicted result, or social-scientific theories are insufficiently precise.⁷ The authors would have us accept that the Lijphart, Allen, and Gourevitch studies—and even more the sweeping inferences that most comparativists drew from them—were bad science; as KKV states explicitly, falsification from a single observation “is not the way social science is or should be conducted” (103).

Rudolf Heberle’s (1963, 1970) justly famous exploration of Nazi support in Schleswig-Holstein is exemplary in doing what KKV calls “making many observations from few” (217); yet Heberle’s research, too, would presumably fail to meet KKV’s standard. Long before Barrington Moore, Jr. (1967) solidified the thesis, analysts had conjectured a close link between labor-repressive agriculture and susceptibility to Fascism. It occurred to Heberle that the north German state of Schleswig-Holstein offered an ideal test of the thesis, containing, as it did, three distinct agricultural regions, characterized respectively by: (1) plantation agriculture on the East Elbian, or the “Junker” model (the Hill district); (2) prosperous family farms like those of western and southwestern Germany (the Marsh); and (3) hardscrabble, quasi-subsistence farming (the *Geest*). The asserted link to feudalism would predict the earliest and strongest Nazi support in the first of these regions; but in fact the Fascist breakthrough occurred in the *Geest*, among the marginalized subsistence farmers; the family farmers came along only considerably later, and the feudal region resisted almost to the end. This brilliantly designed little study thus seriously undermined, even before its precise formulation, what has since come to be known as the “Moore thesis” about the origins of Fascism.

Like Atul Kohli’s (1987) three-state study of poverty policy in India, Heberle’s examination inventively exploits within-country—in Heberle’s case, within-region—variation. Yet KKV dismisses precisely this aspect of Kohli’s analysis, on the ground that the values of both the explanatory and the dependent variables were known in advance; “selection, in effect, is on both the explanatory and dependent variables” so that “the design . . . provides no information about his causal hypothesis” (145). Of course,

7. KKV strictures on the first two points are so sweeping that they must, by implication, include theories and hypotheses in the physical sciences. Hence I take it that KKV would also reject the confirmation of the theory of relativity and other cases alluded to by Hempel (1966: 77), which rested on single observations.

Heberle, by confining his attention to a single state, partially constrained himself against biased selection; but Schleswig-Holstein itself might represent only random variation, and so (KKV would surely say) could not be taken as refuting the hypothesized causal link between feudalism and Fascism. Again, I think, KKV's strictures, taken literally, would dismiss a brilliant study as bad (or at least inadequate) social science.

My final two examples raise the stakes considerably, for they represent, by common consent, the very best of recent work in comparative politics. Yet Peter Katzenstein's *Small States in World Markets* (1985), by KKV's lights, inadmissibly restricts variation on the independent and dependent variables; and Robert Bates's *Markets and States in Tropical Africa* (1981) impermissibly selects on the dependent variable.

Katzenstein, contesting the conventional wisdom that only large states were independent enough to be worth studying, deliberately restricted his focus to the smaller European states and, within that set, to the smaller states that were "close to the apex of the international pyramid of success," thus "excluding Ireland, Finland, and some of the Mediterranean countries" (1985: 21). His reasons were straightforward: (1) the cases that he did study were *anomalous*, for small, price-taking countries were widely supposed to face particular challenges in an uncertain international environment; and (2) they were *forerunners*, in the sense that all countries were rapidly becoming as dependent on international markets as these small ones had long been. To examine why countries that theoretically should not succeed in fact did so (reminiscent of Lijphart's strategy) and to attempt to discern a possible path of adaptation of larger states, seemed, both to Katzenstein and to his generally enthusiastic readership, a sensible strategy. Yet KKV, at least as I read the book, must hold Katzenstein guilty of two cardinal sins that largely vitiate his analysis: (1) instead of choosing his cases to guarantee some range of variation on the independent variable, he restricts his analysis to small (and therefore quite trade-dependent) states; and (2) more seriously, taking economic success or failure as his dependent variable, he looks only at instances of success.

Bates's book is an even clearer case of selection on the dependent variable. Exactly as Michael Porter's *Competitive Advantage of Nations* (1990) examines only cases of economic success and thus draws withering fire from KKV (133–34), Bates focuses almost entirely on cases of economic failure or, more precisely, on the remarkably uniform *pattern* of economic failure among the states of postindependence Africa. He nonetheless develops an account that most readers have found compelling: (1) that the failures all resulted from an economic policy that heavily taxed agricultural exports to subsidize investment in heavily protected manufactures; and (2) that this self-destructive economic policy was the inevitable result of a political constellation in which urban groups were organized and powerful,

rural ones scattered and weak. While Bates supports his analysis by observing that the two African cases of relative economic success (i.e., Kenya and Côte d'Ivoire) were characterized by export-friendly policy and politically more powerful farmers, this part of his discussion is brief and clearly tangential to his main argument.

Why, despite their seemingly egregious sins,⁸ are all of these works believed by most comparativists—rightly, in my judgment—to have provided convincing inferences about their topics of study? Chiefly, I submit, for two reasons, which shed much light on the problems of KKV's account: (1) all of them tested, relied on, or proposed, clear and precise *theories*; and (2) all focused on *anomalies*, either in prevailing theories or in the world—cases that contradicted received beliefs or unexpected regularities that were too pronounced to be accidental.

The theories of cross-cutting cleavages (Truman 1951), atomization (e.g., Kornhauser 1959), world-systems (Wallerstein 1974), and feudal legacy (Moore 1967) had the great advantage of being precise enough to yield implications for single, or for very few, observations. Lijphart, Allen, Gourevitch, and Heberle, respectively, took brilliant scholarly advantage of that precision: (1) to seek out anomalous cases and, usually, (2) to conjecture intelligently about a more satisfactory general theory that could avoid such anomalies.

About small states and heavy reliance on external markets there was less a prevailing theory than a prevailing prejudice—that puniness entailed constraint, insecurity, and (barring extraordinary good luck) economic trouble. By adducing seven cases of small states that had consistently prospered, Katzenstein demonstrated that insecurity and poverty were far from inevitable; by showing that their strategies, in similar circumstances, had differed, he proved that they retained considerable freedom of policy; and by analyzing their marked similarities of historical development and present-day governance, he advanced a plausible (if in this work still conjectural)⁹ theory of situational requisites for highly trade-dependent states.

The African economic devastation that Bates studied was usually “explained” by a *mélange* of misunderstood Marxism and economic illiteracy that stressed the “dependence” of the third world on the first. By invoking standard, simple economics, Bates easily showed that local policy, and

8. As regards selection on the dependent variable, KKV takes a particularly draconian stand: “We can . . . learn nothing about a causal effect from a study which selects observations so that the dependent variable does not vary” (147).

9. To be sure, by looking only at successful small European states, Katzenstein had to leave open the possibilities (1) that unsuccessful small states were also governed corporatively; and (2) that small non-European states had discovered quite different recipes for success.

not first-world plots, must be to blame. If domestic agricultural prices were systematically suppressed, one would expect to see smuggling and rural flight; if domestic industry was protected and subsidized, one would expect cartels, uncompetitive goods, and an overvalued currency; if taxes and controls poured power and resources into the hands of bureaucrats, one would anticipate a bloated public sector and vicious competition for place and favor. In each African case, all of these in fact prevailed, and no amount of external "dependence" could so easily explain this particular concatenation of disasters.

Yet this left a riddle no less profound than the original one: Why should almost all governments of the region have deliberately chosen policies so inimical to aggregate welfare and to long-term growth? Just as a psychologist might become intrigued if all but one or two of the people on a certain street began suddenly to mutilate themselves, Bates pursued a "cluster analysis" (see KKV 148–49) of perverse African policies and reached his highly plausible conjecture that rural weakness produced a fatal "urban bias" (see Lipton 1976) in policy.¹⁰

In the works of Katzenstein and Bates, then, no less than in those cited earlier, the crucial ingredient was clear, precise, powerful ("high leverage") theory with what Lave and March (1975) tellingly called a "sense of process," that is, intuitively plausible causal links. In both accounts, universally accepted economic theory underpinned the critique of received wisdom: if small, price-taking firms survived in uncertain markets, why not small, price-taking countries; if all of the symptoms of the African cases were consistent with systematic price distortions, what other diagnosis was possible? The core of Katzenstein's alternative account was a story about how democratic corporatism facilitated flexible adjustment to external markets; the core of Bates's account, a hypothesized link between power and policy. That both arguments were so clear, plausible, and precise contributed crucially to their persuasiveness.

LESSONS

KKV (127), in contrast, frequently chooses as examples hypotheses that seem obvious or that lack deductive fertility. To prove, for example, that declining Communist societies were more likely to spawn mass movements

10. It is worth noting that Bates has pursued this conjecture *not* through any large-N study, but by close analysis of an apparently anomalous case: Colombia, where dispersed coffee farmers of modest means prevailed politically not only against city dwellers but over concentrated plantation owners of considerable wealth.

of opposition the less repressive the old regime was neither contravenes received wisdom nor carries broader implications for other cases.

The aspects of larger theory and of "sense of process," consequently, seem to be sorely absent from KKV's prescriptions for social inquiry. While the authors are right to fear our natural tendency to see patterns where none exist (21), they emphasize insufficiently the centrality of patterns—indeed, of "paradigms" (Kuhn 1962)—to efficient scientific inquiry. A powerful, deductive, internally consistent theory can be seriously undermined, at least in comparative politics, by even one wildly discordant observation (Lijphart's Netherlands). On the positive side, a powerful theory can, by explaining an otherwise mysterious empirical regularity (European small-state corporatism, African economic failure), gain provisional acceptance at least as a highly plausible conjecture worthy of further research. As most discussions of spurious correlation make clear, we gain confidence in a proposed explanation to the extent that it *both* (1) fits the data *and* (2) "makes sense" in terms of its consistency with other observations and its own deductive implications. KKV, it seems to me, emphasizes the former at the expense of the latter. In consequence, its advice to area specialists focuses almost entirely on "increasing the number of observations" (chap. 6). Many comparativists, I think, would instead counsel: "Choose better theory, which can make better use of few or single observations."¹¹

Valuable as KKV's strictures are, I fear that devout attention to them may paralyze, rather than stimulate, scientific inquiry in comparative politics. The authors write eloquently and insightfully about the trade-offs between close observation of a few cases and more cursory measurement of many (chap. 2, esp. 66–68); I wish they had as perceptively discussed how better theory permits inference from fewer cases, allows restriction on the independent variable, and may even profit from judicious selection on the dependent variable.

In short, I suspect KKV does not mean quite as stern a message as it sends; or perhaps the authors view the studies I have discussed here in a different and more redeeming light. However, the book would have spoken more clearly to comparativists if it had specifically addressed the major literature of the less quantitative tradition.

11. As I note at the outset, KKV does discuss—at some length and quite sensibly—some major characteristics of good theory (section 3.5). The authors seem, however, to despair that social-scientific theories can ever be precise enough to permit valid inference from few cases (210–11); and they explicitly reject parsimony as an inherently desirable property of social-scientific theory (20, 104–5). On neither point, I suspect, will most comparativists find their arguments persuasive; and they seem to me to be refuted by the examples I adduce here.

C. LINKING THE QUANTITATIVE AND QUALITATIVE TRADITIONS

6

Bridging the Quantitative- Qualitative Divide

Sidney Tarrow

In *Designing Social Inquiry* (hereafter KKV), Gary King, Robert O. Keohane, and Sidney Verba have performed a real service to qualitative researchers. I, for one, will not complain if I never again have to look into the uncomprehending eyes of first-year graduate students when I enjoin them—in deference to Przeworski and Teune—to “turn proper names into variables.” The book is brief and lucidly argued and avoids the weighty, muscle-bound pronouncements that are often studded onto the pages of methodological manuals.

But following KKV’s injunction that “a slightly more complicated theory will explain vastly more of the world” (105), I will praise the book no more, but focus on an important weakness in the book: KKV’s central argument is that the same logic that is “explicated and formalized clearly in discussions of quantitative research methods” underlies—or should—the best qualitative research (3). If this is so, then the authors really ought to have paid more attention to the *relations* between quantitative and qualitative approaches and what a rigorous use of the latter can offer quantifiers. While they offer a good deal of generous (if at times patronizing) advice to qualitatively oriented scholars, they say very little about how qualitative approaches can be combined with quantitative research. Especially with the growth of choice-theoretic approaches, whose practitioners often illustrate their theories with narrative, there is a need for a set of ground rules on how to make intelligent use of qualitative data.

KKV does not address this issue. Rather, it uses the model of quantitative

research to advise qualitative researchers on how best to approximate good models of descriptive and causal inference. (Increasing the number of observations is its cardinal operational rule.) But in today's social science world, how many social scientists can simply be labeled "qualitative" or "quantitative"? How often, for example, do we find support for sophisticated game-theoretic models resting on the use of anecdotal reports or on secondary evidence lifted from one or two qualitative sources? More and more frequently in today's social science practice, quantitative and qualitative data are interlarded within the same study. In what follows, I will discuss some of the problems of combining qualitative and quantitative data, as well as some solutions to these problems.

CHALLENGES OF COMBINING QUALITATIVE AND QUANTITATIVE DATA

A recent work that KKV warmly praises illustrates both that its distinction between quantitative and qualitative researchers is too schematic and that we need to think more seriously about the interaction of the two kinds of data. In Robert Putnam's (1993) analysis of Italy's creation of a regional layer of government, *Making Democracy Work*, countless elite and mass surveys and ingenious quantitative measures of regional performance are arrayed for a twenty-year period of regional development. On top of this, he conducted detailed case studies of the politics of six Italian regions, gaining, in the process, what KKV (quoting Putnam) recommends as "an intimate knowledge of the internal political maneuvering and personalities that have animated regional politics over the last two decades" (5) and what Putnam calls "marinating yourself in the data" (KKV: 5; Putnam 1993: 190). KKV (38) uses *Making Democracy Work* to praise the virtues of "soaking and poking," in the best Fenno (1977: 884) tradition.

But Putnam's debt to qualitative approaches is much deeper and more problematic than this; after spending two decades administering surveys to elites and citizens in the best Michigan mode, he was left with the task of explaining the sources of the vast differences he had found between Italy's northcentral and southern regions. In his effort to find them, his quantitative evidence offered only indirect help, and he turned to history, repairing to the halls of Oxford, where he delved deep into the Italian past to fashion a provocative interpretation of the superior performance of northern Italian regional governments vis-à-vis southern ones. This he based on the civic traditions of the (northern) Renaissance city-states, which, according to him, provided "social capital" that is lacking in the traditions of the South (chap. 5). A turn to qualitative history—probably not even in Putnam's

mind when he designed the project—was used to interpret cross-sectional, contemporary quantitative findings.

Putnam's procedure in *Making Democracy Work* pinpoints a question in melding quantitative and qualitative approaches that KKV's canons of good scientific practice do not help to resolve. In delving into the qualitative data of history to explain our quantitative findings, by what rules can we choose the *period* of history that is most relevant to our problem? What *kind* of history are we to use; the traditional history of kings and communes or the history of the everyday culture of the little people? And how can the effect of a particular historical period be separated from that of the periods that precede or follow it? In the case of *Making Democracy Work*, for example, it would have been interesting to know by what rules of inference Putnam chose the Renaissance as determining the Italian North's late twentieth-century civic superiority. Why not look to its sixteenth-century collapse faced by more robust monarchies, its nineteenth-century military conquest of the South, or its 1919–21 generation of Fascism (not to mention its 1980s corruption-fed pattern of economic growth)? None of these are exactly "civic" phenomena; by what rules of evidence are they less relevant in "explaining" the northern regions' civic superiority over the South than the period of the Renaissance city-states? Putnam doesn't tell us; nor does KKV.

To generalize from the problem of Putnam's book, qualitative researchers have much to learn from the model of quantitative research. But quantitative cousins who wish to profit from conjoining their findings with qualitative sources need, for the selection of qualitative data and the intersection of the two types, rules just as demanding as the rules put forward by KKV for qualitative research on its own. I shall sketch some useful tools for bridging the quantitative-qualitative divide from recent examples of comparative and international research (see table 6.1).

TOOLS FOR BRIDGING THE DIVIDE

Tracing Processes to Interpret Decisions

One such tool that KKV cites favorably is the practice of *process tracing* in which "the researcher looks closely at 'the decision process by which various initial conditions are translated into outcomes'" (226; quoting George and McKeown 1985: 35). KKV interprets the advantages of process tracing narrowly, assimilating it to their favorite goal of increasing the number of theoretically relevant observations (227). As George and McKeown actually conceived it, the goal of process tracing was not to increase the number of discrete decision stages and aggregate them into a larger number of data points but to *connect* the phases of the policy process and enable the investi-

gator to identify the reasons for the emergence of a particular decision through the dynamic of events (George and McKeown 1985: 34–41).

Process tracing is different *in kind* from observation accumulation and is best employed in conjunction with it—as was the case, for example, in the study of cooperation on economic sanctions by Lisa Martin (1992) that KKV cites so favorably.

Systematic and Nonsystematic Variable Discrimination

KKV gives us a second example of the uses of qualitative data but, once again, underestimates its particularity. The authors argue that the variance between different phenomena “can be conceptualized as arising from two separate elements: *systematic* and *nonsystematic* differences,” the former more relevant to fashioning generalizations than the latter (56). For example, in the case of Conservative voting in Britain, systematic differences include such factors as the properties of the district, while unsystematic differences could include the weather or a flu epidemic at the time of the election. “Had the 1979 British elections occurred during a flu epidemic that swept through working-class houses but tended to spare the rich,” the

Table 6.1. Tools for Bridging the Qualitative-Quantitative Divide

<i>Tool</i>	<i>Contribution to Bridging the Divide</i>
Process Tracing	Qualitative analysis focused on processes of change within cases may uncover the causal mechanisms that underlie quantitative findings.
Focus on Tipping Points	Qualitative analysis can explain turning points in quantitative time series and changes over time in causal patterns established with quantitative data.
Typicality of Qualitative Inferences Established by Quantitative Comparison	Close qualitative analysis of a given set of cases provides leverage for causal inference, and quantitative analysis then serves to establish the representativeness of these cases.
Quantitative Data as Point of Departure for Qualitative Research	A quantitative data set serves as the starting point for framing a study that is primarily qualitative.
Sequencing of Qualitative and Quantitative Studies	Across multiple research projects in a given literature, researchers move between qualitative and quantitative analysis, retesting and expanding on previous findings.
Triangulation	Within a single research project, the combination of qualitative and quantitative data increases inferential leverage.

authors conclude, "our observations might be rather poor measures of underlying Conservative strength" (56–57).

Right they are, but this piece of folk wisdom hardly exhausts the importance of nonsystematic variables in the interpretation of quantitative data. A good example comes from how the meaning and extension of the strike changed as systems of institutionalized industrial relations developed in the nineteenth century. At its origins, the strike was spontaneous, uninstitutionalized and often accompanied by whole-community "turnouts." As unions developed and governments recognized workers' rights, the strike broadened to whole sectors of industry, became an institutional accompaniment to industrial relations, and lost its link to community collective action. The systematic result of this change was permanently to affect the patterns of strike activity. Quantitative researchers like Michelle Perrot (1986) documented this change. But had she regarded it only as a case of "nonsystematic variance" and discarded it from her model, as KKV proposes, Perrot might well have misinterpreted the changes in the form and incidence of the strike rate. Because she was as good a historian as she was a social scientist, she retained it as a crucial change that transformed the relations between strike incidence and industrial relations.

To put this point more abstractly, distinct historical events often serve as the tipping points that explain the shifts in an interrupted time-series, permanently affecting the relations between the variables (Griffin 1992). Qualitative research that turns up "nonsystematic variables" is often the best way to uncover such tipping points. Quantitative research can then be reorganized around the shifts in variable interaction that such tipping points signal. In other words, the function of qualitative research is not only, as KKV seems to argue, to peel away layers of unsystematic fluff from the hard core of systematic variables; but also to assist researchers in understanding shifts in the values of the systematic variables.

Framing Qualitative Research within Quantitative Profiles

The uses of qualitative data described in the two previous sections pertain largely to aiding quantitative research. But this is not the only way in which social scientists can combine quantitative and qualitative approaches. Another is to focus on the qualitative data, using a systematic quantitative database as a frame within which the qualitative analysis is carried out. Case studies have been validly criticized as often being based on dramatic but frequently unrepresentative cases. Studies of successful social revolutions often focus on characteristics that may also be present in unsuccessful revolutions, rebellions, riots, and ordinary cycles of protest (Tilly 1993: 12–14). In the absence of an adequate sample of revolutionary episodes, no

one can ascribe particular characteristics to a particular class of collective action.

The representativeness of qualitative research can never be wholly assured until the cases become so numerous that the analysis comes to resemble quantitative research (at which point the qualitative research risks losing its particular properties of depth, richness, and process tracing). But framing it within a quantitative database makes it possible to avoid generalizing on the occasional “great event” and points to less dramatic—but cumulative—historical trends.

Scholars working in the “collective action event history” tradition have used this double strategy with success. For example, in his 1993 study of over 700 revolutionary events in over 500 years of European history, Charles Tilly assembled data that could have allowed him to engage in a large-N study of the correlates and causes of revolution. Tilly knows how to handle large time-series data sets as well as anybody. However, he did not believe the concept of *revolution* had the monolithic quality that other social scientists had assigned to it (1993: chap. 1). Therefore, he resisted the temptation for quantification, using his database, instead, to frame a series of regional time-series narratives that depended as much on his knowledge of European history as on the data themselves. When a problem cried out for systematic quantitative analysis (e.g., when it came to periodizing nationalism), Tilly (1994) was happy to exploit the quantitative potential of the data. But the quantitative data served mainly as a frame for qualitative analysis of representative regional and temporal revolutionary episodes and series of episodes.

Putting Qualitative Flesh on Quantitative Bones

An American sociologist, Doug McAdam, has shown how social science can be enriched by carrying out a sustained qualitative analysis of what is initially a quantitative database. McAdam’s 1988 study of Mississippi Freedom Summer participants was based on a treasure-trove of quantifiable data—the original questionnaires of the prospective Freedom Summer volunteers. While some of these young people eventually stayed home, others went south to register voters, teach in “freedom schools,” and risk the dangers of Ku Klux Klan violence. Two decades later, both the volunteers and the no-shows could be interviewed by a researcher with the energy and the imagination to go beyond the use of canned data banks.

McAdam’s main analytic strategy was to carry out a paired comparison between the questionnaires of the participants and the stay-at-homes and to interview a sample of the former in their current lives. This systematic comparison formed the analytical spine of the study and of a series of technical papers. Except for a table or two in each chapter, the texture of *Freedom*

Summer is overwhelmingly qualitative. McAdam draws on his interviews with former participants, as well as on secondary analysis of other people's work, to get inside the Freedom Summer experience and to highlight the effects that participation had on their careers and ideologies and their lives since 1964. With this combination of quantitative and qualitative approaches, he was able to tease a convincing picture of the effects of Freedom Summer activism from his data.

As I write this, I imagine KKV exclaiming, "But this is *precisely* the direction we would like to see qualitative research moving—toward expanding the number of observations and re-specifying hypotheses to allow them to be tested on different units!" (see chap. 7). But would they argue, as I do, that it is the *combination* of quantitative and qualitative methods trained on the same problem (not a move toward the logic of quantitative analysis alone) that is desirable? Two more ways of combining these two logics illustrate my intent.

Sequencing Quantitative and Qualitative Research

The growth industry of qualitative case studies that followed the 1980–81 Solidarity movement in Poland largely took as given the idea that Polish intellectuals had the most important responsibility for the birth and ideology of this popular movement. There was scattered evidence for this propulsive role of the intellectuals; but since most of the books that appeared after the events were written by them or by their foreign friends, an observer bias might have been operating to inflate their importance in the movement vis-à-vis the workers who were at the heart of collective action in 1980–81 and whose voice was less articulate.

Solid quantitative evidence came to the rescue. In a sharp attack on the "intellectualist" interpretation and backed by quantitative evidence from the strike demands of the workers themselves, Roman Laba demonstrated that their demands were overwhelmingly oriented toward trade union issues, and showed little or no effect of the proselytizing that Polish intellectuals had supposedly been doing among the workers of the Baltic coast since 1970 (1991: chap. 8). This finding dovetailed with Laba's own qualitative analysis of the development of the workers' movement in the 1970s and downplayed the role of the Warsaw intellectuals, which had been emphasized in a series of books by their foreign friends.

The response of those who had formulated the intellectualist interpretation of Solidarity was predictably indignant. But there were also more measured responses that shed new light on the issue. For example, prodded by Laba's empirical evidence of worker self-socialization, Jan Kubik returned to the issue with both a sharper analytical focus and better qualitative evidence than the earlier intellectualist theorists had employed, criticizing

Laba's conceptualization of class and reinterpreting the creation of Solidarity as "a multistranded and complicated social entity . . . created by the contributions of various people" whose role and importance he proceeded to demonstrate (1994: 230–38). Moral: a sequence of contributions using different kinds of evidence led to a clearer and more nuanced understanding of the role of different social formations in the world's first successful confrontation with state socialism.

Triangulation

I have left for last the research strategy that I think best embodies the strategy of combining quantitative and qualitative methods—the *triangulation* of different methods on the same problem. Triangulation is particularly appropriate in cases in which quantitative data are partial and qualitative investigation is obstructed by political conditions. For example, Valerie Bunce used both case methodology and quantitative analysis to examine the policy effects of leadership rotation in western and socialist systems. In her *Do New Leaders Make a Difference?* she wrote: "I decided against selecting one of these approaches to the neglect of the other [the better] to test the impact of succession on public policy by employing *both* methodologies" (1981: 39).

Triangulation is also appropriate in specifying hypotheses in different ways. Consider the classical Tocquevillian insight that regimes are most susceptible to a political opportunity structure that is partially open. The hypothesis takes shape in two complementary ways: (1) that liberalizing regimes are more susceptible to opposition than either illiberal or liberal ones; and (2) that within the same constellation of political units, opposition is greatest at intermediate levels of political opportunity. Since there is no particular advantage in testing one version of the hypothesis over the other, testing both is optimal (as can be seen in the recent social movement study, Kriesi et al. 1995).

My final example of triangulation comes, with apologies, from my own research on collective action and social movements in Italy. In the course of a qualitative reconstruction of a left-wing Catholic "base community" that was active in a popular district of Florence in 1968, I found evidence that linked this movement discursively to the larger cycle of student and worker protest going on in Italy at the same time (Tarrow 1988). Between 1965 and 1968, its members had been politically passive, focusing mainly on neighborhood and educational issues. However, as the worker and student mobilization exploded around it in 1968, their actions became more confrontational, organized around the themes of autonomy and internal democracy that were animating the larger worker and student movements around them.

Researchers convinced of their ability to understand political behavior by interpreting "discourse" might have been satisfied with these observations; but I was not. If nothing else, Florence was only one case among potential thousands. And in today's global society, finding thematic similarity among different movements is no proof of direct diffusion, since many movements around the world select from the same stock of images and frames without the least connection among them (Tarrow 1994: chap. 11).

As it happened, quantitative analysis came to the rescue by triangulating on the same problem. For a larger study, I had gathered a large sample of national collective action events for a period that bridged the 1968 Florentine episode. And as it also happened, two Italian researchers had collected reliable data on the total number of religious "base communities" like that in Florence throughout the country (Sciubba and Pace 1976). By reoperationalizing the hypothesis cross-sectionally, I was able to show a reasonably high positive correlation (.426) between the presence of Catholic base communities in various cities and the magnitude of general collective action in each city (Tarrow 1989: 200). Triangulation demonstrated that the findings of my longitudinal, local, and qualitative case study coincided with the results of cross-sectional, national, and quantitative correlations. My inductive hunch that Italy in the 1960s underwent an integrated cycle of protest became a more strongly supported hypothesis.

KKV does not take the position that quantification is the answer to all the problems of social science research. But the book's single-minded focus on the logic of quantitative research (and of a certain *kind* of quantitative research) leaves underspecified the particular contributions that qualitative approaches make to scientific research, especially when combined with quantitative research. As quantitatively trained researchers shift to choice-theoretic models backed up by illustrative examples (often containing variables with different implicit metrics) the role of qualitative research grows more important. We are no longer at the stage when public choice theorists can get away with demonstrating a theorem with an imaginary aphorism. We need to develop rules for a more systematic use of qualitative evidence in scientific research. Merely wishing that it would behave as a slightly less crisp version of quantitative research will not solve the problem.

This is no plea for the veneration of historical uniqueness and no argument for the precedence of "interpretation" over inference. (For an excellent analysis of the first problem, see KKV 42–43; and of the second, see KKV 36–41.) My argument, rather, is that a single-minded adherence to *either* quantitative or qualitative approaches straightjackets scientific progress. Whenever possible, we should use qualitative data to interpret quantitative findings, to get inside the processes underlying decision outcomes, and to investigate the reasons for the tipping points in historical time-series. We should also try to use different kinds of evidence together and in

sequence and look for ways of triangulating different measures on the same research problem.

CONCLUSION

KKV gives us a spirited, lucid, and well-balanced primer for training our students in the essential unity of social science work. Faced by the clouds of philosophical relativism and empirical nominalism that have recently blown onto the field of social science, we should be grateful to its authors. But the book's theoretical effort is marred by the narrowness of its empirical specification of qualitative research and by its lack of attention to the qualitative needs of quantitative social scientists. I am convinced that had a final chapter on combining quantitative and qualitative approaches been written by these authors, its spirit would not have been wildly at variance with what I argue here.

7

The Importance of Research Design

Gary King, Robert O. Keohane, and Sidney Verba

Receiving five serious reviews in this symposium¹ is gratifying and confirms our belief that research design should be a priority for our discipline. We are pleased that our five distinguished reviewers appear to agree with our unified approach to the logic of inference in the social sciences, and with our fundamental point: that good quantitative and good qualitative research designs are based fundamentally on the same logic of inference. The reviewers raise virtually no objections to the main practical contribution of our book—our many specific procedures for avoiding bias, getting the most out of qualitative data, and making reliable inferences.

However, the reviews make clear that although our book may be the latest word on research design in political science, it is surely not the last. We are taxed for failing to include important issues in our analysis and for dealing inadequately with some of what we included. Before responding to the

1. Editors' note: This chapter is reprinted from the 1995 symposium on *Designing Social Inquiry*, published in the *American Political Science Review*. In this chapter, the authors respond to arguments developed in three additional articles in the *APSR* symposium that are reprinted in the present volume: those by Rogowski, Tarrow, and (reprinted in part) Collier. King, Keohane, and Verba likewise respond here to the two other articles in the symposium—by Laitin (1995) and Caporaso (1995)—to which reference is made in the present volume, but which are not included here. The full original citation for this chapter is Gary King, Robert O. Keohane, and Sidney Verba (1995) "The Importance of Research Design in Political Science." *American Political Science Review* 89, no. 2 (June): 475–81. The table of contents, preface, and chapter 1 of *Designing Social Inquiry* are available at pup.princeton.edu/titles/5458.html.

reviewers' most direct criticisms, let us explain what we emphasize in *Designing Social Inquiry* and how it relates to some of the points raised by the reviewers.

WHAT WE TRIED TO DO

Designing Social Inquiry grew out of our discussions while coteaching a graduate seminar on research design, reflecting on job talks in our department, and reading the professional literature in our respective subfields. Although many of the students, job candidates, and authors were highly sophisticated qualitative and quantitative data collectors, interviewers, soakers and pokers, theorists, philosophers, formal modelers, and advanced statistical analysts, many nevertheless had trouble defining a research question and designing the empirical research to answer it. The students proposed impossible fieldwork to answer unanswerable questions. Even many active scholars had difficulty with the basic questions: What do you want to find out? How are you going to find it out? And above all, how would you know if you were right or wrong?

We found conventional statistical training to be only marginally relevant to those with qualitative data. We even found it inadequate for students with projects amenable to quantitative analysis, since social science statistics texts do not frequently focus on research *design* in observational settings. With a few important exceptions, the scholarly literatures in quantitative political methodology and other social science statistics fields treat existing data and their problems as given. As a result, these literatures largely ignore research design and, instead, focus on making valid inferences through statistical corrections to data problems. This approach has led to some dramatic progress; but it slights the advantage of improving research design to produce better data in the first place, which almost always improves inferences more than the necessarily after-the-fact statistical solutions.

This lack of focus on research design in social science statistics is as surprising as it is disappointing, since some of the most historically important works in the more general field of statistics are devoted to problems of research design (see, e.g., Fisher 1935, *The Design of Experiments*). Experiments in the social sciences are relatively uncommon, but we can still have an enormous effect on the value of our qualitative or quantitative information, even without statistical corrections, by improving the design of our research. We hope our book will help move these fields toward studying innovations in research design.

We culled much useful information from the social science statistics literatures and qualitative methods fields. But for our goal of explicating and

unifying the logic of inference, both literatures had problems. Social science statistics focuses too little on research design, and its language seems arcane if not impenetrable. The numerous languages used to describe methods in qualitative research are diverse, inconsistent in jargon and methodological advice, and not always helpful to researchers. We agree with David Collier that aspects of our advice can be rephrased into some of the languages used in the qualitative methods literature or that used by quantitative researchers. We hope our unified logic and, as David Laitin puts it, our “common vocabulary” will help foster communication about these important issues among all social scientists. But we believe that any coherent language could be used to convey the same ideas.

We demonstrated that “the differences between the quantitative and qualitative traditions are only stylistic and are methodologically and substantively unimportant” (KKV 4). Indeed, much of the best social science research can combine quantitative and qualitative data, precisely because there is no contradiction between the fundamental processes of inference involved in each. Sidney Tarrow asks whether we agree that “it is the *combination* of quantitative and qualitative” approaches that we desire (95 this volume). We do. But to combine both types of data sources productively, researchers need to understand the fundamental logic of inference and the more specific rules and procedures that follow from an explication of this logic.

Social science, both quantitative and qualitative, seeks to develop and evaluate theories. Our concern is less with the development of theory than *theory evaluation*—how to use the hard facts of empirical reality to form scientific opinions about the theories and generalizations that are the hoped-for outcome of our efforts. Our social scientist uses theory to generate *observable implications*, then systematically applies publicly known procedures to infer from evidence whether what the theory implied is correct. Some theories emerge from detailed observation, but they should be evaluated with new observations, preferably ones that had not been gathered when the theories were being formulated. Our logic of theory evaluation stresses maximizing leverage—explaining as much as possible with as little as possible. It also stresses minimizing bias. Lastly, though it cannot eliminate uncertainty, it encourages researchers to report estimates of the uncertainty of their conclusions.

Theory and empirical work, from this perspective, cannot productively exist in isolation. We believe that it should become standard practice to demand clear implications of theory and observations checking those implications derived through a method that minimizes bias. We hope that *Designing Social Inquiry* helps to “discipline political science” in this way, as David Laitin recommends; and we hope, along with James Caporaso, that “improvements in measurement accuracy, theoretical specification, and

research should yield a smaller range of allowable outcomes consistent with the predictions made" (1995: 459).

Our book also contains much specific advice, some of it new and some at least freshly stated. We explain how to distinguish systematic from non-systematic components of phenomena under study and focus explicitly on trade-offs that may exist between the goals of unbiasedness and efficiency (KKV chap. 2). We discuss causality in relation to counterfactual analysis and what Paul Holland (1986) calls the "fundamental problem of causal inference" and consider possible complications introduced by thinking about causal mechanisms and multiple causality (KKV chap. 3). Our discussion of counterfactual reasoning is, we believe, consistent with Donald Campbell's "quasi-experimental" emphasis (Campbell and Stanley 1963); and we thank James Caporaso for clarifying this.²

We pay special attention in chapter 4 to issues of what to observe: how to avoid confusion about what constitutes a "case" and, especially, how to avoid or limit selection bias. We show that selection on values of explanatory variables does not introduce bias but that selection on values of dependent variables does so; and we offer advice to researchers who cannot avoid selecting on dependent variables.

We go on in chapter 5 to show that while random measurement error in dependent variables does not bias causal inferences (although it does reduce efficiency), measurement error in explanatory variables biases results in predictable ways. We also develop procedures for correcting these biases even when measurement error is unavoidable. In that same chapter, we undertake a sustained analysis of endogeneity (i.e., when a designated "dependent variable" turns out to be causing what you thought was your "explanatory variable") and omitted variable bias, as well as how to control research situations so as to mitigate these problems. In the final chapter, we specify ways to increase the information in qualitative studies that can be

2. To clarify further, we note that the definition of an "experiment" is investigator control over the assignment of values of explanatory variables to subjects. Caporaso emphasizes also the value of random assignment, which is desirable in some situations (but not in others, see KKV 124–28) and sometimes achievable in experiments. (Random selection and a large number of units are also desirable and also necessary for relatively automatic unbiased inferences, but experimenters are rarely able to accomplish either.) A "quasi-experiment" is an observational study with an exogenous explanatory variable that the investigator does not control. Thus, it is not an experiment. Campbell's choice of the word "quasi-experiment" reflected his insight that observational studies follow the same logic of inference as experiments. Thus, we obviously agree with Campbell's and Caporaso's emphases and ideas and only pointed out that the word "quasi-experiment" adds another word to our lexicon with no *additional* content. It is a fine idea, much of which we have adopted; but it is an unnecessary category.

used to evaluate theories; we show how this can be accomplished without returning to the field for additional data collection. Throughout the book, we illustrate our propositions not only with hypothetical examples but with reference to some of the best contemporary research in political science.

This statement of our purposes and fundamental arguments should put some of the reviewers' complaints about omissions into context. Our book is about doing empirical research designed to evaluate theories and learn about the world—to make inferences—not about generating theories to evaluate. We believe that researchers who understand how to evaluate a theory will generate better theories—theories that are not only more internally consistent but that also have more observable implications (are more at risk of being wrong) and are more consistent with prior evidence. If, as Laitin suggests, our single-mindedness in driving home this argument led us implicitly to downgrade the importance of such matters as concept formation and theory creation in political science, this was not our intention.

Designing Social Inquiry repeatedly emphasizes the attributes of good theory. How else to avoid omitted variable bias, choose causal effects to estimate, or derive observable implications? We did not offer much advice about what is often called the "irrational nature of discovery," and we leave it to individual researchers to decide what theories they feel are worth evaluating. We do set forth some criteria for choosing theories to evaluate—in terms of their importance to social science and to the real world—but our methodological advice about research design applies to any type of theory. We come neither to praise nor to bury rational-choice theory, nor to make an argument in favor of deductive over inductive theory. All we ask is that whatever theory is chosen be evaluated by the same standards of inference. Ronald Rogowski's favorite physicist, Richard Feynman, explains clearly how to evaluate a theory (which he refers to as a "guess"): "If it disagrees with [the empirical evidence], it is wrong. In that simple statement is the key to science. It does not make any difference how beautiful your guess is. It does not make any difference how smart you are, who made the guess, or what his name is—if it disagrees with [the empirical evidence] it is wrong. That is all there is to it" (1965: 156).³

One last point about our goal: we want to set a high standard for research but not an impossible one. All interesting qualitative and quantitative research yields uncertain conclusions. We think that this fact ought not to

3. Telling researchers to "choose better theories" is not much different than telling them to choose the right answer: it is correct but not helpful. Many believe that deriving rules for theory creation is impossible (e.g., Popper, Feynman), but we see no compelling justification for this absolutist claim. As David Laitin correctly emphasizes, "the development of formal criteria for such an endeavor is consistent with the authors' goals."

be dispiriting to researchers but should rather caution us to be aware of this uncertainty, remind us to make the best use of data possible, and energize us to continue the struggle to improve our stock of valid inferences about the political world. We show that uncertain inferences are every bit as scientific as more certain ones so long as they are accompanied by honest statements of the degree of uncertainty entailed in each conclusion.

OUR ALLEGED ERRORS OF OMISSION

The major theme of what may seem to be the most serious criticism offered above is stated forcefully by Ronald Rogowski. He fears that “devout attention” to our criteria would “paralyze, rather than stimulate, scientific inquiry.” One of Rogowski’s arguments, echoed by Laitin, is that we are too obsessed with increasing the amount of information we can bring to bear on a theory and therefore fail to understand the value of case studies. The other major argument, made by both Rogowski and Collier, is that we are too critical of the practice of selecting observations according to values of the dependent variable and that we would thereby denigrate major work that engages in this practice. We consider these arguments in turn.

Science as a Collective Enterprise

Rogowski argues that we would reject several classic case studies in comparative politics. We think he misunderstands these studies and misses our distinction between a “single case” and a collection of observations. Consider two works that he mentions, *The Politics of Accommodation*, by Arend Lijphart (1975 [1968]), and *The Nazi Seizure of Power*, by William Sheridan Allen (1965). Good research designs are rarely executed by individual scholars isolated from prior researchers. As we say in our book, “A single observation can be useful for evaluating causal explanations if it is part of a research program. If there are other observations, perhaps gathered by other researchers, against which it can be compared, it is no longer a single observation” (KKV 211; see also sections 1.2.1 and 4.4.4, the latter devoted entirely to this point). Rogowski may have overlooked these passages. If we did not emphasize the point sufficiently, we are grateful for the opportunity to stress it here.

Lijphart: The Case Study That Broke the Pluralist Camel’s Back

What was once called *pluralist theory* by David Truman and others holds that divisions along religious and class lines make politics less able to resolve political arguments via peaceful means through democratic institu-

tions. The specific causal hypothesis is that the existence of many cross-cutting cleavages increases the level of social peace and, thus, of stable, legitimate democratic government.

In *The Politics of Accommodation*, Arend Lijphart (1975 [1968]) sought to estimate this causal effect.⁴ In addition to prior literature, he had evidence from only one case, the Netherlands. He first found numerous observable implications of his descriptive hypothesis that the Netherlands had deep class and religious cleavages, relatively few of which were cross-cutting. Then—surprisingly from the perspective of pluralist theory—he found considerable evidence from many levels of analysis that the Netherlands was an especially stable and peaceful democratic nation. These descriptive inferences were valuable contributions to social science and important in and of themselves, but Lijphart also wished to study the broader causal question.

In isolation, a single study of the Netherlands, conducted only at the level of the nation at one point in time, cannot produce a valid estimate of the causal effect of cross-cutting cleavages on the degree of social peace in a nation. But Lijphart was *not* working in isolation. As part of a community of scholars, he had the benefit of Truman and others having collected many prior observations. By using this prior work, Lijphart could and did make a valid inference. Prior researchers had either focused only on countries with the same value of the explanatory variable (many cross-cutting cleavages) or on the basis of values of the dependent variable (high social conflict). Previous researchers therefore made invalid inferences. Lijphart measured social peace for the other value of the explanatory variable (few cross-cutting cleavages) and, by using his data in combination with that which came before, made a valid inference.

Lijphart's classic study is consistent with our model of good research design. As he stressed repeatedly in his book, Lijphart was contributing to a large scholarly literature. As such, he was not trying to estimate a causal effect from a single observation; nor was he selecting on his dependent variable. Harvesting relevant information from others' data, although often overlooked, may often be the best way to obtain relevant information.

By ignoring the place of Lijphart's book in the literature to which it was contributing, Rogowski is unable to recognize the nature of its contribution. Rogowski's alternative explanation for the importance of this book and the others he mentions—that "(1) all of them tested, relied on, or proposed, clear and precise *theories*; and (2) all focused on *anomalies*" (95 this

4. Lijphart also went to great lengths to clarify the precise theory he was investigating, because it was widely recognized that the concept of pluralism was often used in conflicting ways, none clear or concrete enough to be called a theory. Ronald Rogowski's description of pluralism as a "powerful, deductive, internally consistent theory" (97 this volume) is surely the first time it has received such accolades.

volume)—suggests one of many possible strategies for choosing topics to research; but it is of almost no help with practical issues of research design or ascertaining whether a theory is right or wrong. Indeed, the only way to determine whether something is an anomaly in the first place is to follow a clear logic of scientific inference and theory evaluation, such as that provided in *Designing Social Inquiry*.

Allen: Distinguishing History from Social Science

The Nazi Seizure of Power is an account of life in an ordinary German community. Allen is not a social scientist: In his book, he proposes no generalization, evaluates no theory, and does not refer to the scholarly literatures on Nazi Germany; rather, he zeroes in on the story of what happened in one small place at a crucial moment in history, and he does so brilliantly. In our terms, he is describing historical detail and occasionally also conducting very limited descriptive inference. We emphasize the importance of such work: “Particular events such as the French Revolution or the Democratic Senate primary in Texas may be of intrinsic interest: they pique our curiosity, and if they were preconditions for subsequent events (such as the Napoleonic Wars or Johnson’s presidency) we may need to know about them to understand those later events” (KKV 36).

In our view, social science must go further than Allen. The social scientist must make descriptive or causal inferences, thus seeking explanation and generalization. Indeed, we think even Rogowski would not accept Allen’s classic work of history as a dissertation in political science. Allen’s work is, however, not irrelevant to the task of explanation and generalization that is of interest to us. In the hands of a good social scientist, who could place Allen’s work within an intellectual tradition, it becomes a single case study in the framework of many others. This, of course, suggests one traditional and important way in which social scientists can increase the amount of information they can bring to bear on a problem: read the descriptive case-study literature.

THE PERILS OF AVOIDING SELECTION BIAS

We agree with David Collier’s observation that, if our arguments concerning selection bias are sustained, then “a small improvement in methodological self-awareness can yield a large improvement in scholarship” (1995: 461). Indeed, because qualitative researchers generally have more control over the selection of their observations than over most other fea-

tures of their research designs, selection is an especially important concern (a topic to which we devote most of our chapter 4).⁵

Rogowski believes that we would criticize Peter Katzenstein's (1985) *Small States in World Markets* or Robert Bates's (1981) *Markets and States in Tropical Africa* as inadmissibly selecting on the dependent variable. We address each book in turn.

Katzenstein: Distinguishing Descriptive Inference from Causal Inference

Peter Katzenstein's (1985) *Small States in World Markets* makes some important descriptive inferences. For example, Katzenstein shows that small European states responded flexibly and effectively to the economic challenges that they faced during the forty years after World War II; and he distinguishes between what he calls "liberal and social corporatism" as two patterns of response. But many of Katzenstein's arguments also imply causal claims—that in Western Europe "small size has facilitated economic openness and democratic corporatism" (1985: 80), and that in the small European states, weak landed aristocracies, relatively strong urban sectors, and strong links between country and city led to cross-class compromise in the 1930s, creating the basis for postwar corporatism (1985: chap. 4).

Katzenstein seeks to test the first of these causal claims by comparing economic openness in small and large states (1985: 86, table 1). To evaluate the second hypothesis, he compares cross-class compromise in six small European states characterized by weak landed aristocracies and strong urban sectors, with the relative absence of such compromise in five large industrialized countries and Austria, which had different values on these explanatory variables. Much of his analysis follows the rules of scientific inference we discuss—selecting cases to vary the value of the explanatory variables, specifying the observable implications of theories, and seeking to determine whether the facts meet theoretical expectations.

But Katzenstein fudges the issue of causal inference by disavowing claims to causal validity: "Analyses like this one cannot meet the exacting standards

5. Selection problems are easily misunderstood. For example, Caporaso claims that "if selection biases operate independently of one's hypothesized causal variable, it is a threat to internal validity; if these same selection factors interact with the causal variable, it is a threat to external validity" (1995: 460). To see that this claim is false, note, as Collier reemphasizes, that Caporaso's "selection factors" can also be seen as an omitted variable. But omitted variables cannot cause bias if they are independent of your key causal variable. Thus, although the distinction between internal and external validity is often useful, it is not relevant to selection bias in the way Caporaso describes.

of a social science test that asks for a distinction between necessary and sufficient conditions, a weighting of the relative importance of variables, and, if possible, a proof of causality" (1985: 138). However, estimating causal inferences does not require a "distinction between necessary and sufficient conditions, a weighting of the relative importance of variables," or an absolute "proof" of anything. Katzenstein thus unnecessarily avoids causal language and explicit attention to the logic of inference which results. As we explain in our book, "avoiding causal language when causality is the real subject of investigation either renders the research irrelevant or permits it to remain undisciplined by the rules of scientific inference" (KKV 76).

Remaining inexplicit about causal inference makes some of Katzenstein's claims ambiguous or unsupported. For example, his conclusion seems to argue that small states' corporatist strategies are responsible for their postwar economic success. But because of the selection bias induced by his decision to study only successful cases, Katzenstein cannot rule out an important alternative causal hypothesis—that any of a variety of other factors accounts for this uniform pattern. For instance, the postwar international political economy may have been benign for small, developed countries in Europe. If so, corporatist strategies may have been unrelated to the degree of success experienced by small European states.

In the absence of variation in the strategies of his states, valid causal inferences about their effects remain elusive. Had Katzenstein been more attentive to the problems of causal inference that we discuss, he would have been able to claim causal validity in some limited instances, such as when he had variation in his explanatory and dependent variables (as in the 1930s analysis). More importantly, he would also have been able to improve his research design so that valid causal inferences were also possible in many other areas.

Rogowski is not correct in inferring that we would dismiss the significance of *Small States in World Markets*. Its descriptions are rich and fascinating, it elaborates insightful concepts such as liberal and social corporatism, and it provides some evidence for a few causal inferences. It is a fine book, but we believe that more explicit attention to the logic of inference could have made it even better.

Bates: How to Identify a Dependent Variable

Rogowski claims that Robert Bates's purpose in *Markets and States* was to explain economic failure in tropical African states, and that by choosing only states with failed economies and low agricultural production, Bates biased his inferences. If agricultural production were Bates's dependent variable, Rogowski would be correct, since (as we argue in *Designing Social Inquiry*; see also Collier 1995) using—but not correcting for—this type of

case selection does bias inferences. However, low agricultural production was, in fact, not Bates's dependent variable.

Bates's book makes plain his two dependent variables: (1) the variations in *public policies* promulgated by African states and (2) differences in the *group relations* between the farmer and the state in each country. Both variables vary considerably across his cases. Bates also proposed several explanatory variables, which he derived from his preliminary descriptive inferences. These include (1) whether state marketing boards were founded by the producers or by alliances between government and trading interests, (2) whether urban or rural interests dominated the first postcolonial government, (3) the degree of governmental commitment to spending programs, (4) the availability of nonagricultural sources for governmental funds, and (5) whether the crops produced were for food or export. These explanatory variables do vary, and they helped account for the variations in public policy and state-farmer relations that Bates observed.

As such, Bates did not select his observations so they had a constant value for his dependent variable. Moreover, he did not stop at the national level of analysis, for which he had a small number of cases and relatively little information. Instead, he offered numerous observable implications of the effects of these explanatory variables at other levels of analysis within each country. As with many qualitative studies, Bates had a small number of cases but an immense amount of information. We believe one of the reasons Bates's study is—and should be—so highly regarded is that it is an excellent example of a qualitative study that conforms to the rules of scientific inference. In sum, Rogowski says that Bates wrote an excellent book that we would reject. If the book were as Rogowski describes it, we very well might reject it. Since it is not—and indeed is a good example of our logic of research design—we join Rogowski in applauding it.⁶

TRIANGULAR CONCLUSIONS

We conclude by emphasizing a point that is emphasized both in *Designing Social Inquiry* and in the reviews. We often suggest procedures that qualitative researchers can use to increase the amount of information they bring to bear on evaluating a theory. This is sometimes referred to as "increasing the num-

6. Subsequently, Bates pursued the same research program. For example, in *Essays on the Political Economy of Rural Africa* he evaluated his thesis for two additional areas—colonial Ghana and Kenya (1983: chap. 3). So Bates did exactly what we recommend: having developed his theory in one domain, he extracted its observable implications and moved to other domains to see whether he observes what the theory would lead him to expect.

ber of observations." As all our reviewers recognize, we do not expect researchers to increase the number of full-blown case studies to conduct a large-N statistical analysis: our point is not to make quantitative researchers out of qualitative researchers. In fact, most qualitative studies already contain a vast amount of information. Our point is that appropriately marshaling all the thick description and rich contextualization in a typical qualitative study to evaluate a specific theory or hypothesis can produce a very powerful research design. Our book demonstrates how to design research in order to collect the most useful qualitative data and how to restructure it even after data collection is finished, to turn qualitative information into ways of evaluating a specific theory. We explain how researchers can do this by collecting more observations on their dependent variable, by observing the same variable in another context, or by observing another dependent variable that is an implication of the same theory. We also show how one can design theories to produce more observable implications that then put the theory at risk of being wrong more often and easily.

This brings us to Sidney Tarrow's suggestions for using the comparative advantages of both qualitative and quantitative researchers. Tarrow is interested specifically in how unsystematic and systematic variables and patterns interact, and seems to think that principles could be derived to determine what unsystematic events to examine. We think that this is an interesting question for any historically sensitive work. Many unsystematic, nonrepeated events occur, a few of which may alter the path of history in significant ways; and it would be useful to have criteria to determine how these events interact with systematic patterns. We expect that our discussions of scientific inference could help in identifying which apparently random, but critical, events to study in specific instances, and we are confident that our logic of inference will help determine whether these inferences are correct; Tarrow or others may be able to use the insights from qualitative researchers to specify them more clearly. We would look forward to a book or article that presented such criteria.

Another major point made by Tarrow is that all appropriate methods to study a question should be employed. We agree; a major theme of our book is that there is a single unified logic of inference. Hence it is possible effectively to combine different methods. However, the issue of triangulation that Tarrow so effectively raises is not the use of different logics or methods, as he argues, but the triangulation of diverse *data sources* trained on the same problem. Triangulation involves data collected at different places, sources, times, levels of analysis, or perspectives, data that might be quantitative, or might involve intensive interviews or thick historical description. The best method should be chosen for each data source. But more data are better. Triangulation, then, refers to the practice of increasing the amount of information brought to bear on a theory or hypothesis, and that is what our book is about.

D. DIVERSE TOOLS, SHARED STANDARDS

8

Critiques, Responses, and Trade-Offs: Drawing Together the Debate

David Collier, Henry E. Brady, and Jason Seawright

The past two decades have seen the emergence of an impressive spectrum of new techniques for quantitative analysis, as well as the strong resurgence of interest in developing and refining the tools of qualitative research. The intellectual vitality of these two traditions, along with the apparent divergences between them, has sharply posed the challenge of evaluating their respective strengths and weaknesses, producing a major new methodological dialogue. The present volume seeks to extend and refine this dialogue.

A basic point of reference in this discussion has been King, Keohane, and Verba's *Designing Social Inquiry* (KKV), which has broken new ground in the ongoing effort to develop a shared framework for both quantitative and qualitative analysis. Compared to KKV, the present volume places far greater emphasis on the limitations of quantitative tools and on the contributions of qualitative methods to addressing these limitations.

The chapters in the present volume present diverse perspectives on this debate. Chapters 3 and 4 by Brady and Bartels, respectively, draw in part on insights from what we have referred to as statistical theory. They argue that the perspective of mainstream quantitative methods advocated by KKV is an inadequate foundation for a general methodological framework. Chapters 5 and 6 by Rogowski and Tarrow, as well as online chapters 1–4 by Collier, Mahoney, and Seawright, Munck, Ragin, and McKeown, offer insights more centrally drawn from the qualitative tradition. These chapters systematically review methodological tools employed by qualitative researchers and maintain that our understanding and evaluation of these

tools cannot simply be subordinated to the framework of mainstream quantitative methods, as they argue KKV proposes. In chapter 7, King, Keohane, and Verba's interim response (reprinted from an earlier review symposium) focuses on key issues in this discussion of quantitative versus qualitative methods, questioning arguments made in other chapters regarding theory, concepts, selection bias, no-variance designs, and the evaluation of evidence from case studies. Their chapter, like several others, underscores the importance of linking quantitative and qualitative methods in the framework of careful attention to research design.

We now synthesize and push further this discussion. We first revisit four critiques of KKV, concerning the challenge of doing research that is "important," conceptualization and measurement, selection bias, and probabilistic versus deterministic models of causation. Given our concern with finding new ways to bridge alternative methodological traditions, we consider statistical responses that might be made to each critique and the overall conclusions that may be drawn. In the final part of the chapter, given that these critiques and responses often hinge on contending goals of research, we explore the basic theme that methodology involves fundamental trade-offs. A major concern of research design should be with managing these trade-offs. Chapter 9 then further develops our conclusions to the book by focusing on alternative sources of leverage in causal inference.

CRITIQUES AND STATISTICAL RESPONSES

In addressing broad issues of methodology, KKV relies centrally on the framework of mainstream quantitative methods. The book has attracted wide attention in part because this framework provides a standardized perspective and vocabulary for addressing many methodological questions. Given that the quest for shared standards of methodology and research design is an abiding concern in the social sciences, KKV's framework appropriately commands great attention. For example, David Laitin (1995: 454), in his review essay on KKV, underscores the book's potential role in "disciplining political science."

In light of the positive reception accorded to KKV, how are we to evaluate the diverse critiques that have been offered in the present volume—critiques that incorporate both a qualitative perspective and statistical arguments? One option is to ask: Can we gain additional leverage by stepping back and further exploring these critiques of KKV from the standpoint of statistical theory? The following sections adopt this approach to reviewing four significant critiques. For each of these four topics, we first present a brief synopsis of KKV's position, occasionally adding examples or points of clarification. We then summarize the critiques of KKV presented in the

chapters above, which combine the broader statistical perspective offered by Brady and Bartels and the qualitative perspective that is central to the other chapters. Occasionally, we supplement this discussion by reference to additional writings of our authors, or closely related critiques made by other scholars. Finally, we explore further responses to the critique that could be made from the viewpoint of statistical theory.

For two of the topics addressed—the challenges of doing research that is “important” and of evaluating deterministic models of causation—we find that the statistical response calls into question some aspects of the qualitative critique of KKV, and we seek to reconcile these alternatives. By contrast, for two other topics—conceptualization and measurement and selection bias—we find arguments from a statistical perspective that reinforce the critiques.

Within the larger framework of this book, the discussion of these critiques shows how perspectives drawn from statistical theory can potentially offer shared standards for accommodating the claims advanced by both quantitative and qualitative methodologists.

Doing Research That Is Important

KKV briefly argues (see chap. 2) that scholars should study topics that are important, both in the real world and in relation to a given scholarly literature. But KKV does not provide guidance for how to choose important topics; nor does the book address the concern that the methodological norms it advocates might make it *harder* to do research that is important, which would of course represent a major trade-off in research design. This section reviews these concerns, takes a close look at the statistical rationale for KKV’s deliberately limited attention to theory, and considers the most appropriate balance between these alternative views.

Establishing that research is substantively “important”—or theoretically “innovative” or “creative”—is a complex matter. For the purpose of this discussion, studies that address questions evaluated as being of great normative significance would be considered important—as in Bates’s (1981) study, discussed below, which seeks to explain a pervasive pattern of failed economic growth and human misery across an entire continent. Likewise, studies that help advance theory in a way that gives scholars new leverage in conceptualizing and explaining significant outcomes would also be considered important. For example, recent advances in Downsian spatial modeling provide valuable new tools for analyzing dramatic change in party systems (e.g., Kitschelt 1994; Greene 2002). By contrast, some critiques of KKV raise the concern that, in adopting the book’s framework, scholars may sharply narrow their substantive research questions, thus producing studies that are less important.

Critique

A recurring theme in the critiques of KKV is that the book provides little guidance in how to achieve major advances in our substantive and theoretical understanding of politics and society. Rogowski argues that KKV's approach is, in general, insufficiently theory driven. He draws on ideas about the philosophy and practice of science to develop his thesis. Rogowski suggests that KKV's framework fails to account for the achievements of many well-known studies that have greatly advanced theory, even though they do not follow KKV's guidelines. His examples include such influential works as William Sheridan Allen's (1965) *The Nazi Seizure of Power* and Arend Lijphart's (1975 [1968]) *The Politics of Accommodation*, as well as Bates's study noted above.¹ Rogowski points out that these studies do not meet the methodological standards proposed by KKV, in that they lack variance on the dependent variable, which should, in turn, undermine causal inference. King, Keohane, and Verba (118–21 this volume) disagree with Rogowski's interpretation of some of these studies, arguing, for example, that Bates did have variance on some dependent variables.² Notwithstanding these specific disagreements, Rogowski's overall argument stands: We sometimes do face a conflict between (a) the methodological goals of improving descriptive and causal inference on the basis of empirical data, and (b) the objective of studying humanly important outcomes and developing theory that helps us to conceptualize and explain them.

McKeown raises the concern that KKV provides no heuristics for theory construction (chap. 4, online). Ragin suggests that KKV's warning against

1. In addition to Rogowski's summary of these books, see the discussion by King, Keohane, and Verba (116–18 this volume).

2. We wish to comment here on alternative interpretations of Bates's study. Rogowski's (94 this volume) position is that Bates lacks variance on his main dependent variable, in that he focuses on "cases of economic failure, or, more precisely, on the remarkably uniform *pattern* of economic failure among the states of post-independence Africa." By contrast, King, Keohane, and Verba (120–21 this volume) argue that a number of key factors in Bates's study do vary, including the two factors they identify as his dependent variables. In our view, Bates develops a complex, multistep causal argument, and some of the variables in that argument certainly do vary across his cases. For example, Bates finds that in Ghana, a small group of wealthy farmers receives a disproportionate amount of government aid compared to the many poor farmers (Bates 1981: 54–61). However, other dependent variables of the study, such as "the apparent shortfalls in agricultural production in Africa" (Bates 1981: 2), are treated as constant across the cases. Our overall conclusion is that although Bates essentially treats his principal dependent variable as not varying, there is variance on additional dependent variables included in his argument. Thus, Rogowski, as well as King, Keohane, and Verba, focusing on different parts of Bates's argument, both have a point.

the use of “no-variance” research designs would preclude a valuable method for gaining new theoretical understanding (chap. 3, online). Analysts may observe telling commonalities within a set of cases that all share the relevant outcome, and subsequent efforts to explain these commonalities can generate new theoretical insights (chap. 3, online). Ragin (2000: 88–104), for example, has presented a method for theoretically generalizing this kind of insight. Munck (chap. 2, online), and also Collier, Mahoney, and Seawright (chap. 1, online), likewise argue that no-variance research designs can be a valuable source of insight if the scholar employs within-case analysis.

Statistical Response

In formulating a statistical response, we first underscore KKV’s emphasis on the goals of descriptive and causal inference, as well as the book’s statements about what it is not trying to accomplish. KKV is quite explicit about the fact that it is not attempting to provide guidelines for theoretical innovation, quoting Popper’s statement that “there is no such thing as a logical method of having new ideas. . . . Discovery contains ‘an irrational element,’ or a ‘creative intuition’” (KKV 14). Although KKV (38) allows that any definition of science must have “room for ideas regarding the generation of hypotheses,” the book maintains a strict separation between this process and the procedures of “valid scientific inference,” which are its main focus. For example, when the authors (chap. 7, this volume) reject no-variance designs, the book does so on grounds wholly unrelated to the goals of generating hypotheses and learning about unfamiliar phenomena. Instead, it rejects no-variance designs because they provide a weak basis for causal inference. In their response to commentators, King, Keohane, and Verba (114–15 this volume) reiterate their goal: to improve inference, not to provide guidelines for generating theory. As these authors formulate it in KKV (16), “[t]his book offers no advice on becoming brilliant.”

From a statistical perspective, KKV’s advice need not be understood as identifying the only types of studies that can lead to productive findings. Indeed, any given piece of research may yield correct inferences or incorrect inferences, regardless of the procedures used in conducting that research. What statistical reasoning seeks to provide are guidelines that *increase the probability* of generating a correct inference, as well as tools for estimating that probability. Therefore, very crucially, an appropriate way to judge KKV’s procedures is not to compare them with those employed in producing the most innovative works in political science. Rather, it is to inquire whether following their advice will, on average, produce superior inferences.

A closely related statistical rationale for KKV’s approach is that the book’s

framework for descriptive and causal inference provides a standard by which other scholars can evaluate a given study. Thus, scholars may evaluate an inference by judging whether it was made using appropriate methodological tools. KKV's (7–9) definition of scientific research emphasizes public scrutiny of research procedures, and the book's tools for inference represent a valuable step toward a framework that may help scholars meet this standard.

Finally, we wish to insist that any conflict between achieving inferential goals and carrying out theoretically productive research is not just a dilemma in KKV. Rather, it poses a dilemma for all researchers. Further, this is not merely a dilemma that arises in conjunction with specific issues such as selection bias, but rather is a much more general methodological problem. For example, in our discussion in the next chapter of determinate versus indeterminate research designs, we argue that KKV's legitimate objectives of avoiding multicollinearity and increasing the number of observations may pull scholars away from the most direct possible test of their theoretical ideas. This points to the issue of trade-offs: we may face a basic trade-off between attention to certain standards of good inference and the broader priorities of pursuing interesting theoretical ideas.

The Challenge of Promoting Creativity

If we can establish standards for improving and evaluating inference, can we also establish procedures that promote theoretical creativity and lead to important research? On the one hand, the view that we lack systematic procedures for generating novel insights into political phenomena is widely held. As noted above, KKV explicitly states that it does not intend to provide advice on how to be brilliant. Making a parallel argument, a leading advocate of the systematization of case studies, Harry Eckstein, similarly writes that “the Tocquevilles or Bagehots might have been successful in spawning plausible theories without writing case studies, since their imagination and incisiveness clearly matter more than the vehicles chosen for putting them to work” (1975: 146). A researcher may be inspired to think of a new variable that helps explain the outcome of interest by reading Aristotle, Borges, Conan Doyle, or even John Grisham—in addition to gaining insight through carrying out counterfactual thought experiments, or by employing no-variance research designs. The research community should hardly expect hard-and-fast guidelines about how to be creative.

On the other hand, there is good reason to believe that some research practices are more likely to produce theoretical insights than others. Formal, deductive theory can make valuable contributions, although a significant component of the insight associated with such theory depends on substantive insights derived from sources other than the deductive proce-

dures (Powell 1999: chap. 1; Munck 2001: 193–94). Inductive tools for gaining new insights are also well established. Older approaches include Lazarsfeld's elaboration model (Lazarsfeld 1955; Babbie 2004: chap. 15), grounded theory methodology (Glaser and Strauss 1967; Strauss and Corbin 1994), and the procedure of "replacing proper names" of political systems with relevant analytic variables (Przeworski and Teune 1970: 26–30). A more recent formulation of inductive procedures is found in Ragin's (chap. 3 online; also 1987, 2000) methods of "qualitative comparative analysis," including the use of no-variance research designs.

Moreover, specific research activities can be especially useful stimuli for theoretical innovation, even if such activities by no means guarantee inspiration. For example, field research has produced many fundamental insights. Prominent scholars such as Campbell (1975: 182–85) and Piore (1979: 560–61) have underscored the role of fieldwork in overturning established understandings and generating new ideas. Collier's (1999) discussion of the research practice of "extracting new ideas at close range" likewise suggests how field research can generate novel findings. A careful exploration of the specific ways in which field research produces theoretical insights would represent a genuine contribution to social science methodology.

Some of the chapters in the present volume suggest valuable starting points for a broader exploration of techniques that contribute to theoretical innovation. For example, Rogowski (91–96 this volume) emphasizes the value of studying anomalous cases. He discusses famous single-case studies that focus on "most-likely" cases—that is, cases that *should* fit the predictions of an established theory. Such studies can be especially fruitful for gaining insight if these cases turn out *not* to fit, thereby pointing to analytically revealing exceptions to the theory. In a similar vein, Munck (chap. 2 online) discusses several approaches to how case-study research can help analysts generate new theories and hypotheses.

Overall, although no one has an exact formula for being creative, we can certainly identify specific research practices that contribute to creativity.

Innovative Research, Trade-Offs, and KKV's Framework

Scholars can identify research practices that contribute either to improving inference or to promoting theoretical innovation, but not necessarily to both. Hence, we may often face a trade-off in pursuing these alternative goals. KKV's framework for improving causal inference can distract researchers from expanding the range of substantive questions that social science seeks to address. Given that, as McKeown (chap. 4 online) observes, modern social science does not possess "a huge backlog of attractive, highly developed theories that stand in need of testing," this trade-off between theory building and testing is well worth pondering.

This trade-off is made more complex by the fact that theory is routinely seen as a prerequisite for good empirical inference, in that theory generally plays a central role in specifying the models that are tested. For example, theory plays a central role in dealing with the problems of inference highlighted by conditional independence and related assumptions (chap. 2, guideline 26; and Brady 76 this volume). Adequately addressing these assumptions requires, for example, heavily theory-dependent choices about including and excluding variables. Consequently, procedures for improving causal inference that hinder the development of theory may, in turn, impede causal inference.

These potential tensions and complementarities between achieving good inference and developing strong theory also raise issues for how we define “science.” As noted in chapter 2 above, KKV does not merely discuss inference, but also raises a much larger set of issues involved in carrying out “scientific research.” KKV’s carefully formulated definition of scientific research includes the stipulations that “[t]he goal is inference” and “[t]he content is the method” (7, 9). The book could equally well have stated that both the goal and the content of science is theory. The theories employed in different domains of science are certainly heterogeneous, but so also are the methods. There is no reason to think that method, any more than theory, is the essence of science. Both are fundamental, and scholars must recognize the value of both goals.

Conceptualization and Measurement

KKV devotes chapter 2 to descriptive inference, and both there and in many other parts of the book the authors make a number of recommendations about conceptualization and measurement. These recommendations include brief, general advice about the validity and reliability of measurement, the effects of measurement error on causal inference, the kinds of concepts that should be studied, and typologies (see guidelines in chap. 2, this volume). Thus, KKV (25, italics omitted) states that scholars should “maximize the validity of . . . measurements,” and they should use reliable data-collection procedures that, if applied again, would yield the same data. The book (157–68) discusses the impact of measurement error on descriptive and causal inference, pointing, for example, to the relatively familiar claim that whereas error in measuring the dependent variable does not bias causal estimates, error in the independent variable biases causal estimates toward zero.

Regarding the selection of concepts, KKV urges researchers to “choose observable, rather than unobservable, concepts wherever possible” (109). Specifically, “[a]ttempting to find empirical evidence of abstract, unmeasurable, and unobservable concepts will necessarily prove more difficult and

less successful than for many imperfectly conceived specific and concrete concepts" (110). KKV also expresses strong skepticism about the use of typologies: "in general, we encourage researchers *not* to organize their data in this way" (48). Further, the book claims that "it is easiest to maximize validity by adhering to the data and not allowing unobserved or unmeasurable concepts [to] get in the way" (25).

KKV provides brief but useful comments on trade-offs in conceptualization and measurement. Regarding the issue of generality versus concreteness in concepts and theory, the book comments on the tension between the effort to "maximize the concreteness" of our theories (109–12) and the priority that theories should be stated in the most encompassing way feasible (113–14). KKV likewise notes the trade-off, in the use of nominal categories as opposed to higher levels of measurement, between "descriptive richness and facilitation of comparison" (154), as well as the familiar trade-off between measurement validity, on the one hand, and reliability and precision on the other (152).

In the present section, we focus on general issues of conceptualization and measurement. The question of trade-offs is explored later in this chapter.

Critique

The authors in the present volume have several concerns about KKV's approach to conceptualization and measurement. First, in a book of KKV's scope, such topics require extensive attention, rather than brief commentary. Conceptualization and measurement are, after all, basic to the way scholars frame topics and establish procedures for making observations. Furthermore, the validity of causal inference often depends just as much on conceptualization and measurement as it does on KKV's central concerns with having adequate variance, sufficient degrees of freedom, and well-specified models.

Yet Brady observes that, notwithstanding the importance of conceptualization and measurement, in KKV's framework "the problems of theory construction, concept formation, and measurement recede into the distance" (77 this volume). Bartels likewise suggests that KKV's methodological framework neglects research aimed at refining concepts (85 this volume), and Laitin's (1995: 455–56) review essay similarly underscores KKV's inattention to conceptual issues. Overall, commentators believe that research focused on concepts makes just as big a contribution to advancing knowledge as empirical research that seeks to make descriptive or causal inferences.

Second, regarding KKV's advice to employ concepts that readily lend themselves to operationalization, Brady (77 this volume) underscores the

central methodological challenge of coming to grips with difficult concepts such as civil society, deterrence, democracy, nationalism, material capacity, corporatism, group-think, and credibility. Successful measurement always depends on having a well-developed understanding of the concept we want to measure, and efforts at conceptualization and measurement routinely need to tackle theoretical concepts such as these. Laitin (1995: 455–56), in his commentary on KKV, likewise calls attention to the complex concepts with which scholars routinely work: charisma, hegemony, political culture, social mobilization, and division of labor, as well as exit, voice, and loyalty. Serious attention to the methodological challenges inherent in conceptualizing and measuring complex concepts is imperative if they are to be useful in political research.

Third, KKV's skeptical advice about typologies is seen as striking at the heart of the qualitative enterprise, in much the same way that KKV's recommendations about increasing the number of observations are seen as a mandate for qualitative, small-N researchers to give up the kind of research they do.³ Munck emphasizes the importance of typologies as a fundamental tool in political analysis. Typologies play a central role not only in areas in which their use is familiar—for example, delineating types of national political regimes and types of international systems—but also in other domains: for example, Sundquist's (1973: chap. 2) typology of electoral realignment, Collier and Collier's (1991: 7–8, 15–18, 162–68) typology of labor incorporation, and Boix's (1998: chap. 1) typology of economic growth strategies. Further, Brady emphasizes the importance of typological thinking as an explanatory tool (71 this volume).

Fourth, other concerns focus on the treatment of measurement. Bartels (85–86 this volume) finds KKV's discussion of measurement error "incomplete and unrealistically optimistic." He suggests that the book's observations concerning the effect of random measurement error in the independent variable pertain only in the bivariate case. In the multivariate case, error in the estimate for any one variable can produce complex forms of error in the estimates for other variables, even if these other variables are measured without error (see also Bollen 1989: 154–67). Brady likewise discusses the broader literature on measurement and measurement theory, arguing that KKV's framework inappropriately neglects basic ideas and research tools in this literature. He suggests that the leverage methodologists can bring to reasoning about the differences between quantitative and qualitative research would be greatly strengthened by close attention to these ideas and tools (76–81 this volume).

KKV pays almost no attention to contextual specificity of conceptualiza-

3. This concern about KKV's advice regarding the number of observations is expressed by Brady (69–70 this volume) and Munck (chap. 2, online).

tion and measurement. This key issue arises not only in broad cross-national comparisons, but also in disaggregated comparisons of subunits and in comparisons of change over time. This lack of concern with contextual specificity leads to strong misgivings about several of KKV's recommendations, especially the recurring advice to increase the number of observations. Increasing the N has a downside—specifically, it may take the analysis outside of the domain where given concepts are appropriate and measurements remain valid. This may occur either when the analyst moves to a new spatial or temporal domain of cases, or when researchers focus on subunits within an established domain. These subunits may in effect involve a different context, due to heterogeneity within units.

Ragin and Munck devote considerable attention to this question of contextual specificity. One issue they discuss is conceptual stretching, which occurs when, in a new empirical context, the phenomena to which the component attributes of the concept refer are sufficiently different that an established operationalization no longer yields valid measurement. Two well-known means of avoiding conceptual stretching and establishing analytic equivalence are to restrict the domain of cases and, alternatively, to adapt the concept to fit a wider range of cases. Munck (chap. 2 online) points to another option: establishing equivalence by employing system-specific or context-specific indicators, that is, indicators that tap the underlying concept by measuring it in different ways in different contexts. This approach, which remains a basic tool of comparative analysis, has recently been extended by Adcock and Collier (2001: 534–36).

Statistical Response

In light of these critiques, it is productive to consider the response that might be advanced from the standpoint of statistical and psychometric reasoning about these issues. Ideas will also be drawn from the perspective of mathematical measurement theory—including the work of Carl Hempel, whose writings encompass early efforts to formalize basic ideas about measurement.⁴

The very existence of a substantial literature on psychometrics and measurement theory is a useful reminder that conceptualization and measurement are fundamental methodological topics in the social sciences. The perspective that emerges from these literatures generally supports the critiques just discussed, reinforcing arguments about the need for close attention to concept formation, measurement validity, and the contextual specificity of measurement.

With regard to concept formation, the psychometrics literature under-

4. The following discussion incorporates some ideas from Collier and Adcock (1999) and Adcock and Collier (2001).

scores the importance of careful formulation of concepts as a prerequisite for measurement. Shepard (1993: 417) suggests that careful work with concepts should include the specification of both the internal dimensions of a concept and its relationship to other, closely connected concepts. Bollen's (1989: vi, 185–86, 194) analysis, which bridges structural equation modeling and the tradition of content validation,⁵ emphasizes the need for careful analysis focused on the meaning of concepts. He stresses that sophisticated quantitative forms of validity assessment—such as covariance structure models, which he labels structural equation models with latent variables⁶—stand on weak foundations unless basic conceptual questions are resolved. These models provide tools for making choices about what are potentially numerous alternative indicators of a given concept. Bollen argues that, “[j]ust as a nonrepresentative sample of people can lead to mistaken inferences to the population, a nonrepresentative sample of measures can distort our understanding of a concept” (1989: 186). Bollen therefore calls for careful examination of theory and concepts, along with detailed substantive knowledge, to ensure that the set of indicators analyzed is appropriate to the concept. This in turn is essential to achieving valid measurement.

Mathematical measurement theory likewise offers valuable lessons for understanding the relationship between quantitative and qualitative approaches to measurement. These lessons suggest a different perspective about this relationship than that proposed by KKV, which is centrally focused on applying quantitative tools to qualitative research. By contrast, measurement theory comes closer to emphasizing a perspective that might be adopted by qualitative researchers. A fundamental theme in measurement theory is that all quantitative research, in its logical foundations, is ultimately based on qualitative, pairwise comparisons. Measurement theory rests on the appraisal of different logical relations—for example, coincidence, precedence, additivity, reflexivity, symmetry, and transitivity—to establish whether they validly characterize similarities and contrasts within pairs of observations. Reasoning about larger numbers of observations and about higher levels of measurement logically depends on establishing the validity of claims about simple paired comparisons and then aggregating these claims. For example, if the complex requirements of ordinal measurement are not met for two cases, then they certainly are not met for one thousand cases. A major statement of this fundamental idea in measurement theory is found in Krantz, Luce, Suppes, and Tversky (1971: 1–6).⁷

5. Content validation focuses on whether the indicators used to measure a concept are judged to correspond to the substantive “content” of the concept.

6. Other standard labels for these techniques are MIMC (multiple-indicator multiple-cause) models and LISREL-type models.

7. Useful overviews of these issues are found in Coombs, Dawes, and Tversky (1970); Roberts (1976); and Michell (1990: 165–75).

Brady and Ansolabehere (1989) provide a substantive illustration of how ideas about ordinal relationships drawn from measurement theory can be used to evaluate the ordinality assumptions behind the concept of preference, which is central to many lines of inquiry, including, for example, rational choice theory. Their analysis focuses on complex differences in the kinds of ordinality that emerge in respondents' preference orderings regarding candidates in U.S. presidential primaries—involving what are called linear, weak, semi-, interval, partial, and sub-orderings. Distinctions of this kind are standard in the field of psychometrics (Michell 1990: 165–75).

We are convinced that quantitative social scientists should, in general, pay more attention to the foundations of measurement. Further, the procedures through which some qualitative researchers build up their concepts and comparisons on the basis of careful analysis of a few cases is, in many respects, closer to fundamental ideas in measurement theory. An example, drawn from comparative research on democracy, is provided by discussions of how qualitative researchers develop “diminished” subtypes that designate specific forms of “partial” democracy—for example, illiberal democracy or one-party democracy. These subtypes may capture gradations vis-à-vis the concept of democracy more validly than do multistep ordinal scales, which sometimes make the mistake of aggregating nonequivalent gradations of democracy.⁸

Another basic argument in the psychometric tradition is that theory and measurement validity are mutually dependent.⁹ Measurement validity is not an inherent property of a particular indicator. Rather, validity entails a specific understanding of that indicator in relation to a given conceptual and theoretical framework. The reconceptualization of validity by psychometricians in recent years thus embraces a more “theory-based view” that measurement validation must be strongly linked to the analyst's theoretical concerns (Shultz, Riggs, and Kottke 1998: 270; see also Moss 1995: 6; Shepard 1993: 406). Thus, a measure of “democracy” that is appropriate for a scholar seeking to conceptualize, observe, and explain transitions from authoritarian to democratic rule could be quite different from that employed by a scholar focused on conceptualizing, observing, and explaining contrasts in “democracy” in advanced industrial countries.

Further, KKV's warnings about avoiding unobserved and unmeasurable variables would seem to be at odds with the three-decades-long tradition of research identified with what are now called covariance-structure models, as well as the hundred-year-long tradition of work on factor analysis.

8. Collier and Adcock (1999: 560–61); Collier and Levitsky (1997).

9. KKV does recognize one aspect of the way in which descriptive inference is theory dependent (e.g., 55–63), but this topic could have received a more thorough treatment.

Both factor analysis and covariance-structure models are based on the recognition that scholars often work with concepts that cannot be directly measured.¹⁰ In these traditions of research, which make an effort to merge insights drawn from psychometrics and econometrics, unmeasured concepts, that is, latent variables, are the point of departure for both descriptive and causal inference. This represents a different perspective from that embodied in KKV's suggestion, noted above, that validity can be maximized by sticking to the data and avoiding unobservable or unmeasured concepts.

Notwithstanding KKV's advice to avoid difficult-to-operationalize concepts, the book (chap. 3) does in fact follow the approach laid out by statistical theorists (e.g., Neyman 1990 [1923]; Rubin 1974, 1978; Pratt and Schlaifer 1984; Rosenbaum 1984; Holland 1986; and Stone 1993) by putting in the painstaking work required to arrive at a plausible systematization of one of the hardest concepts of all—the concept of causation. Thus, the majority of KKV's advice focuses on how to conceptualize and measure causation. Some scholars in fact believe it is simply too hard, and hence an unproductive enterprise, to conceptualize causation or to measure it in the sense of making adequate causal inferences. However, that is not KKV's position, and it is certainly not ours. Conceptualizing and measuring causation unquestionably deserves the sustained attention it receives both in KKV and in the present volume. Our point is simply that many other difficult concepts similarly require such sustained attention.

Regarding the argument that KKV is excessively optimistic about addressing issues of measurement error, we would note that Bartels's critique (85–86 this volume), discussed above, builds directly on standard statistical treatments of this topic. Evaluating the consequences of measurement error for any particular study is difficult, not only in qualitative research, but also in quantitative research. Quantitative researchers do of course have tools for addressing such error. These include reliability indices, regression using instrumental variables, factor analysis, and, more broadly, covariance structure models, which subsume many other approaches. Such tools are relatively easy to apply, and having some tools available is definitely better than having none. Yet in practice, these tools necessarily provide imperfect estimates, given that they depend on complex and often unverifiable assumptions about the underlying causal structure of the data (Kim and Mueller 1978: 43–46; Bollen 1989: 40–80, 179–223; Greene 2000: 375–86).

10. For a historical overview, see Bollen's (1989: 1–9) discussion regarding the development of covariance-structure models. Obviously, making inferences with these techniques requires a great many assumptions, and these assumptions should be treated with the same caution that we advocate in addressing, for example, the specification assumption in regression analysis.

If these tools for addressing measurement error are subject to major limitations in quantitative analysis, attempts to apply them would seem to pose even greater problems for qualitative researchers, in that they rely on quantitative procedures that are often inapplicable in this latter tradition. However, this gap may not be as great as it appears. Whereas qualitative researchers may not think of themselves as working with the multiple indicators that are essential to these techniques, in making choices about measurement these researchers do often consider alternative indicators. Indeed, these choices can be made in a self-conscious way that at least implicitly utilizes some of the underlying ideas about validation employed by quantitative researchers (Adcock and Collier 2001: 536–43).

KKV's skepticism about typologies likewise seems surprising from the standpoint of the broader statistical tradition discussed here. Relevant statements range from Hempel's (1965: chaps. 6 and 7) discussion of the role played by taxonomy and typological methods in the natural and social sciences, to Bailey's (1994) book *Typologies and Taxonomies*, which provides an overview of statistical procedures for developing classifications. Furthermore, a wide range of common quantitative tools, such as regression with dummy variables and multinomial logit analyses, have been developed for the specific purpose of causal inference with categorical/typological independent and dependent variables.

With regard to the qualitative critics' concern with the contextual specificity of measurement, this idea is also central to measurement theory and psychometrics. Measurement theory treats the notion of a specified domain of applicability as essential to reasoning about conceptualization and measurement, and specifically as a requirement for working with the logical relations that underlie measurement, as discussed above. Hempel's classic *Fundamentals of Concept Formation* designates this domain as "D," and he treats it as the starting point for constructing arguments about different levels of measurement (1970 [1952]: 703–20, 723). As Roberts puts it, "a relation is not properly defined without giving its underlying set" (1976: 476; see also Coombs, Dawes, and Tversky 1970: 13; Michell 1990: 165–66). Thus, the claim that arguments about measurement must be developed in relation to specific contexts or domains is not solely a preoccupation of qualitative researchers who undertake comparisons across diverse cultures and political systems.

Psychometricians likewise argue that the validity of a given indicator must always be treated as context-specific, in that it pertains to a particular domain of cases. The late Samuel Messick, a leading specialist in psychological and educational testing, argues that the validity of a measure should be understood in relation to the specific domain of cases analyzed in the process of validation. The measure should not be generalized to other contexts until the researcher has evidence of its validity in those contexts (Messick

1989: 14–15; 1975: 956; see also Moss 1992: 236–38). For example, a measure of deference to authority that has been exhaustively validated among American college undergraduates is not necessarily valid for Liverpool dockworkers or Brazilian politicians.

To summarize, writing linked to the traditions of psychometrics, mathematical measurement theory, and statistics supports the critics of KKV with respect to conceptualization and measurement validity. Careful decisions about conceptualization and measurement are crucial for empirical research, and these decisions must be a central concern in discussions of methodology and research design.

Finally, we should note that King, Keohane, and Verba (114–15 this volume) respond to concerns about the role of concepts in KKV by suggesting that tools for “concept formation and theory creation,” while valuable, are not emphasized because of the book’s central focus on “empirical research designed to evaluate theories . . .,” that is, on descriptive and causal inference. On the one hand, this is a plausible justification. Concept formation is, in part, an element of theory building. As discussed in the section above on doing research that is important, KKV deliberately chooses not to emphasize theory building, so inattention to concept formation might seem justified and reasonable. On the other hand, as just discussed, concept formation is also a step in the process of operationalization and is therefore central to descriptive inference—and, by extension, causal inference. In this sense, the additional perspectives on conceptualization and measurement offered in the present section are essential in moving beyond KKV’s excessively limited treatment of these topics.

Selection Bias

KKV presents strong and detailed advice about selection bias, framing it as a central problem in causal inference (128–39). Selection bias arises either when cases are selected according to an unrepresentative sampling rule, or when some unknown, nonrandom process assigns causes to cases. This bias can result from selection procedures employed by the investigator, from self-selection of individuals or other units of analysis into the sample, or from self-selection of the cases under study into the categories of a major independent variable.¹¹ Under any of these conditions, tests of explanatory hypotheses can suffer from systematic error.

11. Of these three sources of bias, the problem of the deliberate selection of cases on the dependent variable by the investigator is of particular concern in the present volume. Another principal source of bias, which involves the self-selection of cases specifically into the categories of an independent variable, is explained below.

KKV specifically focuses on the problem of investigator-induced selection bias. The book argues that using any truncated sample will yield causal inferences that, on average, underestimate the importance of the independent variable or variables being evaluated (130). Further, KKV suggests that research designs in which all cases included in the analysis exhibit just one outcome on the dependent variable—for example, a revolution or a severe international crisis—suffer from “extreme selection bias,” and hence “[w]e will not learn about causal effects from them” (130). At the same time, KKV provides advice about appropriate ways to select on the dependent variable, arguing that researchers should select cases across the entire range of that variable.¹²

Critique

A recurring concern of the present volume is that, in making recommendations for qualitative researchers, KKV overextends rules and norms identified with conventional quantitative research. Perhaps in part because “selection bias” sounds like an especially grave error in research design, it has become a catchphrase that lends itself to emphatic advice that further encourages this overextension.

These issues are explored in the chapters by Rogowski and the online chapter by Collier, Mahoney, and Seawright. Several arguments will be reviewed here. First, concern with selection bias should often be considered in light of trade-offs vis-à-vis other methodological and theoretical priorities, as emphasized by Rogowski (97 this volume; see also 131–32 this chapter).

Second, Collier, Mahoney, and Seawright ask whether qualitative research based on cross-case analysis and within-case analysis is subject to selection bias. Qualitative researchers must recognize that such bias can be an issue for cross-case analysis. However, when within-case analysis is based on causal-process observations, selection bias need not arise. Hence, with regard to selection bias, the analogy between regression analysis and these qualitative tools is flawed.

Third, KKV’s treatment of no-variance research designs (i.e., designs focused only on cases with positive scores on the dependent variable) as an extreme case of selection bias is correct for regression analysis, but it provides an inadequate perspective on the application of other analytic tools to such designs. Within-cases analysis based on causal-process observations can be fruitfully employed in what from a regression perspective are no-variance designs (Collier, Mahoney, and Seawright chap. 1; Munck chap. 2; Ragin chap. 3, all online).

12. King, Keohane, and Verba (114 this volume) again call attention to the idea of criteria for selecting on the dependent variable.

Fourth, the very definition of selection bias depends on how the universe of cases is defined. The idea that a researcher is working with a truncated sample only makes sense in relation to a well-defined universe, in relation to which the sample is nonrandom and unrepresentative. Yet defining the universe can be highly problematic, depending as it does on the researcher's assumptions about causal homogeneity and measurement validity, and relatedly on the substantive research question. These issues are of great concern to many qualitative researchers, as emphasized especially in Munck's and Ragin's chapters. It may not be meaningful to raise questions of selection bias until such issues are resolved.

Compared to KKV, commentators in the present volume thus offer a different view of studies focused on extreme cases: They argue that the concern with selecting extreme values on the dependent variable has been oversold, and qualitative researchers have distinctive tools for making valid causal inferences, even if they are dealing with a truncated sample.

Statistical Response

Statistical arguments offer support for KKV's basic claims about selection bias in regression analysis. At the same time, a statistical perspective likewise provides an underpinning for the critiques focused on the application of KKV's ideas to qualitative research.

Statistical theory endorses KKV's argument that regression analysis is useless for the analysis of no-variance designs. When researchers select only cases with one fixed value (which we will call C , for constant) on the dependent variable, they force the error term for each case to be equal to the difference between the causal effect of the independent variable and C . If the causal relationship is positive, this creates a negative relationship between the error term and the independent variable that is exactly equal in magnitude to the positive relationship between the independent variable and the dependent variable. Regression confounds these two relationships, so the overall estimate of the causal effect is zero. This argument generalizes to multivariate regression.¹³

This argument suggests that KKV's claim that designs with no variance in the dependent variable make it impossible to evaluate any causal effect is therefore imprecise. With a no-variance design on the *independent* variable, it is indeed impossible to carry out a regression analysis at all because the

13. In the context of a regression model where $Y = Xb + e$, choosing only cases where Y is equal to the fixed value, C , completely determines the value of the error term. Stated another way, $e = C - Xb$. Therefore, the regression normal equations, $Y = Xb + e$, are equivalent to $X'Y = X'Xb + X'C - X'Xb = X'C + X'X(0)$. As a result, regression will estimate the slopes associated with each independent variable as zero.

matrix containing the independent variable will be impossible to invert. By contrast, no such mathematical disaster occurs when there is no variance on the dependent variable. Instead, the causal estimates go to zero due to selection bias. Thus, the regression produces an estimate of the causal effects—but that estimate is wrong. KKV is right to state that regression cannot produce useful estimates of any causal effect with a no-variance design—although the book is technically incorrect in saying that regression-based inferences are impossible with such a design.

Statistical ideas likewise support several arguments about selection bias advanced by qualitative researchers. Discussions of selection bias presuppose a stable, precise definition of the universe of cases. Freedman, Pisani, and Purves (2007: 353–54 and chap. 19 *passim*) argue that many issues of bias cannot be addressed without having a clear prior understanding of the relevant population, and Stolzenberg and Relles (1990: 408), writing from the standpoint of quantitative sociology, observe that our conception of selection bias depends entirely on our conception of the population to which we wish to make inferences.

Finally, there is a sound statistical basis¹⁴ for the claim that conventional quantitative discussions of selection bias do not directly consider the potential contribution of qualitative no-variance designs to the broader goals of theoretical and substantive learning. Specifically, these goals are hard to quantify, so they are not included in the equations behind claims about selection bias. In other words, quantitative analysis can produce specific figures that represent the magnitude of bias associated with a given research design, but such analysis cannot describe the amount of new theoretical and substantive knowledge the design will produce. Hence, qualitative judgment is required if we are to consider these broader goals.

Drawing together these arguments, we conclude that ideas drawn from statistical theory support several of the critiques. Issues of investigator-induced selection bias sometimes arise in quantitative research and in qualitative cross-case analysis—although not for within case analysis. However, other issues need to be addressed before conclusions can be drawn about this kind of selection bias in any particular study.

In concluding this discussion, a broader concern should be raised: for a discipline such as political science, prominent warnings about investigator-induced selection bias may have been something of a red herring. While truncation is in theory a major problem for many statistical tools, it is in practice relatively uncommon for quantitative researchers in the social sciences to deliberately use truncated samples. Likewise, as discussed in chap-

14. We view the following as a statistical argument because it reflects the basic idea that a statistical equation cannot capture the relevance of a variable that is not included in that equation.

ter 1 online, it appears that for qualitative research, concerns about selection bias due to truncation have been seriously overstated. Hence, warnings about this source of selection bias may have distracted scholarly attention from other forms of selection bias which, overall, may be far more prevalent. Specifically, from the standpoint of broader statistical thinking, selection bias that arises either from political and social processes, or through a mismatch between the analytic models employed by the researcher and empirical reality, is almost certainly a more serious and prevalent concern in the social sciences than selection bias due to deliberate truncation.

The problem of self-selection of individuals into the categories of included (independent) variables routinely arises in observational studies in the social sciences. For example, Heckman (1990) has explored this challenge in efforts to assess the impact of unionism on wage differentials, given that workers' decisions about taking unionized jobs generally involve a component of self-selection. The problem of self-selection can also arise at the level of macrocomparative analysis whenever cases are selected into different categories of the included variables through social and political processes that are, inevitably, beyond the investigator's control. For example, Przeworski et al.¹⁵ suggest that democracies may be more likely than authoritarian regimes to break down in the face of poor economic performance. If this is true, then some countries will be "selected in" to the categories of the explanatory variable (regime type) due to their scores on the outcome variable (economic performance). The expected result is an incorrect causal attribution, due to selection bias, concerning the relation of regime type and economic growth.

Selection bias may likewise occur when individuals or other units are selected into or out of the sample through a nonrandom process. Manski (1995: 21) discusses the obvious example of survey research, given that large numbers of potential respondents routinely choose not to participate in surveys. This problem has become particularly severe in telephone surveys. Manski (1995: 21–22) points to other examples as well, including the partially related problem that arises in longitudinal panel surveys, as well as in research on how schooling influences wages, how welfare programs influence labor supply, and how sentencing influences the commission of crimes. In all these areas, the self-selection of some individuals out of the sample forces researchers to make causal inferences through extrapolating from the data about those who participated in the study to those who did not. If, as is likely, these two groups of people are different in substantively

15. See Przeworski (1995); and Przeworski, Alvarez, Cheibub, and Limongi (2000: 9).

relevant ways, adequate extrapolation from one group to the other may be difficult.

In summary, although poor decisions about case selection can sometimes induce selection bias in both quantitative research and qualitative cross-case analysis, selection bias produced by social and political processes is probably a more important problem. In observational studies, when researchers cannot control the processes through which cases are selected into categories on the independent variables (i.e., in observational studies), such bias can severely distort causal inferences because some unmeasured variables may affect both the dependent variable, on the one hand, and the process of assignment to categories of the independent variable, on the other. In essence, this is the problem of the specification assumption—which we discuss in the next chapter—viewed from the standpoint of selection issues.

Probabilistic versus Deterministic Models of Causation

KKV adopts an exclusively probabilistic model of causation, arguing that “the world, at least as we know it, is probabilistic rather than deterministic” (89 n. 11). This focus leads the book (87–89, 204–5, 209–12) to reject techniques for causal assessment that use a “deterministic” perspective.

Before we discuss these issues, a point of terminology must be clarified. In statistics, “deterministic causation” sometimes designates the broad set of models in which the error variance is specified to be zero—that is, models that have no random component. In the vocabulary of qualitative methodologists, by contrast, “deterministic causation” often refers to models of necessary and/or sufficient causation, which represent a subset of the causal models that are deterministic according to the statistical definition. In this section, we follow traditional qualitative usage and treat deterministic causation as referring to necessary and/or sufficient causation.¹⁶

Critique

Some authors are convinced that KKV is limited by its inattention to deterministic models of causation. Munck (chap. 2 online) expresses concern about approaches like KKV’s, which rely on standard regression models and assume a probabilistic approach. KKV’s approach fails to recognize the importance in qualitative research both of hypotheses about determin-

16. We emphasize the distinction between deterministic and probabilistic causal models. Some scholars instead emphasize the contrast between linear models of causation, as opposed to models of necessary and/or sufficient causation. The main idea in this section is that necessary and/or sufficient causation is both deterministic and nonlinear.

istic causation, and of the effort to develop tools that directly test such hypotheses. McKeown (chap. 4 online) also expresses misgivings about KKV's strictly probabilistic perspective, and Ragin (chap. 3 online) maintains that deterministic causation requires scholarly attention (see also Ragin 1987: 39–44, 54–55, 113–18; 2000: 95–96).

Further, critics argue that KKV's recommendation to seek variance on the independent and dependent variable may impede efforts to test deterministic causal models (Ragin chap. 3 online; see also Ragin 2000: 96–99). If the independent and the dependent variables are dichotomous, these authors suggest that the cases providing the main test of necessary causation are those in which the outcome occurs (see cells A and B in figure 8.1), based on what may be called a "positive on outcome" design; further, the cases providing the main test for sufficient causation are those in which the hypothesized cause occurs (cells A and C in the figure), based on what may be called a "positive on cause" design. This approach is a major challenge to KKV's contention that variance on both the independent and dependent variables is essential to causal assessment. More specifically, the argument of these critics challenges KKV's (130) warning that designs lacking variance on the dependent variable (i.e., include only observations in cells A and B, and not in C and D) always constitute an extreme case of selection bias and should be avoided.

Before we turn to the statistical response, it is useful to provide a brief further introduction to deterministic causation, given that this topic may be relatively unfamiliar to some readers. Examples of familiar research procedures that presume deterministic causation include Harry Eckstein's crucial case studies, John Stuart Mill's methods of difference and agreement, and Ragin's method of qualitative comparative analysis. The application of these procedures depends in part on the idea that, in a given bivariate analysis,¹⁷ if a single case deviates from a hypothesized causal pattern, this finding casts serious doubt on the hypothesis. Thus, within a deterministic causal framework, a single variable on its own is hypothesized to have a distinctive causal impact. The variable's presence inevitably causes an outcome if it is a sufficient cause, and its absence definitively prevents an outcome if it is a necessary cause, regardless of the values of other variables. By contrast, a researcher employing a standard probabilistic, multivariate model may be more strongly inclined to treat a deviant case as the result of excluded variables, or as a random outlier.

17. Of course, the scholar may be concerned with multiple explanatory variables. The point is that the hypothesis of deterministic causation posits a decisive relationship between *each* explanatory variable and the outcome variable. Hence, within this framework, each bivariate relationship can productively be evaluated in isolation.

Figure 8.1. Evaluating Necessary and/or Sufficient Causes

Dependent Variable	Occurs	Cell A	Cell B
	Does Not Occur	Cell C	Cell D
		Occurs	Does Not Occur
		Independent Variable	

Research Designs for Testing Necessary and Sufficient Causes

1. Positive on Outcome Design, for Assessing a Necessary Cause: A design with no variance on the *dependent* variable, focusing on cells A and B. Hypothesis is supported by observations in cell A and rejected if observations are found in cell B.

2. Positive on Cause Design, for Assessing a Sufficient Cause: A design with no variance on the *independent* variable, focusing on cells A and C. Hypothesis is supported by observations in cell A and rejected if observations are found in cell C.

3. All Cases Design, for Assessing Necessary or Sufficient Causes: A design in which all cases in the relevant universe (i.e., cells A, B, C, and D) can be included. If cases are found in cell B, necessary causation is ruled out. If cases are found in cell C, sufficient causation is ruled out. All cases that do not rule out a particular causal hypothesis are treated as evidence in favor of that hypothesis.

Note: Adapted from Seawright 2002a: 180.

The other background point that should be underscored is that deterministic causes are increasingly viewed as substantively important in the social sciences. Scholars who have addressed deterministic causation from both Bayesian and non-Bayesian statistical perspectives maintain that deterministic causes play a significant role in political and social theory. Dion (1998: 141) and Seawright (2002a: 180–81) present numerous examples of influential hypotheses about necessary or sufficient causes, and Goertz (2003) has compiled a remarkable inventory of 150 examples of claims about necessary causes, many drawn from prominent authors. A frequently cited example is Wickham-Crowley’s (1992: 9) comparative study of modern revolutions in Latin America. He finds that specific weaknesses of “patrimonial praetorian regimes” are a necessary (and nontautological) requisite for

revolution. This study (1992: 312, 316–18) further hypothesizes that a withdrawal of U.S. support for the existing regime is a necessary cause of social revolution in the region (i.e., cell B in figure 8.1 is empty). In another example, Migdal (1988: 269–71) hypothesizes that, over a long time horizon, weak societal networks are a sufficient cause of state-building (i.e., cell C is empty). It is against this background that Munck and Ragin, in their contributions to this discussion, argue that deterministic causation is neglected in KKV.

Statistical Response

A statistical response to the debate about KKV's position on necessary and sufficient causes provides some support for KKV's critics, but also some support for KKV's perspective. We will present the response in three steps, focusing on the problems that arise if probabilistic tests are employed in assessing what in fact prove to be deterministic causes; the issue of selection bias; and the challenge of finding the most efficient test for assessing necessary and/or sufficient causation.

Probabilistic Tests of Deterministic Causes. Statistical arguments support the position of KKV's critics by showing that, if a deterministic cause is indeed present, then a researcher who only considers a probabilistic model may make invalid inferences. Braumoeller and Goertz (2000: 846–47) provide a statistical demonstration of this point. Unless the hypothesis of necessary causation is explicitly modeled, which would depart from the probabilistic approach of mainstream quantitative methods, then quantitative tools are biased toward inferring that there is some likelihood of the outcome in the absence of the necessary cause. Yet in fact, that probability is zero (i.e., cell B is empty). Such inferential errors occur because some variables that are correlated with the outcome will usually be present to at least some degree, even when the necessary cause is absent. Adopting a conventional quantitative approach based on multivariate linear regression and probabilistic causation invites such errors.

It is therefore essential to use tests that explicitly consider necessary and/or sufficient causation. Dion (1998), Ragin (2000), Braumoeller and Goertz (2000), and Seawright (2002a), drawing in part on Bayesian analysis, suggest that this challenge can be addressed by a variety of research designs and statistical tools. For example, Braumoeller and Goertz (2000) offer a specific procedure for assessing the probability that a given independent variable is a necessary, rather than a probabilistic, cause of the dependent variable. This procedure, which takes the important step of directly testing the hypothesis that the outcome is impossible without the cause, starts with assumptions about the underlying sampling distribution and then estimates the level of measurement error. When confronted with a case

that appears to disconfirm the hypothesis of necessary causation (i.e., a case in cell B of figure 8.1), Braumoeller and Goertz's approach provides criteria for deciding whether the evidence is consistent with necessary causation, given potential problems of sampling error and measurement error; or, alternatively, whether the evidence should count against the hypothesis of necessary causation.

Necessary and/or Sufficient Causes and Selection Bias. Several of the research designs just discussed involve testing a deterministic causal model with no-variance research designs, thereby violating some of KKV's basic precepts. Thus, a test for a necessary cause that focuses solely on cells A and B (figure 8.1), that is, the positive on outcome design, lacks variance on the dependent variable. Likewise, a test for a sufficient cause that focuses only on cells A and C, that is, the positive on cause design, lacks variance on the independent variable.

These designs would therefore appear to pose a major dilemma. KKV argues that research designs which allow no variance on the dependent variable suffer from extreme selection bias (129–30). Yet Ragin, Dion, and Braumoeller and Goertz are correct in ignoring the issue of selection bias in this instance. As discussed in chapter 4 online, selection bias from truncation arises when the mechanism of selection generates a correlation between the error term in the causal model and the independent variable. However, this problem is irrelevant in research based on a deterministic model, because the variance of the error term in such a model is zero—that is, there is no error term. Hence, no matter how cases are selected, there cannot be a correlation between the independent variable and the error term.

To put this point more intuitively, selection bias distorts inferences in regression analysis by overrepresenting atypical cases. However, with a deterministic model, it is irrelevant whether atypical cases are overrepresented or not, since deterministic causal models require even atypical cases to follow the overall pattern. Hence, the unusual pattern of cases generated by truncated sampling does not distort the conclusions that can be drawn about deterministic causation.

*Identifying the Most Efficient Test.*¹⁸ Apart from the issue of selection bias, the question remains of whether, in general, no-variance designs are the most productive way to assess deterministic causation. This issue is currently the subject of an interesting debate, which points to the possibility that KKV's original advice to seek variance on the dependent and independent variable is effectively correct, though for different reasons than the book suggests.

We address this question using the example of necessary causation—

18. This section draws heavily on Seawright (2002a, 2002b).

although a parallel argument can be made for sufficient causation. Ragin (2000: 96–99), Dion (1998: 128–29), and Braumoeller and Goertz (2000: 846, 852–56) argue, following the positive on outcome design, that only cases actually manifesting the outcome being explained (cells A and B) are relevant to assessing a necessary cause. The hypothesis of necessary causation asserts that only cases experiencing the cause (cells A and C) can possibly experience the outcome. Hence, an appropriate test of this hypothesis consists of examining all cases that experience the outcome and verifying that they all experience the cause. Thus, all cases should be in cell A, cell B should be empty, and cells C and D are irrelevant to the assessment.

Is this type of no-variance design the only way to assess necessary causation? In fact, it is not. Seawright (2002a) uses a simple Bayesian analysis to demonstrate that research designs based on sampling from all available cases (including cells C and D) are also a statistically appropriate test of necessary causation. Further, he argues that, on the basis of the standard of statistical efficiency,¹⁹ this “all-cases” design may sometimes be preferable (see figure 8.1). This is particularly true in the study of relatively rare phenomena, for example, the three revolutions studied by Skocpol. She argues that these are the only social revolutions that have occurred in the large domain of historical cases that she identifies as proto-bureaucratic autocracies, located in agrarian societies that have not experienced colonial domination (1979: 40–41). Analysts who study such phenomena may quickly run out of cases that experienced the outcome, yet, using an all-cases design, they can potentially draw on a large pool of analytically equivalent cases where the outcome did not occur. The point here is that any one of these cases might have fallen in cell B, but did not. Other things being equal (for example, the appropriateness of the cases to the analytic question), considering these additional cases therefore yields a stronger inference.

Given that drawing the sample from the entire universe of cases can produce a more efficient causal inference, the central issue is whether or not all cases are in fact relevant as tests of the hypothesis that the causal process is deterministic. As noted above, Dion, Ragin, and Braumoeller and Goertz argue that, for necessary causation, the most appropriate test focuses on cases that experience the outcome (cells A and B), while another possible test focuses on cases that do not experience the cause (cells B and D). Cases that experience the cause but not the outcome (cell C) are irrelevant to both types of tests. These researchers start by conditioning on, or treating as fixed in advance, either the value of the dependent variable or the value of the independent variable, and then considering whether or not the values of the *other* variable confirm or negate the hypothesis of necessary causation.

19. Efficiency is the extent to which a given analytic procedure fully utilizes available evidence to maximize inferential leverage.

On the basis of this reasoning, cases that experience the cause but not the outcome (cell C) are not relevant for falsifying the hypothesis and hence do not constitute tests (e.g., Ragin 2000: 96; Braumoeller and Goertz 2002).

However, Seawright (2002a: 187–89; 2002b: 205–6) argues that it is inappropriate, in working with observational data, to claim that the value of either variable must be treated as fixed in advance. Thus, it is not mandatory that the researcher condition on either the independent or the dependent variable. In observational studies, the scores on the independent and dependent variables are not assigned by the researcher; thus, it is not logically necessary to take either as fixed. Rather, all cases assume their values on the independent and dependent variables through the unfolding of the political and social processes, and all cases are free to assume any combination of values on these variables. Hence, any of the cases could, *a priori*, have falsified the hypothesis, and the examination of any of the cases (cell C, as well as A, B, and D) constitutes a test of the hypothesis. A parallel argument can be made for sufficient causation.

Additional advantages of the all-cases design should be noted. If analysts find evidence against the hypothesis of deterministic causation, they can use the data already collected to estimate the strength of the probabilistic association between the two variables. By contrast, with a positive on outcome or positive on cause design, they cannot. Relatedly, the all-cases design is also more productive if it turns out that: (1) a necessary or sufficient cause ultimately turns out to fit the hypotheses of both necessary and sufficient causation; (2) what was initially hypothesized to be a necessary cause proves to instead be sufficient; or (3) what was initially thought to be a sufficient cause proves instead to be necessary. In any of these situations, if the researcher limits case selection to a positive on outcome or positive on cause design, it is impossible to do further hypothesis testing without collecting additional data. These are important drawbacks of no-variance designs.

This discussion demonstrates that a number of statistical tools are available for empirically testing hypotheses of deterministic causation against probabilistic alternatives.²⁰ Moreover, researchers are working to refine the statistical foundations of these tools (e.g., Clarke 2002; Braumoeller and Goertz 2002; and Goertz and Starr 2003). As just discussed, recent work suggests that the strongest tests of deterministic hypotheses may in fact include variance on both the independent and the dependent variables. In this respect, the more traditional advice of mainstream quantitative meth-

20. The tests discussed in this section are incapable of distinguishing among probabilism due to unobserved variables, measurement error, or a genuinely probabilistic causal process. However, they do distinguish between these three forms of probabilism, on the one hand, and deterministic hypotheses on the other.

ods remains relevant to the study of these distinctive forms of causation, although conventional regression analysis does not provide an appropriate test. Rather, analysts should use statistical techniques, such as those discussed above, that directly evaluate hypotheses about necessary and/or sufficient causation.

The Statistical Responses: Some Conclusions

One of our goals, both in this section and in this book overall, is to explore a range of methodological issues from three different perspectives: mainstream quantitative methods, qualitative methods, and statistical theory. KKV presents a synthesis of mainstream quantitative methods. The four critiques just discussed draw heavily on the perspective of qualitative methodologists, although they include commentaries by Brady and Bartels that, to a significant degree, employ the broader perspective of statistical theory. In response to each critique, we introduce additional arguments from statistical theory in order to gain new leverage for addressing each concern.

For two of the topics addressed in this chapter—that is, doing research that is important and probabilistic versus deterministic views of causation—we find that statistical responses in some respects support KKV. For the question of doing research that is important, the statistical perspective calls attention to the potential trade-off between striving for importance, as opposed to valid inference. With regard to testing hypotheses about deterministic causation, the no-variance designs employed for this purpose have been criticized as being subject to extreme selection bias. On the one hand, the discussion above shows that KKV's rejection of no-variance designs is based on a regression perspective that is not appropriate for evaluating necessary and sufficient causes. On the other hand, all-cases designs, with variance on the independent and dependent variables, can in fact be more efficient than no-variance designs, a conclusion that more nearly supports KKV's priorities in research design, though for different reasons than those offered by KKV.

For the other two topics—conceptualization and measurement and selection bias—the responses drawn from statistical theory either directly reinforce the critiques advanced by qualitative researchers, or make parallel arguments that push the discussion in the same direction. This is particularly the case with regard to conceptualization and measurement. With regard to selection bias, we point to statistical arguments, beyond the mainstream quantitative arguments advanced by KKV, that can provide valuable guidance to qualitative researchers. Scholars who use statistical tools, based on detailed and precise arguments about evidence and inference, thus reach the same conclusions about these issues as the qualitative critics. This

points to a convergence between qualitative and statistical perspectives on research design, yet a convergence quite distinct from the imposition of quantitative norms on qualitative research proposed by KKV.

In sum, perspectives drawn from statistical theory sometimes reinforce the views of qualitative methodologists and sometimes those of mainstream quantitative methodologists. Statistical theory can thus provide an independent standard for adjudicating these methodological debates.

TRADE-OFFS IN RESEARCH DESIGN

The critiques and statistical responses concerning these four major topics point to the fact that, in social science methodology, all good things do not necessarily go together. Indeed, research involves fundamental trade-offs. An unusually effective introduction to the idea of trade-offs is found in John Gerring's (2001) *Social Science Methodology: A Criterial Framework*. Gerring explores the complex trade-offs entailed in working with concepts, in developing propositions, and in the design of research. With regard to choices about refining concepts, he explores, for example, trade-offs among differentiation, operationalizability, familiarity, parsimony, resonance, and theoretical utility.²¹ Our goal in this section is to situate trade-offs within the more balanced view of methodology we advocate in this volume.

Trade-Offs, Goals, and Tools

Trade-offs may involve conflicts among the *goals* pursued by researchers. Trade-offs also arise with respect to the *tools* employed in pursuing these goals. It is likewise productive to contrast *overarching* and *intermediate* goals, as we explain below. These distinctions will also help us in developing a further theme of this book: the idea that working with diverse tools does not preclude establishing shared standards for evaluating research.

In the methodological framework of the present volume, one overarching goal is to seek valid descriptive and causal inferences about important phenomena in the political and social world. This goal is clearly shared

21. For an overview, see Gerring (2001: 22–26 and 234–39). Other valuable statements about trade-offs are found in Sartori's (1970: 1040–46) discussion of a trade-off between the *intension* (i.e., the meaning) and *extension* (i.e., the range of corresponding observations) of concepts; Ragin's (1987: chaps. 3 and 4) account of case-oriented versus variable-oriented research; and Coppedge's (1999) distinction between concepts and theories that are thick and thin. Sil (2000) discusses a fundamental trade-off between analytic alternatives that broadly parallel those of Ragin. See also the discussion of trade-offs by Przeworski and Teune, Cohen, and Blalock cited in the text below.

with KKV. The pursuit of this goal can be advanced through a second overarching goal: refining theory, in order both to gain leverage in establishing what is important, and to strengthen these descriptive and causal inferences.²² Some scholars may use a different vocabulary in discussing these two overarching goals, but we are convinced that these goals are widely shared in contemporary social science.

Of course, scholars make different choices about how they pursue these overarching goals, and these choices are usefully understood at the level of intermediate goals, which involve more specific research objectives. We noted above David Laitin's priority of "disciplining political science," and we believe that a promising source of such discipline is to be found in the careful discussion of how these intermediate goals can serve the overarching goals.

With regard to intermediate goals related to descriptive inference, according to Cohen (1989: 31–36) scholars may alternatively seek precise communication, empirical import, or fertility in the application of concepts; and, according to Blalock (1982: 27–31), generalizability, simplicity, and precision in conceptualization and measurement. In causal assessment, scholars may strive for generality, parsimony, accuracy, and/or causality²³ (Przeworski and Teune 1970: 20–23). The potential diversity of intermediate goals might be an obstacle to the coherence of scholarship. Yet this obstacle may be overcome: Studies that pursue divergent intermediate goals can make complementary contributions to achieving the overarching goals.

Tools, on the other hand, are specific research practices and procedures aimed at achieving intermediate goals, and through them the overarching goals. Some tools are highly systematized and have elaborate statistical and mathematical underpinnings. Other tools, more commonly found in qualitative research, involve practices and procedures that were not developed with explicit statistical or mathematical justifications—although, as we suggest at various points in this book, statistical justifications can serve to illuminate the leverage provided by these tools. Methodology is concerned both with developing tools and with reasoning about how particular tools succeed or fail in achieving research goals. For example, Rogowski argues that an emphasis on narrow methodological criteria for case selection may distract scholars from a larger focus on theoretical innovation and generating valuable substantive insights into politics and society.

Rogowski's concern is one of many demonstrations that goals and tools involve trade-offs. At the level of intermediate goals, for example, the pur-

22. KKV has been criticized for neglecting theory. Yet as can be seen in the guidelines in chapter 2, the book does consider the links between the methodological issues they discuss and questions of theory.

23. By causality they mean a fully specified causal model.

suit of one particular objective may make it harder to achieve another. In promoting the idea of shared standards that is a basic theme in the present volume, our purpose is to encourage recognition that different choices at the level of intermediate goals may constitute legitimate, alternative means of pursuing the overarching goals. In the examples noted above, in the application of concepts we may encounter a trade-off among precise communication, empirical import, and fertility. Likewise, Przeworski and Teune's formulation constitutes a major example of a trade-off among intermediate goals. They argue, for example, that more general theories are often less accurate and parsimonious. These trade-offs are often quite real, and scholars must recognize that different combinations of generality, parsimony, and accuracy, or of precision and fertility, can be productive in pursuing the overarching goals of causal and descriptive inference.

At the level of tools, trade-offs are also fundamental. For example, in a regression analysis, a no-variance or "low-variance" research design may be a poor choice from the standpoint of concern with selection bias. Yet it can be a good choice in a research domain where basic descriptive information is lacking, and a scholar is using within-case analysis to unearth new information. KKV discusses the strength of nominal categories in terms of "descriptive richness," yet also calls attention to their relative weakness in the "facilitation of comparison" (154). Similarly, cross-national regression analysis based on cross-sectional data has the virtue of providing a concise summary of the relationships among a set of variables across many contexts and of testing the "comparative statics" of theories, that is, contrasts among cases at a given point in time. Yet large-N, cross-national studies too often give insufficient attention to causal mechanisms and to hypotheses about the development of phenomena over time, and such studies may also depend heavily on untested assumptions. In the face of these trade-offs, the idea of shared standards becomes relevant. Thus, it is necessary not merely to criticize given tools in light of their weaknesses, but also to carefully weigh their strengths against these weaknesses in light of what the investigator is trying to accomplish.

In developing what we view as a more balanced approach to the relation between quantitative and qualitative methodology, we are centrally concerned with maintaining this distinction between overarching goals, intermediate goals, and tools, and with focusing on the trade-offs that arise among them. Seeking shared standards for research is much easier if scholars recognize the distinctions among these levels—and if they acknowledge the overarching goals that they share.

A central focus on trade-offs is indispensable, given the tensions among alternative intermediate methodological goals. If we pretend that trade-offs do not exist, it is impossible to have an informed discussion of the objectives being pursued in a given study. Further, the exploration of trade-offs

is not a formula for methodological anarchy. Rather, it is a step toward avoiding anarchic situations where scholars are simply talking past one another. The notion of trade-offs rests on the idea that we do have standards; and we need to be explicit about goals, as well as strengths and weaknesses of alternative means for pursuing these goals. As Gerring emphasizes (2001: 26), the number of criteria relevant to evaluating research is relatively limited. Raising the issue of trade-offs challenges us to specify the criteria we are emphasizing, and to justify our choices.

Trade-Offs in KKV

We see a striking contrast between this focus on trade-offs and the position of KKV. In most research, some methodological goals are simply incompatible. By contrast, KKV's central argument is that scholars should adopt a set of tools that is presumed to meet almost all major methodological priorities; only secondarily does the book mention trade-offs among those priorities.

In fact, scattered throughout the book, KKV does briefly discuss five basic trade-offs. With regard to descriptive inference, KKV briefly comments on the trade-off (just noted above) between measurement validity and precision (152). The trade-off between "descriptive richness" in the use of nominal categories, and "facilitation of comparison" in higher levels of measurement, is mentioned (154). The authors note the tension between the advice to "maximize the concreteness" of theories (109–12) and the suggestion to make them as encompassing as is feasible (113–14). Concerning issues that arise in both descriptive and causal inference, KKV comments, for example, on the trade-off between maximizing observable implications and studying cases that are sufficiently independent of one another to add new information to the analysis (222–23). The book also discusses the trade-off that sometimes arises between minimizing the variance of estimators and achieving unbiasedness in both descriptive and causal inference (66–71, 97).²⁴ However, these are in every case isolated observations. The reader finds no suggestion that a central challenge in methodology is to address choices among potentially incompatible goals, or to evaluate these trade-offs in light of alternative goals.

Placing Trade-Offs at the Center of Attention

We are convinced that making choices among potentially incompatible goals is, in fact, the essence of research design. A major challenge for meth-

24. King, Keohane, and Verba (114–15 this volume) again underscore the importance of this particular trade-off.

odologists is to do a better job of recognizing and explicating the trade-offs they inevitably encounter.

The first section of this chapter focused on the complex trade-off between theoretical innovation and rigorous testing. Additional trade-offs include the five to which KKV refers, as well as the many trade-offs identified by Przeworski and Teune, Blalock, Cohen, and Gerring (see above). We would draw attention to three further trade-offs that are central to this debate: between the precision and generality offered by quantitative tools and the reliance on the often untested assumptions required by these tools; between seeking to avoid bias by including all relevant independent variables in an analysis and seeking to maintain inferential leverage by limiting the number of independent variables; and between the representativeness and interpretability of quantitative tests associated with random sampling, versus the close focus on theoretically relevant comparisons (involving both similarities and contrasts) afforded by careful, nonrandom case selection.

However, for several critics, the most fundamental trade-off raised by KKV's recommendations is between increasing the number of observations and other significant goals. As Brady (69–70 this volume) and Munck (chap. 2 online) observe, this recommendation appears to suggest that qualitative, small-N researchers should solve their basic research problems by ceasing to be small-N researchers. In discussing these trade-offs, we first emphasize that within KKV's framework, increasing the N does serve several legitimate purposes. As noted in chapter 2 above, KKV argues that increasing the N can help in strengthening falsifiability, enhancing explanatory leverage, and addressing indeterminacy and multicollinearity (guidelines no. 4a, 6b, 9a, 30a). Thus, KKV proposes increasing the number of observations in pursuit of legitimate goals.

Yet increasing the number of observations may have serious disadvantages. First, it may take the analysis to a domain that is not appropriate to the research question. In making the case in favor of sticking to observations that are theoretically relevant and appropriate to the research question, KKV does usefully quote Lieberman's (1985: chap. 5) incisive statement regarding this priority. The book fails, however, to mention that Lieberman's argument is a critique of a study in which a researcher sought to greatly increase the N by switching the level of analysis to subunits that Lieberman saw as inappropriate to the research question. Further, KKV does not really follow Lieberman's advice. For example, KKV (24–25) at one point advocates an enormous shift in the domain of analysis in order to add observations to the test of a given hypothesis. Specifically, KKV suggests that scholars might study topics in economics such as pricing strategies and entry into markets as a means of testing the theory of deterrence in international politics. Comparing these different domains might be useful as a

source of hypotheses, but there is no reason to believe that the same causal processes will operate in each of these domains. These comparative “leaps” can involve a major trade-off: they may move scholars too far away from the original research question.

A closely related disadvantage of increasing the number of observations concerns concepts, measurement validity, and causal homogeneity. Overextending concepts to domains in which they are inappropriate is a recurring methodological problem. Measurement validity is context specific, and extending the research domain to increase the number of observations can impose a high cost in terms of validity and reliability. Extending the research domain can likewise make it more difficult to maintain causal homogeneity. The quest to increase the *N* can too easily lead a researcher to introduce cases with different causal structures from those that are central to the research question. The resulting loss in validity of causal inference may more than offset any gain in leverage from having a larger *N*.

Increasing the *N* also makes it more difficult to maintain knowledge of the context. In chapter 2 under guideline no.17, we quoted KKV's (43) forceful statement on the importance of deep knowledge of the research context. Yet this priority receives little attention in the book. Rich background knowledge can be difficult and time-consuming to acquire. Thus, a key question concerns the number of cases for which it can in fact be acquired. Further, scholars face a trade-off between obtaining rich, unstructured knowledge of the context and treating either geographic or temporal subunits of cases as the unit of analysis. After all, cultures and the relevant aspects of history change in complex ways within a society over time, and they may vary in equally intricate ways within each subunit of a society. Obtaining detailed background knowledge of observations at other levels of analysis adds to the cost of research in terms of time and other resources, as does adding new cases. Therefore, seeking to increase the number of observations and also achieve deep knowledge involves a fundamental trade-off.

Finally, as KKV (222–23) does note, multiplying observations can pose a trade-off in relation to the independence of observations. A focus on temporal or spatial subunits can add observations that are not independent either from the initial set of observations, or from one another. Hence, adding observations that are not independent creates a misleading appearance of a bigger *N*, leading, for example, to incorrect estimates of statistical significance.

The trade-offs discussed in the previous paragraphs involve several major intermediate goals that become more difficult to achieve when scholars increase the number of observations. Seeking to increase leverage by moving to a larger *N* may come at a high price. Scholars should be very clear about this trade-off when designing research.

The existence of such trade-offs means that no one set of methodological guidelines can ensure that researchers will do good work. Diverse methodological tools will always be relevant to any substantive problem. The best approach to trade-offs is to recognize them explicitly, to acknowledge that there is usually no single “correct” resolution, and to identify the strengths and weaknesses of different combinations of goals and tools.

CONCLUSION

Given the pervasive role of trade-offs, we argue that several methodological issues are far more complex than they appear in KKV. We have placed particular emphasis on dilemmas related to the book’s most frequently repeated piece of advice: increase the number of observations. The five corresponding trade-offs summarize part of the reason why choices about the N are complex. More broadly, the pervasive importance of trade-offs in research design means that methodological advice must be presented more cautiously than it is in KKV.

We have likewise argued that descriptive inference entails hard decisions about concepts, typologies, measurement relations, and domains of measurement validity. Decisions such as these are largely neglected by KKV. Finally, in our discussions of deterministic causation and selection bias, we have emphasized that advice about causal inference that is valuable in some situations may be counterproductive in others. Methodologists should be careful to tailor their advice to the actual inferential situation of the researcher, a norm that KKV largely disregards.

The goal of the final chapter in Part I of this volume (chap. 9), which follows, is to further refine both the statistical and the qualitative perspective on these dilemmas. We offer a new conceptualization of the different kinds of observations employed in causal inference and in research design more broadly. A central goal is to illustrate how diverse tools can be evaluated in terms of shared standards and overarching goals. Specifically, we show how an emphasis on the goal of valid causal inference can lead to fundamental critiques of mainstream quantitative methods, and to a renewed focus on alternative tools that grow out of the qualitative tradition.

9

Sources of Leverage in Causal Inference: Toward an Alternative View of Methodology

David Collier, Henry E. Brady, and Jason Seawright

The challenge of identifying, assessing, and eliminating rival explanations is a fundamental concern in social research.¹ The goal of this chapter is to synthesize the view of methodology offered in the present volume by considering further the contribution of alternative quantitative and qualitative tools in evaluating rival explanations.

We seek to clarify several methodological distinctions that are essential to understanding causal inference. We also propose a new distinction: between data-set observations and causal-process observations. Our discussion considers the contrasting, yet complementary, forms of inferential leverage provided by each type of observation. In the final section of the chapter, we offer some observations about balancing methodological priorities in the face of the ongoing technification of method and theory in many branches of the social sciences.

REVISITING SOME KEY DISTINCTIONS

Understanding the leverage for causal inference provided by different styles of research requires close attention to several basic distinctions. If these are

1. Snyder (1984/85: 91–92), in contrast to KKV (7–9), explicitly makes the elimination of rival explanations one of his criteria for the scientific method.

not treated carefully, conclusions about alternative sources of leverage may be misleading.

Two broad contrasts are indispensable to the argument we seek to develop: between experiments and observational studies, and between mainstream quantitative methods and perspectives drawn from statistical theory. We then consider three other distinctions, involving more specific statistical issues: determinate versus indeterminate research designs, data mining vis-à-vis specification searches, and the assumptions of conditional independence versus the specification assumption. Readers may refer to the glossary for a compilation of the definitions we employ.

Experiments, Quasi-Experiments, Observational Studies, and Inferential Monsters

As is well known, in experiments analysts randomly assign cases to different treatments, that is, to different values of the key independent variable. In observational studies, by contrast, analysts observe the values that the independent variables acquire through the unfolding of political and social processes. For the purpose of evaluating rival explanations, the most fundamental divide in methodology is neither between qualitative and quantitative approaches, nor between small-N and large-N research. Rather, it is between experimental and observational data. All researchers know this, but they often do not give adequate attention to the severe inferential problems that arise with observational data. In addition to differing on the explanatory variables of interest, such real-world cases may also differ in many other ways that the researcher cannot measure and control for, and that can distort causal inference.²

Concern with these severe inferential problems has led the econometrician Edward Leamer to underscore “the truly sharp distinction between inference from experimental and inference from non-experimental data. . . .” He points out that with the latter, “there is no formal way to know what inferential monsters lurk beyond our immediate field of vision” (Leamer 1983: 39).

Given this apparently sharp dichotomy between experimental and observational data, what are we to make of the intermediate, or hybrid category, the “quasi-experiment,” popularized by Campbell and Stanley (1963: 34–64)?³ A quasi-experimental design is typically based on time-series data,

2. Important problems of causal inference also arise in experiments. External validity is a recurring issue, and obstacles to internal validity can arise as well. Nonetheless, problems of causal inference are far more severe in observational studies.

3. A second legacy of Campbell’s work has been the emergence, under the broad heading of quasi-experiments, of a renewed emphasis on “natural experiments,” in which the mechanisms through which cases receive a value on the main explanatory variable are demonstrably unrelated to the error term. Hence, some of Camp-

involving a sequence of observations focused on the outcome being explained. At some point within this time series, an event, policy innovation, or other change occurs, and the analyst examines prior and subsequent values of the dependent variable in an effort to infer the impact of this event. This design is sometimes called an “interrupted time-series.” A stunning exemplar is Campbell and Ross’s (1968) study of the crackdown on speeding in the state of Connecticut. They explore the surprising difficulties of causal inference encountered in assessing the impact of this crackdown on death rates in automobile accidents. Many of the obstacles to good causal inference they consider are parallel to those confronted in experiments, which reinforces the idea that this design is in many ways like an experiment—hence, quasi-experimental. Although the idea of quasi-experiments is strongly identified with Campbell, he subsequently had misgivings about this hybrid category. He recognized that the studies he had included in this category were actually observational studies, and that it had been misleading to suggest that there is an intermediate type between observational and experimental research. With characteristic humor and irony, Campbell suggests that:

It may be that Campbell and Stanley (1966) should feel guilty for having contributed to giving quasi-experimental designs a good name. There are program evaluations in which the authors say proudly, “We used a *quasi*-experimental design.” If responsible, Campbell and Stanley should do penance, because in most social settings, there are many equally or more plausible rival hypotheses. . . . (Campbell and Boruch 1975: 202)

The central legacy of Campbell’s work on these issues, as both Brady (76–77 this volume) and Caporaso (1995: 459) emphasize, is Campbell’s insightful inventory of threats to validity in observational studies (Campbell and Stanley 1966: 5–6; Cook and Campbell 1979: 51–55). This inventory points to the surprisingly large number of things that can go wrong in making causal inferences from what may initially appear to be relatively straightforward observational data.

These words of caution from both Leamer and Campbell are crucial in assessing KKV’s methodological framework. KKV provides recommendations for researchers engaged in observational studies, yet the book’s discussion of causation takes as a point of departure an experimental model. KKV employs the counterfactual definition of causation, grounded in the model of experiments introduced by Neyman (1990 [1923]), Rubin (1974, 1978), and Holland (1986). We think that this definition is indeed valuable in helping scholars to reason about causation as an abstract concept. How-

bell’s threats to validity (see below) are at least partially averted. Unfortunately, it is often hard to find research contexts in which this criterion is met.

ever, Neyman, Rubin, and Holland intended their definition primarily for application to experimental research. They express skepticism about causal inference based on observational data (Rubin 1978; Holland 1986: 949), and their initial discussions of causation were only secondarily concerned with the challenges faced by researchers who use such data.⁴ An account of causal inference in the social sciences must explicitly consider obstacles to causal inference in observational studies and address their practical implications for research. Yet Brady (73–75 this volume) is concerned that KKV does not adequately address these issues.

As Brady observes, KKV could have been more careful about distinguishing between the methodological strengths of experiments and those of quantitative observational studies. In fact, KKV sometimes seems to confound the tools relevant to experiments and those relevant to conventional quantitative research. For example, KKV is not clear enough in distinguishing between the independence assumption and conditional independence (Brady 74–75 this volume; see also the discussion later in this chapter), the former being relevant to experiments, and the latter applying primarily to observational studies.

Relatedly, KKV offers a somewhat confusing statement about the relationship between randomization and the quantitative/large-N versus qualitative/small-N distinction.⁵ The book argues that:

Randomness in selection of units and in assigning values to explanatory variables is a common procedure used by some quantitative researchers working with large numbers of observations to ensure that the conditional independence assumption is met. . . . Unfortunately, random selection and assignment have serious limitations in small-*n* research. (KKV 115; see also 94)

In this statement, KKV overstates the role of random assignment in conventional quantitative research and in effect lumps together random selection and random assignment, thereby merging the characteristic strengths of experimental design and of quantitative analysis. The book thus comes too close to making it appear as if the main divide is between these two approaches, on the one hand, and small-N, qualitative studies, on the other.

4. Rubin (1980) developed the “stable-unit-treatment-value assumption” (SUTVA) as a formalization of one situation in which observational studies can be analyzed as if they were experiments. This initial move in the direction of discussing causal inference in observational studies is perhaps especially valuable as a statement of the difficulties involved in such inference.

5. In this sentence, we refer to quantitative/large-N versus qualitative/small-N to accommodate the combined usage in the following quotation from KKV. For further discussion of these distinctions, see 178–80 in this chapter.

Caporaso's (1995: 459) commentary on KKV, by contrast, emphasizes the importance of sharply separating these two types of randomization: the random *assignment* carried out in most experiments, versus the random *sampling* that is often used in quantitative observational studies. Caporaso emphasizes that, while random assignment does indeed eliminate several challenges to causal inference, "[r]andom sampling does not solve the problems of drawing inferences when numerous causal factors are associated with outcomes" (1995: 459). Thus, large-N quantitative studies—which rarely employ random assignment—are still left with the basic inferential problem faced by small-N studies.

In sum, experimental and observational studies are profoundly different. The traditions of scholarship discussed in the present volume are based on observational data; quantitative and qualitative researchers therefore face the same fundamental problems of inference. KKV's effort to address the major inferential challenges of small-N, qualitative research—based on the norms and practices of large-N, quantitative research—thus faces a major obstacle: Large-N, quantitative methods confront many of the same inferential challenges as qualitative observational studies. In important respects, quantitative researchers do not have strong tools for solving these dilemmas, as Bartels (84–87 this volume) emphasizes above.

Mainstream Quantitative Methods versus Statistical Theory

Given that our basic concern is with challenges to causal inference that arise in analyzing observational data, where can we turn for help in identifying and dealing with these inferential monsters discussed by Leamer? This question points to the need to distinguish two alternative views of how effective quantitative analysis can be in achieving valid inference: first, the perspective of mainstream quantitative methods in political science, which is at times insufficiently attentive to the difficulty of using quantitative tools; and second, perspectives drawn from statistical theory, which sometimes express serious warnings about these tools.

Mainstream quantitative methods are a subset of applied statistics. In the years before the publication of KKV, a central focus in political science methodology was the refinement and application of regression analysis and related econometric techniques. This body of work has been influential across several social science disciplines, and it is a major source of KKV's methodological advice. When commentators argue that KKV adopts a quantitative perspective, they should be understood as referring to mainstream quantitative methods in this sense. Chapter 2 above (e.g., table 2.1) seeks to provide a summary of KKV's quantitative tools.

The main point, for present purposes, is that mainstream quantitative

methods and important currents of thinking in statistical theory have adopted quite different perspectives on the feasibility of effectively eliminating rival hypotheses in observational studies through regression-based tools. Within political science, mainstream quantitative methods have been associated with the advocacy of quantitative approaches—treating them as a set of research tools that provide superior leverage in both descriptive and causal inference. We view KKV as a clear expression of such advocacy, a form of advocacy that is also strongly reflected in the standards for “good research” applied by many political science departments and disciplinary journals.

By contrast, according to arguments that can be made from the standpoint of statistical theory, the superiority of quantitative methods is less clear. Such statistical arguments place far greater emphasis on the many assumptions and preconditions required to justify the use of specific quantitative tools, suggesting that these tools may often be inapplicable in observational research.⁶ As emphasized above, more skeptical norms about inference are also fundamental to the work of Campbell.⁷

Statistical ideas quite distinct from those presented in KKV are also found in psychometrics and mathematical measurement theory (20, 129–31 this volume). These fields offer valuable insights into concepts, the foundations of measurement, the complex assumptions required in justifying higher levels of measurement, and the contextual specificity of measurement claims—insights that present a different picture than that offered by mainstream quantitative methodology.

Bayesian statistical analysis is likewise a relevant branch of statistical theory largely neglected by KKV,⁸ as McKeown (chap. 4 online) emphasizes. Ideas drawn from Bayesian analysis, which have recently come to be more widely used in political science methodology, provide tools for estimating uncertainty that are relevant for several problems of research design that KKV discusses.

For example, KKV argues that qualitative researchers are often better off

6. Important examples include Liu (1960), Leamer (1983), Dijkstra (1988), Manski (1995), McKim and Turner (1997), and Berk (2004). See also Lucas (1976); Cox (1977); Copas and Li (1997); Lang, Rothman, and Cann (1998); and Scharfstein, Rotnitzky, and Robins (1999). Within political science, work that reflects this broader statistical perspective includes Achen (1986, 2000, 2002), Bartels (1991), and Wallerstein (2000). Within sociology, relevant examples are Lieberston (1985), Goldthorpe (2001), and Ní Bhrolcháin (2001).

7. Campbell and Stanley (1963); Campbell and Ross (1968); Cook and Campbell (1979).

8. The authors of KKV (102 n. 13) state that they adopt a “philosophical Bayesian” approach; yet Bayesian analysis plays no discernible role in the book’s recommendations.

not working with random samples. Yet many of the book's statements in favor of estimating uncertainty would seem to rely on procedures for testing statistical significance originally designed for use with inferences from a random sample to a universe of cases. Unfortunately, extending significance tests to situations where the data are not a random sample from a larger universe may not be justified. As Freedman, Pisani, and Purves (2007: 556) put it, "[i]f a test of significance is based on a sample of convenience, watch out." While significance tests can be an appropriate way to handle forms of randomness other than sampling error, Greene (2000: 147) argues that standard interpretations of statistical significance tests in such situations require that the test statistic be random. When the data are a random sample, this requirement is automatically satisfied; it may not be met under other circumstances. Overall, scholars should heed Freedman and Lane's (1983) warning against using conventional significance tests as a general tool for estimating uncertainty.

Bayesian statistics definitely cannot solve all the problems of making descriptive and causal inferences with a nonrandom sample. Yet these tools do provide a framework for evaluating uncertainty that may sometimes allow researchers to incorporate more kinds of uncertainty, and more detailed information about the sampling process, than do traditional significance tests. Thus, while KKV's emphasis on estimating uncertainty is laudable, this goal might be better accomplished using insights based on a Bayesian perspective.

Another reason a Bayesian perspective may be relevant for thinking about small-N research is that it systematizes a research strategy noted briefly by KKV (211): overcoming the small-N problem by situating small-N findings within a larger research program. Bayesian ideas help in reasoning about the relation between the findings of prior research and the insights generated by any given small-N study. As we have argued above, Bayesian analysis also provides tools for evaluating arguments about necessary and sufficient causation (148–49 this volume), and thus specifically for improving the practice of qualitative research. In some of these situations, a full Bayesian framework, including formalization of prior beliefs about all parameters, may be quite useful. More generally, however, informal applications of the central Bayesian insight—that is, that inferences should be evaluated in light of the data *and* of prior knowledge—can provide a useful corrective to the sometimes inappropriate use of significance tests in causal inference.

Overall, from this wider perspective of statistical theory, the tools emphasized by KKV are properly seen as just one option—an option that perhaps needs to be approached with greater recognition of its limitations and of available alternatives. In order to further illustrate why such caution is needed, we now discuss two additional distinctions: between determi-

nate and indeterminate research designs, and between data mining and specification searches.

Determinate versus Indeterminate Research Designs

In discussing the challenge of eliminating rival explanations, KKV distinguishes between “determinate” and “indeterminate” research designs.⁹ The book designates as “determinate” those designs that meet the standards of: (a) having a sufficient N in relation to the number of explanatory parameters being estimated, and (b) avoiding the problem that two or more explanatory variables are perfectly correlated—that is, perfect multicollinearity (KKV 119, 150; see also 120).¹⁰ Meeting these standards gives the researcher stronger tools for adjudicating among rival hypotheses. By contrast, designs that fail to meet these standards are called “indeterminate” (118–24, 145, 228). Such designs do not consider enough data¹¹ to distinguish the causal impact of alternative independent variables, which is one aspect of the problem of unidentifiability (KKV 118 n. 1).¹² As a consequence, the data under consideration are compatible with numerous interpretations. KKV goes so far as to state: “[a] determinate research design is the sine qua non of causal inference” (116).¹³ By contrast, for research

9. This distinction, of course, involves quite different issues from the contrast between deterministic and probabilistic causation discussed in chapter 8.

10. KKV (122) uses the term “multicollinearity” in discussing this problem. The definition of multicollinearity that KKV offers is, however, stronger than most definitions of the term in statistics (see, e.g., Vogt 1999: 180). Therefore, we have used the term “perfect multicollinearity” in discussing this issue.

11. Perfect multicollinearity is a problem of insufficient data, in the sense that the analyst lacks data that can distinguish between the effects of two (or more) explanatory variables. Adding such data by finding cases in which the explanatory variables are not perfectly correlated would, of course, eliminate the perfect multicollinearity.

12. Unidentifiability also involves other important issues that KKV does not discuss. In structural equation modeling, problems of unidentifiability arise in several different situations. This problem arises if all the variables are endogenous because they appear as both independent and dependent variables within the same system of equations. In this case everything affects everything else, and there is no way of finding a “prime mover” to pin down causal relationships. It also arises if, for a particular endogenous variable of interest, there is no exogenous (i.e., truly independent) variable that affects only the endogenous variable directly (and there is no other identifying information). In this case the researcher has no way to isolate the endogenous variable’s impact on the other endogenous variables. These aspects of unidentifiability are key challenges in using statistical tools to address endogeneity and selection bias (Achen 1986: 38–39; Greene 2000: 663–76).

13. At a later point, KKV (150) does soften this statement by discussing ways in which a determinate research design can produce invalid inferences.

designs that are indeterminate, “virtually nothing can be learned about the causal hypotheses” (118).

The distinction between a determinate and an indeterminate research design relies on the standard idea of the power of statistical tests. Discussions about the power of a test are useful for focusing on the degree to which the analysis is capable of rejecting the null hypothesis when that hypothesis is in fact false, under the assumption that the model is correct and only random error is at stake. This is a useful, but narrow, idea.

Correspondingly, we find the distinction between determinate and indeterminate research designs somewhat misleading. It is true that researchers must think carefully about the size of the N , given that it is the principal source of leverage in dealing with the issue of sampling error. Yet the size of the N is hardly the only source of inferential leverage, and sampling error is certainly not the only challenge to causal inference. Correspondingly, KKV’s distinction gives these specific concerns too much weight.

Further, it seems particularly inappropriate to argue that a determinate research design in this sense is the sine qua non of causal inference, whereas an indeterminate design contributes little. This claim can be seen as reifying the small- N problem, in the specific sense that it establishes a vivid dichotomy, in relation to which the small- N researcher is always on the wrong side.

The strong contrast that KKV draws between determinate and indeterminate research designs runs the risk of obscuring the broader, and much more important, contrast between experimental and observational studies discussed above. From this broader point of view, all inferences drawn from observational data share fundamental problems of alternative explanations and misspecified models. These problems pose a much greater challenge to the validity of causal inference than the problem of insufficient data—above all the small- N problem—emphasized by the idea of a determinate research design. In the realm of observational studies, the conclusions drawn from research are always partial, uncertain, and dependent on meeting underlying analytical assumptions, as KKV (passim) acknowledges.

To put this another way, we find it problematic to suggest that any observational study can ever be “determinate,” given this term’s questionable implication that the “inferential monsters” to which Leamer refers can definitively be ruled out. We doubt they can. Further, if no observational research design is ever really determinate, then the concept of an indeterminate research design is also misleading when applied to observational studies. All such studies can be understood as involving indeterminate research designs. For this reason, we suggest avoiding the distinction between determinate and indeterminate research designs, while recognizing the issues

raised as an unavoidable aspect of the larger problem of identifiability in research design.

In addition, we are concerned that KKV's use of the label "determinate research design" focuses attention on issues of identifiability to an extent that implicitly advocates an inversion of what we see as the most productive relationship between theory and testing. Avoiding multicollinearity and large numbers of explanatory variables vis-à-vis the N are obviously important for regression analysis, and such issues should be a concern in small- N analysis as well. However, an excessive focus on these objectives may push analysts toward redesigning theory to be conveniently testable, instead of searching for more rigorous tests of the theories that scholars actually care about.

We would argue that, in situations where researchers are trying to test well-developed theories against clear alternative explanations, adopting an approach to testing that first requires modifications of the theories in question gives up a lot. In such circumstances, it is usually best to establish the testing requirements in light of the theory and the relevant alternative explanations: only in this way can we effectively adjudicate among these alternatives. If a hypothesis is difficult to test against the relevant alternative hypotheses with the existing data, then the best approach is to find new data and new approaches to testing, not to modify the hypotheses until it is easy to test them. Hence, to reiterate, the term "determinate" emphasizes the standards of identifiability and statistical power in a way that can distract analysts from testing the theories that often motivate research to begin with.

Rather than evaluating research designs as being determinate or indeterminate, it may be more productive to ask a broader question: Are the findings and inferences yielded by a given research design *interpretable*,¹⁴ in that they can plausibly be defended? The interpretability of findings and inferences can be increased by many factors, including a larger N , a particularly revealing comparative design, a rich knowledge of cases and context, well-executed conceptualization and measurement, or an insightful theoretical model. If the research question has been modified in order to make it more testable, then the findings may be less interpretable in relation to the original research question, and inferential leverage has probably been lost, not gained. This focus on interpretable findings broadens KKV's idea of a determinate research design by recognizing multiple sources of inferential leverage.

Data Mining versus Specification Searches

Many researchers seek to evaluate competing explanations through intensive analysis of their data; however, this practice often raises the con-

14. See, for example, Stone's (1985: 689) discussion of interpretability as a central characteristic of statistical models.

cern that researchers have engaged in “data mining” (KKV 174) or “data snooping” (Freedman, Pisani, and Purves 2007: 547) and have thereby exhausted the inferential leverage provided by the data. If researchers try out enough different combinations of explanatory variables, they will eventually find one that fits the data—even if the data are random.¹⁵ Data mining is therefore seen as an undesirable research practice that weakens causal inference. Concerns about different forms of this problem recur in the guidelines, presented in chapter 2 above, that summarize KKV’s framework. Guideline no. 27 is concerned with the problem that researchers run “regressions or qualitative analyses with whatever explanatory variables [they] can think of” (KKV 174). No. 34, the injunction to test theory with data other than that used to generate the theory, and no. 35, the recommendation that theory should generally not be reformulated after analyzing the data, also address concerns related to data mining.

We find it striking that the related, partially inductive, econometric practice of “specification searches” is, by contrast, viewed favorably by methodologists as an unavoidable step in making causal inferences from observational data. The literature on specification searches has proposed systematic approaches to the iterated process of fitting what are inevitably incomplete models to data. The main ideas in this literature implicitly point to the dilemma that treating these inductive practices as a problem can be misleading, if not counterproductive, in establishing criteria for good research. Such a dilemma can be seen, first of all, in quantitative research that uses complex explanatory models. In the social sciences, such models are virtually never sufficiently detailed to tell us *exactly* what should be in the regression equation. Scholars who wish to test these models are forced to make decisions about the underspecified elements of the model and, in actual practice, they almost never stop after running the first regression that seems reasonable to them. It is the myth that these multiple tests do not occur that leads Leamer to worry about “the fumes which leak from our computer labs” (1983: 43). Rather than pretending that they do not occur, Freedman, Pisani, and Purves specifically urge analysts to report “how many tests they ran before statistically significant [results] turned up” (2007: 547).

Because we usually do not know the correct specification of a model, stopping with the first specification is methodologically problematic, just as it would be unjustified to stop with the specification that most favors the

15. Thus, if a researcher who is running bivariate regressions successively regresses a purely random dependent variable on each of one hundred purely random independent variables, on average five of the resulting bivariate relationships will be statistically significant at the .05 level. This is true by the definition of significance tests.

working hypothesis. The methodology of specification searches is concerned with systematic procedures for deciding where to start, when to stop, how to report the steps in between, and when we should believe the results of this overall process. Some scholars present elaborate justifications for beginning with the simplest plausible model and then engaging in “fragility testing” or “sensitivity analysis” by adding variables that may change the coefficients of interest (Leamer 1983: 40–42; 1994 [1986]; Levine and Renelt 1992). Other scholars work from the other side: they begin with the most elaborate plausible model and eliminate elements of the model that prove to have little explanatory power (Hendry 1980; Hendry and Richard 1982; White 1994; see Granger 1990 for statements from both sides of this debate). These two approaches both use induction to test the plausibility of findings under divergent sets of methodological assumptions. The specification searches literature thus takes a position on induction that is radically different from the simple mandate not to reformulate theory after looking at the data.

The idea of specification searches is, of course, just one facet of a much larger concern with the inductive component of research. Both quantitative and qualitative researchers routinely adjust their theories in light of the data—often without taking the further step of moving to new data sets in order to test the modified theory. Whether this inductive component involves completely overturning previous models or refining them in the margins, such inductive practices are widely recognized as an essential part of research. For example, Ragin and Munck (chaps. 2 and 3 online) devote extensive attention to procedures for inductive analysis.

To conclude, data mining can certainly be a problem. Yet the misleading pretense that they are not routinely utilized, and even worse, the indiscriminate injunction against inductive procedures, is at least as big a problem in social research.

Conditional Independence or the Specification Assumption

Two alternative formulations of key assumptions underlying causal inference are the assumption of conditional independence and the specification assumption. The issue here is how to conceptualize and label the set of assumptions used to justify causal inference based on observational data. Rather than conceptualizing the most important of these several assumptions in terms of conditional independence—the concept employed by KKV—we find it productive to frame these issues in terms of the specification assumption. In discussing the choice between these alternative overarching concepts, it is essential to recognize that they are fundamentally similar. Given this similarity, this section conveys a suggestion, and simul-

taneously sounds a note of caution, about the focus and emphasis entailed in these alternative assumptions.

Our basic point in the discussion that follows is that, while the assumption of conditional independence is rooted in an analogy to experiments, the specification assumption more directly reflects the situation of a researcher seeking to analyze observational data. For this reason, we find the specification assumption to be more helpful—at the same time that we recognize the underlying similarities between the two assumptions.

As discussed in greater depth in chapter 2, the assumption of conditional independence builds on an analogy involving a counterfactual understanding of causation and treats every causal inference as a partial approximation of an ideal experiment. For the purpose of explicating the contrast with the specification assumption, in this section we briefly summarize conditional independence. We begin by discussing the basic thought experiment behind the idea of conditional independence, which serves as the foundation for introducing the assumption of “independence of assignment and (potential) outcomes.” We use this assumption in defining *conditional* independence, and we then discuss why it is particularly relevant for observational studies. In comparison with the discussion in chapter 2, our goal here is particularly to discuss the range of issues that are highlighted by these conceptualizations, rather than to present the more general framework they represent.

The assumption of conditional independence posits that each case can be understood as having a value (which may or may not actually be observed—hence, this is in effect a hypothetical variable) on an outcome variable, Y_t , that reflects the outcome that case would experience if given an experimental *treatment*; and likewise a value (which, again, may or may not be observed) on a second variable, Y_c , that reflects the outcome the case would experience if it were the *control* in an experiment. The causal effect of the treatment relative to the control for this case is the (hypothetical) difference between its values on these two variables.

In the real world, even in randomized experiments, the value of only one of these variables can actually be observed for each case at any point in time. Through some process (i.e., through randomization in experiments, or, in an observational study, through a real-world process that may or may not be known to the researcher), any given case is, in effect, assigned either the treatment, or the control. A given case cannot simultaneously be assigned to both. For example, an individual can either be exposed to a political message, or not be exposed to it; or a democratic country can either use proportional representation to elect its officials, or use some other electoral method.

Because we cannot empirically observe what would have happened to the same individual or country at any one point in time both with and

without the treatment, causal inference routinely relies on real-world comparisons of cases that receive the treatment with other cases that do not receive the treatment. The comparison of these observed treated cases with the observed control cases substitutes for the hypothetical comparison of each case with and without the treatment. Comparing two real-world groups of cases that do and do not receive the treatment yields a good causal inference, provided that these two groups are similar in the sense that both have the same mean values of the (hypothetical) variable Y_t , and also the same mean values of the (hypothetical) variable Y_c . With a large enough sample, randomization of assignment, as in a well-designed experiment, ensures that this condition will be met.

With observational data, however, this standard, which is called *independence* of assignment and outcome,¹⁶ is usually not met. Furthermore, there is no way to test whether independence is satisfied—because only Y_t or Y_c , but not both, is observed for each case. Although we can calculate the mean value of Y_t for the cases that are actually assigned to the treatment, we cannot do so for the cases assigned to the control. Similarly, although we can calculate the mean value of Y_c for the cases assigned to the control, we cannot do so for the cases that are assigned to the treatment. Consequently, we cannot know if the treatment cases would have had the same average on Y_c (if they had been assigned to the control) as the cases that were actually assigned to the control. Further, we cannot establish whether the control cases would have had the same average on Y_t (if they had been assigned to the treatment) as the cases that were actually assigned to the treatment. In short, no test will allow us to establish whether the standard of independence holds for a given set of cases.

The assumption of conditional independence becomes relevant if this criterion of independence is not met. Conditional independence means that there is another variable or set of variables, which serve as “statistical controls,” such that by controlling for—or *conditioning* on—these variables, the treatment group and the control group come to have the same mean values on both Y_t and Y_c . If the researcher uses quantitative techniques that control for these variables, such as stratification,¹⁷ conditional indepen-

16. More precisely, as noted in chapter 2, this standard in fact involves mean independence of assignment and outcome, and the standard of conditional independence of concern here is mean conditional independence of assignment and outcome.

17. Regression analysis employs assumptions that some readers may view as similar to the assumption of conditional independence, in that these assumptions stress the importance of control variables in causal inference. At a general level, this understanding is probably adequate; however, it is important to remember that analytic techniques (e.g., stratification versus regression) differ, sometimes substantially, in the details of the assumptions they depend on.

dence is thereby satisfied and an important criterion for good causal inference has been met. In effect, by introducing statistical controls into the analysis and then assuming conditional independence, the researcher turns the observational study into something akin to an experiment. However, it is obviously vital to remember that the assumption of conditional independence, like the assumption of independence, is hard to test.

Unlike conditional independence, which is rarely mentioned in econometrics textbooks, the specification assumption is frequently discussed in econometric and statistical work on regression analysis.¹⁸ The specification assumption has the major advantage that it starts with what is typically the actual situation of the researcher—that is, having an explanatory model of unknown usefulness—and then specifies the criteria that must be met to move in the direction of causal inference. The name of this assumption refers directly to this process of specification.

Thus, the starting point for the specification assumption is not the metaphor of an experiment, but rather the model that researchers use to organize their hypotheses. In the simplest case, this model consists of a dependent variable and a set of independent variables in a single regression equation. More generally, it may explicitly include an equation for the process of assignment to treatment, as well as for the outcome variable. The specification assumption focuses attention on what must be true—concerning the relationships between the included explanatory variables and the unobserved error terms in the model—in order to make unbiased inferences about the strength of the associations predicted by these relationships.

In the context of a regression model, the specification assumption is the claim that the included independent variables are statistically unrelated to the error term that derives from a (hypothetical) comparison between the regression model and the true causal equation.¹⁹ One major threat to the

18. See, e.g., Greene (2000: 219–20); Kennedy (1998: chaps. 3 and 5); Mirer (1995); Darnell (1994: 369–73); Gujarati (1988: 57–60, 166, 178–82); and Wonnacott and Wonnacott (1979: 413–19). Treatments by political scientists include Achen (1982: chap. 5; 1986: 12, 27); and Hanushek and Jackson (1977: 79–86). For a highly accessible statement, see Vogt (1999: 271–72). Stone (1993) discusses the relationships among the specification assumption (which he calls “no confounding”), conditional independence, and mean conditional independence (which he calls “no mean effect”).

19. The specification assumption as defined here is sometimes confused with the much weaker assumption that the expectation of the residuals in a regression analysis is zero, conditional on the included variables. This second assumption, which is *not* the specification assumption, focuses on whether the included right-hand side variables successfully capture all predictive information that these variables provide about the dependent variable. For example, the heights of sisters can provide an excellent prediction of their brothers’ heights even though the correlation is causally

specification assumption is omitting variables that ought to be included—and therefore relegating the effects of those variables to the error term, sometimes producing missing variable bias (the central, direct concern of conditional independence). A second major threat is including variables that are endogenous, that is, are statistically related to the part of the dependent variable that is not caused by the included variables. Including such variables that have a direct connection with the error term yields endogeneity bias. When a model has either of these problems, the estimated causal effects of the included variables will be biased because the included variables will stand in for (or proxy for) either missing variables or the error term.

A further benefit of discussing these issues in terms of the specification assumption—in addition, as noted above, to focusing attention more directly on the actual situation of the researcher—is that this term is directly linked to other standard methodological labels: model specification, specification error, specification analysis, the specification problem, misspecification, and specification searches.

While we believe that the framework of the specification assumption brings basic issues of causal inference into sharper focus, it also has a major limitation—which it shares with the assumption of conditional independence. Both assumptions are hard to test, and no analyst can ever prove that an observational study meets either assumption. Leamer's inferential monsters may always be lurking beyond the researcher's immediate field of vision. This is one of the reasons why, in order to supplement correlation-based causal inference, scholars turn to alternative sources of inferential leverage such as experiments or causal-process observations.

To reiterate the point made at the start of this section, our argument here is neither that the assumption of conditional independence is misleading in any fundamental sense, nor that meeting the specification assumption solves all problems of causal inference. Rather, we believe that the analogy

spurious. Because no causal connection is implied by this assumption, researchers can always meet this standard without introducing additional right-hand side variables (although they may have to add nonlinear transformations of the included variables).

By contrast, the specification assumption means that there is no statistical relationship between the included independent variables and any excluded variables that causally affect the dependent variables. Often, meeting this assumption would require analysts to include more independent variables. Thus, in a regression equation that predicts brothers' heights from sisters' heights, the specification assumption fails because there is a correlation between the sisters' heights—the included independent variable—and the parents' heights, excluded variables that causally affect brothers' heights. Only by including these missing variables can the researcher meet the specification assumption.

behind conditional independence may focus too much attention on control variables as a solution to problems of causal inference based on observational data. By contrast, the specification assumption focuses more directly on problems of endogeneity and misspecified relationships *among* measured variables, as well as other inadequacies of our causal models.

Taken together, our observations about these five distinctions considered in this section help to spell out the perspective on causal inference that we have adopted, which clearly differs from that of KKV. We now turn to some additional distinctions that help to develop further our overall argument about sources of leverage in causal inference: qualitative versus quantitative research, cases versus observations, and data-set observations versus causal-process observations.

FOUR APPROACHES TO THE QUALITATIVE VERSUS QUANTITATIVE DISTINCTION

Debates about sources of leverage for eliminating rival explanations in causal inference—and obviously also about tools for descriptive inference—are routinely framed in terms of the relative strengths and weaknesses of qualitative and quantitative research. Yet this distinction needs to be disaggregated if it is to play a useful role in thinking about research design. In conjunction with this distinction, we do not find two neatly bounded categories, but rather four overlapping categories (see table 9.1). However, notwithstanding this complexity, it is still useful for many purposes to use the dichotomous labels of qualitative versus quantitative.

Level of Measurement

One distinction concerns the level of measurement. Here we find ambiguity regarding the cut-point between qualitative and quantitative, and also contrasting views of the leverage achieved by different levels of measurement. Some scholars label data as qualitative if it is organized at a nominal level of measurement and as quantitative if it is organized at an ordinal, interval, ratio, or other “higher” level of measurement (Vogt 1999: 230). Alternatively, scholars sometimes place the qualitative-quantitative threshold between ordinal and interval data (Porkess 1991: 179). This latter cut-point is certainly congruent with the intuition of many qualitative researchers that ordinal reasoning is central to their enterprise (Mahoney 1999: 1160–64). With either cut-point, however, quantitative research is routinely associated with higher levels of measurement.

Higher levels of measurement are frequently viewed as yielding more analytic leverage, because they provide more fine-grained descriptive differ-

entiation among cases. However, higher levels of measurement depend on complex assumptions about logical relationships—for example, about order, units of measurement, and zero points—that are sometimes hard to meet. If these assumptions are not met, such fine-grained differentiation can be illusory, and qualitative categorization based on close knowledge of cases and context may in fact provide more leverage. In any case, careful categorization is a valuable, indeed essential, analytic tool.

Size of the N

A second approach is to identify the qualitative-quantitative distinction with the contrast between small-N and large-N research. Here we will treat the question of the “N” as a relatively straightforward matter involving the number of observations on the main dependent variable that the researcher seeks to explain, understood at the level of analysis that is the principal focus of the research.²⁰ In a subsequent section, we will explore the complex issues that can arise in establishing the N.

The N involved in a paired comparison of Japan and Sweden, or in an analysis of six military coups, would routinely be identified with the qualitative tradition. By contrast, an N involving hundreds or thousands of observations would routinely be identified with the quantitative tradition. Although there is no well-established cut-point between qualitative and quantitative in terms of the N, such a cut-point might be located somewhere between ten and twenty.

However, some studies definitely break the methodological stereotypes: that is, those with a larger N that in other respects adopt a qualitative approach; as well as those with a relatively small N that in other respects adopt a quantitative approach. Examples of qualitative studies which have a relatively large N include Rueschemeyer, Stephens, and Stephens’s (1992) *Capitalist Development and Democracy* (N = 36), Tilly’s (1993) *European Revolutions, 1492–1992* (hundreds of cases), and R. Collier’s (1999) *Paths toward Democracy* (N = 27). Wickham-Crowley’s (1992) *Guerillas and Revolution in Latin America* focuses on twenty-six cases: he carries out a qualitative/narrative analysis, based on detailed discussion of thirteen cases, and he analyzes thirteen additional cases using dichotomous/categorical variables and Boolean methods.

Some studies that rely heavily on statistical tests in fact have a smaller N than these qualitative studies. Examples are found in the literature on advanced industrial countries: a study with an N of eleven focused on the impact of partisan control of government on labor conflict (Hibbs 1987);

20. Obviously, the unit of analysis, as well as the number of cases being studied, may change in the course of research.

Table 9.1. Four Approaches to the Qualitative-Quantitative Distinction

<i>Approach</i>	<i>Defining Distinction</i>	<i>Comment</i>
1. Level of Measurement	Cut-point for qualitative vs. quantitative is nominal vs. ordinal scales and above; alternatively, nominal and ordinal scales vs. interval scales and above.	Lower levels of measurement require fewer assumptions about underlying logical relationships; higher levels yield sharper differentiation among cases, provided these assumptions are met.
2. Size of the N	Cut-point between small N vs. large N might be somewhere between 10 and 20.	A small N and a large N are commonly associated with contrasting sources of analytic leverage, which correspond to the third and fourth criteria below.
3. Statistical Tests	In contrast to much qualitative research, quantitative analysis employs formal statistical tests.	Statistical tests provide explicit, carefully formulated criteria for descriptive and causal inference; a characteristic strength of quantitative research. Yet this again raises question of meeting relevant assumptions.
4. Thick vs. Thin Analysis^a	Central reliance on detailed knowledge of cases vs. more limited knowledge of cases.	Detailed knowledge associated with thick analysis is likewise a major source of leverage for inference; a characteristic strength of qualitative research.

^a This distinction draws on Coppedge's (1999) discussion of thick versus thin concepts. See also note 22 in the text below.

and studies with an N of fifteen focused on the influence of corporatism and partisan control on economic growth (Lange and Garrett 1985, 1987; Jackman 1987, 1989; Hicks 1988; and Hicks and Patterson 1989; Garrett 1998). Likewise, quantitative research that seeks to forecast U.S. presidential and congressional elections routinely employs an N of eleven to thirteen (e.g., Lewis-Beck and Rice 1992; J. Campbell 2000; Bartels and Zaller 2001). Choices about the N are thus at least partially independent from choices about other aspects of a qualitative or quantitative approach.

Scholars decide on the N according to many different criteria, including the availability of analytically relevant data and a concern with the alternative sources of inferential leverage associated with a small N and a large N. The third and fourth criteria for qualitative versus quantitative, presented below, address these alternative sources of leverage.

Statistical Tests

The third approach focuses on the use of statistical tests.²¹ An analysis is routinely considered quantitative if it employs statistical tests in reaching its descriptive and explanatory conclusions. By contrast, qualitative research typically does not employ such tests. While the use of statistical tests is generally identified with higher levels of measurement, the two are not inextricably linked. Quantitative researchers frequently apply statistical tests to nominal variables. Conversely, qualitative researchers often analyze data at higher levels of measurement without utilizing statistical tests. For example, in the area studies tradition, a qualitative country study may make extensive reference to ratio-level economic data.

Statistical tests are a powerful analytic tool for evaluating the strength of relationships and important aspects of the uncertainty of findings in a way that is more difficult in qualitative research. Yet, as with higher levels of measurement, statistical tests are only meaningful if complex underlying assumptions are met. If the assumptions are not met, alternative sources of analytic leverage employed by qualitative researchers may in fact be more powerful.

Thick versus Thin Analysis

Finally, we distinguish between “thick” and “thin” analysis.²² Qualitative research routinely utilizes *thick analysis*, in the sense that analysts place great reliance on a detailed knowledge of cases. Indeed, some scholars consider thick analysis the single most important tool of the qualitative tradition. One type of thick analysis is what Geertz (1973) calls “thick description,” that is, interpretive work that focuses on the meaning of human behavior to the actors involved. In addition to thick description, many forms of detailed knowledge, if utilized effectively, can greatly strengthen description and causal assessment.²³ By contrast, quantitative researchers routinely

21. We intend the present usage of “statistical tests” somewhat broadly, including techniques of parameter estimation as well as tools of statistical inference.

22. This distinction draws on Coppedge’s (1999) discussion of thick versus thin concepts. Neither our distinction nor that of Coppedge should be confused with Geertz’s (1973) distinction between “thick description,” which focuses on the meaning of human behavior to the actors involved, as opposed to “thin description,” which is not concerned with this meaning. With the expression “thick analysis,” we mean research that focuses closely on the details of cases. These details may or may not encompass subjective meaning. In this sense, Geertz’s thick description, and also constructionism, is a specific type of what we call thick analysis.

23. This should not be taken to imply that researchers pursuing the goal of thick description must always use tools of thick analysis. For example, survey researchers may seek to gain insights into the subjective meaning of respondents’ behavior, at

rely on *thin analysis*, in that their knowledge of each case is typically far less complete. However, to the extent that this thin analysis permits them to focus on a much larger N, they may benefit from a broader comparative perspective, as well as from the possibility of using statistical tests. Whereas the precision and specificity of statistical tests are a distinctive strength of quantitative research, the leverage gained from thick analysis is a characteristic strength of qualitative research.

The distinction between thick and thin analysis is closely related to Ragin's (1987) discussion of case-oriented versus variable-oriented research. Of course, qualitative researchers do think in terms of variables, and quantitative researchers do deal with cases. The point is simply that qualitative researchers are more often immersed in the details of cases, and they build their concepts, their variables, and their causal understanding in part on the basis of this detailed knowledge. Such researchers seek, through their in-depth knowledge of cases, to carefully rule out alternative explanations until they come to one that stands up to scrutiny. Detailed knowledge of cases does sometimes play a role in quantitative research. Indeed, some quantitative research employs thick analysis. However, in-depth knowledge is far more common in qualitative research and much less common among quantitative researchers, who tend to rely on statistical tests.

Drawing Together the Four Criteria

As this section illustrates, there is no single, sharp distinction that consistently differentiates qualitative and quantitative research—and that unambiguously sorts out the most important sources of inferential leverage. We would certainly classify as qualitative a study that places central reliance on nominal categories, focuses on relatively few cases, makes little or no use of statistical tests, and places substantial reliance on thick analysis. By contrast, a study based primarily on interval- or ratio-level measures, a large N, statistical tests, and a predominant use of thin analysis is certainly quantitative. Both types of study are common, which is why it makes sense, for many purposes, to maintain the overall qualitative-quantitative distinction.

However, an adequate discussion of inferential leverage requires careful consideration not only of these polar types, but also of the intermediate alternatives. For example, a particularly strong form of inferential leverage may be gained by combining statistical tests with thick analysis, bringing together their complementary logics in what may be called “nested inference.”²⁴ This relationship between qualitative and quantitative methods is

the same time that they may have a selective and in some ways superficial overall level of knowledge about each respondent.

24. This term is adapted from Coppedge's (2001) “nested induction” and from Lieberman's (2003a) “nested analysis.”

very different from that proposed by KKV, because with nested inference the characteristic strengths of each approach supplement and enhance research based on the other approach.

CASES VERSUS OBSERVATIONS

Well-understood definitions of “case” and “observation” are essential in discussing sources of inferential leverage in qualitative and quantitative research, yet finding adequate definitions of these terms is a serious challenge. Indeed, the question “what is a case?” is the title of an entire book (Ragin and Becker 1992).

Cases

We understand a case as one instance of the unit of analysis employed in a given study. Cases correspond to the political, social, institutional, or individual entities or processes about which information is collected. For example, the cases in a given study may be particular nation-states, social movements, political parties, trade union members, or episodes of policy implementation. The number of cases is conventionally called the “N.”

It is productive to think about cases in relation to a “rectangular data set”—that is, a matrix or uniform array of data in which the rows correspond to cases and the columns correspond to variables. The pieces of data aligned in a single row in the data set pertain to a particular case, and the number of rows corresponds to the number of cases (the N). The pieces of data aligned in a single column in the data set pertain to a particular variable, and the number of columns corresponds to the number of variables. The information in a rectangular data set may be either quantitative or qualitative—that is, it may consist of scores on variables at any level of measurement.

Observations

We now present a definition of the term observation that serves to underscore the importance of this second, horizontal slice. “Observation,” of course, has a commonsense meaning: it is an insight or piece of information recorded by the researcher about a specific feature of the phenomenon or process being studied. This usage is widespread, and it is found, for example, in KKV (57). In the language of variables, an observation in this sense is a single piece of data that constitutes the value of a variable for a given case. The commonsense meaning also includes other kinds of information that might not conventionally be thought of as a score on a vari-

able—for example, information about context that makes the phenomenon under study intelligible and that helps the researcher avoid basic mistakes in interpreting it.

A fundamentally different meaning of observation, which is standard in quantitative analysis, refers to a row in a rectangular data set. According to this meaning, an observation is the collection of scores for a given case, on the dependent variable *and* all the independent variables (KKV 117; also 53, 209). In other words, an observation is “all the numbers for one case,” that is, all the scores within any given row of the data set. In relation to this definition of observation, a “case,” which also corresponds to a row in the data set, should be understood as the larger setting from which the numbers in each row are drawn.²⁵

The second definition may initially seem counterintuitive for scholars not oriented toward thinking about rectangular data sets and matrix algebra. Whereas the commonsense meaning of observation refers only to one score, this second meaning involves two or more scores. A useful way of clarifying this second usage is to think about it as a “data point,” which in a two-dimensional scatterplot corresponds to the scores of the independent and dependent variables. The data point is an observation whose meaning depends on simultaneously considering the scores for both variables.²⁶ The cluster of information contained in a data point plays a central role in causal inference by focusing our attention simultaneously on the scores for the independent and dependent variables. This same idea can be extended to the analysis of more than two variables (as in scatterplots with three or more dimensions), and the purpose of this second definition of observation is to highlight that central inferential role. As with the rectangular data set, the data entailed in an observation of this type may be either quantitative or qualitative.

This second meaning of observation serves a useful methodological purpose. For example, it can clarify the meaning of the well-known “many-variables, small-N problem” (Lijphart 1971: 685–91). In debates on methodology, increasing the number of observations is routinely understood as a basic solution to this problem. Obviously, the content of this recommendation depends on our definition of an observation. For instance, if we score the cases on an additional variable, we add observations in the sense

25. KKV (52–53, 117–18, 217–18) makes a parallel distinction between case and observation. While the book mainly uses observation in the sense of data-set observation, see also KKV (57), which refers to observation as a score.

26. The term “data point” is also sometimes used informally to mean the score for a given variable on a given case (Vogt 1999: 71). However, for any scholar who has worked with scatterplots, the meaning given in the text above more directly conveys the intuitive idea of a data point in a scatterplot.

of the ordinary language usage noted above—that is, we introduce one new piece of data for each case. However, adding a variable generally makes the many-variables, small-N problem worse, because it reduces the degrees of freedom. In this sense, increasing the number of observations does not help the problem concerning the degrees of freedom.

By contrast, using the second definition of observation, it makes sense to say that increasing the number of observations addresses the many-variables, small-N problem. Adding observations—in the sense of adding “all the numbers” for one or more new cases—increases the number of rows in the matrix.

This usage thus clarifies a basic piece of methodological advice. At various points in the present volume, we argue that “increasing the number of observations,” as KKV frequently recommends, may not always be a good idea. However, taking one position or the other on this issue makes little sense as long as there is ambiguity about whether one is referring to adding “pieces of data” or adding cases to the analysis.

Given that it is confusing when the same term carries two meanings, we adopt the following usage. When we mean observation in the first, commonsense usage discussed above, we refer to a score, or to a piece of data or information. To highlight the second meaning of observation, we propose the expression “data-set observation.”

DATA-SET OBSERVATIONS VERSUS CAUSAL-PROCESS OBSERVATIONS

We thus introduce the label “data-set observation” to refer to observation in the sense of a row in a rectangular data set. At the same time, we do not want to lose sight of the critical role played in causal inference by information that is not part of a row in a data set. We therefore introduce the expression “causal-process observation” to emphasize the role such pieces of information play in causal inference (table 9.2). Whereas data-set observations lend themselves to statistical tests within the framework of what we have called “thin analysis,” causal-process observations offer an alternative source of inferential leverage through “thick analysis,” as discussed above.

A causal-process observation is an insight or piece of data that provides information about context or mechanism and contributes a different kind of leverage in causal inference. It does not necessarily do so as part of a larger, systematized array of observations. Thus, a causal-process observation might be generated in isolation or in conjunction with many other causal-process observations—or it might also be taken out of a larger data

Table 9.2. Data-Set Observation versus Causal-Process Observation

	<i>Data-Set Observation</i>	<i>Causal-Process Observation</i>
Corresponding Root Meaning of “Observation”	Standard quantitative/statistical meaning. Thus, all the scores for a given case; a row in a rectangular data set.	Ordinary language meaning. Thus, a piece of data or information; a datum.
Contribution to Causal Inference	The foundation for correlation-based causal inference. Provides the basis for tests of overall relationships among variables.	The foundation for process-oriented causal inference. Provides information about mechanism and context.

set. In the latter case, it yields inferential leverage on its own.²⁷ In doing so, a causal-process observation may be like a “smoking gun.” It gives insight into causal mechanisms, insight that is essential to causal assessment and is an indispensable alternative and/or supplement to correlation-based causal inference.

Part of the contrast between data-set observations and causal-process observations is that these two expressions utilize different root meanings of the term “observation” (table 9.2). Because the idea of “observation” is so closely tied in the minds of many quantitatively oriented scholars to data in a rectangular matrix, we might have chosen the expression “causal-process *information*.” However, we deliberately introduce the expression “causal-process *observation*” to emphasize that this kind of evidence merits the same level of analytic and methodological attention as do “data-set *observations*.”

While we can distinguish these two types of observations, we also find connections between them. For example, a scholar who has discovered a fruitful causal-process observation in one case—involving, for example, a causal mechanism that links two variables—might then proceed to systematically score many cases on this same analytic feature and add the new scores to an existing collection of data-set observations. Thus, the discovery of a causal-process observation can motivate the systematic collection of new data. Alternatively, a researcher who has done an analysis based on data-set observations may turn to causal-process observations to provide evidence about causal mechanisms. Thus, inference may be strengthened by movement in either direction.

27. Knowledge about the place of a causal-process observation within a larger data set can certainly influence how a scholar interprets this observation. Yet that is a different matter from relying on covariation within the data set to make causal inferences. And of course, causal-process observations are routinely studied in conjunction with an analysis of data-set observations based on such covariation.

The idea of causal-process observations is intended to make explicit the source of leverage in causal inference that lies at the heart of a long tradition of within-case analysis in qualitative research, a tradition discussed above by Rogowski and Tarrow and also in the online chapters by Collier, Mahoney, and Seawright, Munck, and McKeown. As discussed in the first of the online chapters, this tradition dates back at least to the 1940s and has, over the years, employed a number of different labels in the effort to pinpoint the distinctive analytic leverage offered by this approach. Recent writing on “mechanisms” is a valuable extension of this tradition.²⁸

Although the role of causal-process observations in qualitative research may be fairly obvious, their contribution to quantitative work should be underscored. Goldthorpe (2001), developing a line of argument that explicitly builds on the work of statisticians,²⁹ pinpoints this contribution in his important article “Causation, Statistics, and Sociology.” He uses the label “generative process” in referring to the linkage mechanisms that play an essential role in giving causal interpretations to quantitative associations. Goldthorpe contrasts this focus on generative processes with attempts to demonstrate causation through experiments or regression models.

This idea of causation [that] has been advanced by statisticians does not . . . reflect specifically [quantitative] thinking. It would appear to derive, rather, from an attempt to specify what must be *added* to any [quantitative] criteria before an argument about causation can convincingly be made. (Goldthorpe 2001: 8)³⁰

This procedure assumes that in quantitative analysis, an association

is created by some “mechanism” operating “at a more microscopic level” than that at which the association is established. In other words, these authors would alike insist . . . on tying the concept of causation to some process existing in time and space, even if not perhaps directly observable, that actually

28. Among many authors, see Elster 1999: chap. 1; McAdam, Tarrow, and Tilly 2001: chaps. 1–3; Tilly 2001.

29. Goldthorpe (2001: 8–9) cites various authors who have embraced this perspective, including Hill (1991 [1937]); Simon and Iwasaki (1988); Freedman (1991, 1992a, b); Cox (1992); and Cox and Wermuth (1996). See also Rosenbaum (1984).

30. In this and the following block quotation, the word “statistical” has been replaced (in brackets) by the word “quantitative.” The goal is to make clear the extent to which Goldthorpe’s argument converges with the argument of the present volume. Specifically, Goldthorpe is using ideas from statistical theory to argue that findings from the branch of applied statistics that we are calling mainstream quantitative methods analysis must be supplemented by qualitative insights.

generates the causal effect of X on Y and, in so doing, produces the [quantitative] relationship that is empirically in evidence. . . . [This mechanism can] illuminate the “black boxes” left by purely [quantitative] analysis. . . . (Goldthorpe 2001: 9)³¹

We see a sharp contrast between (a) Goldthorpe’s assertion that inference based on causal-process observations does *not* involve the approach of what we are calling mainstream quantitative methods; and (b) KKV’s approach, which explicitly seeks to subordinate this form of causal inference to its quantitative framework. KKV argues, in discussing the inferences drawn from “process tracing” (226), “historical analysis,” and “detailed case studies” (86), that these inferences must be treated through the framework for inference discussed throughout their book (85–87; see also 226–28). King, Keohane, and Verba reemphasize this point in chapter 7 above (111, 121–22 this volume). Yet KKV’s framework is designed for analyzing data-set observations and not causal-process observations, and the book’s recommendations therefore effectively treat causal-process observations as if they were data-set observations.

Our point, by contrast, is that causal-process observations offer a *different* approach to inference. Causal-process observations are valuable, in part, because they can fill gaps in conventional quantitative research. They are also valuable because they are an essential foundation for qualitative research. One goal of the present discussion is to strengthen the methodological justification for that foundation. Because inferences based on data-set and causal-process observations are fundamentally different, one promising direction of research is to combine the strengths of both types of observation within a given study. In the present volume, Tarrow presents an invaluable inventory of practical suggestions for how this may be accomplished.³² We would call attention to two of Tarrow’s techniques, which he labels “sequencing qualitative and quantitative research” and “triangulation.”³³ These utilize the distinctive strengths of alternative tools for data collection and inference. Tarrow (107 this volume) cites research on Poland’s Solidarity Movement as an example of the kind of fruitful

31. Goldthorpe goes on to point out that these efforts to establish causation “can never be taken as definitive” and must always be open to further empirical testing. “[F]iner-grained accounts, at some yet deeper level, will in principle always be possible” (2001: 9). Note that the quotation in the text above is in part Goldthorpe’s summary of arguments made by these statisticians, but Goldthorpe clearly intends this as a statement of his own position.

32. See also Bennett and George (1997a); Wallerstein (2001); and *APSA-CP* (2003).

33. King, Keohane, and Verba (121–22 this volume) conclude their chapter by endorsing a related concept of triangulation.

exchange that may take place between analysts using data-set observations and others relying on causal-process observations. Tarrow also points to the complementarities that result when elements of both approaches are combined in a given study.

In sum, both data-set observations and causal-process observations can play a role in both qualitative and quantitative research. The rich causal insights that qualitative researchers may gain from thick analysis can often be supplemented by systematic cross-case comparison using data-set observations, statistical tests, and thin analysis. Similarly, the correlation-based inferences that quantitative researchers derive from data-set observations can often be enhanced by causal-process observations.

Examples of Causal-Process Observations

Three brief, schematic illustrations of causal-process observations will help to clarify their contribution to causal inference. Because we seek to underscore the contrast with data-set observations, we present examples of studies in which both data-set and causal-process observations are employed.³⁴

The first example focuses on the use of causal-process observations to discredit the findings of a time-series cross-sectional regression analysis, based on data-set observations. In an article that became an important part of the political debate after the 2000 U.S. presidential election, John R. Lott (2000) used regression to conclude that at least 10,000 votes for Bush were lost in the Florida panhandle because the media declared Gore the winner in Florida shortly before the polls had closed in this region, which, unlike the rest of the state, is on Central Standard Time. Brady (chap. 12, this volume) employs causal-process observations, focused on the actual events of election day, to demonstrate that this inference is implausible. Brady shows that the maximum number of votes that Bush could have lost was 224, and that the actual loss was probably just a few dozen votes. Brady's causal-process observations draw on diverse sources of data to establish several pertinent facts: the number of last-minute voters, the proportion of this group of voters exposed to the media, the further proportion who would specifically have heard media predictions of the outcome, and the likely impact of this prediction on their vote. Although he could have addressed this question through a broader analysis based on data-set observations,

34. For other examples in which the contributions of these two kinds of observations are juxtaposed, see Tarrow's chapter above (especially 105–10 this volume). Of course, many case-study researchers carry out extended analyses based on causal-process observations without relying in any substantial way on data-set observations.

Brady is convinced that he got better answers using causal-process observations focused sharply on what actually happened that day in the Florida panhandle.

Another example is Susan Stokes's (2001) analysis of the dramatic economic policy shifts toward neoliberalism initiated by several Latin American presidents between 1982 and 1995. These presidents had campaigned strongly against neoliberalism. Yet, shortly after being elected, they abruptly embraced neoliberalism. Stokes's question is whether the presidents opted for neoliberalism on the basis of (a) considered views about the consequences for the economy and the functioning of the state in their countries if they failed to implement neoliberal reform, or (b) a narrower rent-seeking calculation regarding short-term economic or social payoffs from powerful market actors. Stokes systematically compares thirty-eight Latin American presidents, some of whom switched and some of whom did not. She scores them on a series of explanatory variables, as well as on the outcome variable, that is, the adoption of neoliberal policies, thus using dataset observations. This approach, employing both a probit model (93–101) and more informal comparative analysis, yields evidence favoring the first explanation, that is, that the choice was based on the conviction that neoliberalism would solve a series of fundamental national problems.

Stokes supplements this large-N analysis by examining a series of causal-process observations concerning three of the presidents, who abruptly switched from populist campaign rhetoric to neoliberal policies after winning the election. In this small-N analysis, her inferential leverage derives from the direct observation of causal links. In one of these analyses, Stokes offers an intriguing step-by-step account of how Peruvian President Fujimori decided to abandon the more populist rhetoric of his campaign and adopt a package of neoliberal reforms (2001: 69–73). Stokes shows that, just after Fujimori's electoral victory, a sequence of encounters with major international and domestic leaders exposed him to certain macroeconomic arguments, and these arguments convinced him that Peru's economy was headed for disaster if neoliberal reforms were not adopted. Causal-process observations thus provide valuable evidence for the argument that Fujimori's decision was driven by this conviction, rather than by the rent-seeking concerns identified in the rival hypothesis.

A final example of the distinctive contribution of causal-process observations comes from Nina Tannenwald's (1999) analysis of the role played by normative concerns in U.S. decisions about the use of nuclear weapons. Tannenwald hypothesizes that decisions about nuclear weapons have been guided by a "nuclear taboo," that is, a normative stigma against nuclear weapon use, which she hypothesizes to have been a powerful influence on U.S. decision making during the decades since the invention of nuclear weapons. She frames her discussion around the important competing

hypothesis that decisions about nuclear weapons were guided exclusively by considerations associated with deterrence theory.

Tannenwald uses a small-N, qualitative test based on data-set observations to evaluate the hypothesis that the nuclear taboo has had a causal impact on U.S. decision making. In comparing U.S. decisions about nuclear weapons during World War II, the Korean War, the Vietnam War, and the Gulf War, Tannenwald controls for deterrence, since none of these conflicts involved an opponent with the capacity for nuclear retaliation. Because nuclear weapons were only used during World War II, when the broad tradition of negative world public opinion about such weapons had not yet formed, Tannenwald's data-set observations are compatible with the nuclear taboo hypothesis. This comparison of four different wars thus provides some initial evidence in favor of Tannenwald's argument. However, the N is only four, so the comparison yields relatively little analytic leverage.

To gain additional leverage, Tannenwald devotes most of her analysis to the historical record, in search of evidence regarding the actual priorities of key political leaders during decisions about nuclear weapon use in each crisis. Since the nuclear taboo hypothesis implies that decision makers would be both aware of and explicitly concerned about such a taboo, causal-process observations focused on decision-making processes during each war can provide a useful test of the hypothesis. If the historical record shows that decision makers actually discussed constraining effects of a nuclear taboo, then Tannenwald has found important evidence in favor of the hypothesis.

In fact, Tannenwald finds many such statements in accounts of the relevant decision-making processes. To cite a few representative examples, when discussing the Korean War, Tannenwald presents documentary evidence that key U.S. decision makers thought the use of nuclear weapons would be a disaster in terms of world public opinion (1999: 444) and, in the words of one prominent decision maker, "offensive to all morality" (1999: 445). In parallel top-level debates on the potential use of nuclear weapons during the Vietnam War, one key meeting reached the conclusion that "use of atomic weapons is unthinkable" (1999: 454) for normative reasons.

Of course, this evidence could be accounted for in other ways than by the nuclear taboo hypothesis. For example, the statements she quotes might be strategic misrepresentations of political leaders' real agendas, or the beliefs and priorities of these leaders may in some way have been irrelevant to the decisions that they ultimately adopted. However, to the extent that researchers find alternative accounts such as strategic misrepresentation less plausible, Tannenwald's causal-process observations provide valuable support for her argument.

In discussing these three examples, we certainly do not claim to have dis-

covered a new type of evidence for use in political and social research. Such evidence is obviously familiar to scholars who use process tracing, within-case analysis, and related techniques. Our goals in this discussion are, first, to argue that these many forms of analysis employ a similar kind of evidence; and, second, to give this type of evidence, based on causal-process observations, a methodological status parallel to that of data-set observations.

Further, these three examples illustrate an important complementarity between data-set observations and causal-process observations. In all three examples, the causal-process observations focus on ideas or priorities that must be held by actors in order for the hypothesis associated with the data-set observations to be correct. They identify indispensable steps in the causal process, without which the hypothesis does not make sense.

In the following section, we explore the analytic leverage that derives from these two types of observations.

Implications of Contrasting Types of Observations

The distinction between data-set observations and causal-process observations helps to clarify several methodological issues. These include differences between qualitative and quantitative research; the implications of adding different kinds of data for the N , for degrees of freedom, and for inferential leverage; the consequences of missing data; the tools of causal inference employed in quantitative analysis; and advice about increasing the number of observations. These issues will now be explored in turn.

Qualitative versus Quantitative

Large- N quantitative researchers may routinely use large numbers of data-set observations and many fewer causal-process observations. By contrast, small- N qualitative researchers may use few data-set observations and a great many causal-process observations. These qualitative researchers use causal-process observations, as we put it above, to slowly but surely rule out alternative explanations until they come to one that stands up to scrutiny. This is a style of causal inference focused on mechanisms and processes, rather than on covariation among variables.

At the same time, we do not wish to narrowly identify the qualitative versus quantitative distinction with the causal-process versus data-set distinction. The two types of observations, used together, can provide strong inferential leverage in both traditions of research. For example, within the framework of Alexander George's "method of structured, focused comparison," which has played a central role in defining the comparative case-study tradition, researchers ask "a set of *standardized, general questions* of each

Table 9.3. Adding Different Forms of Data: Consequences for Causal Inference

<i>Adding Data</i>	<i>Consequences for Causal Inference</i>		
	<i>For the N</i>	<i>For Degrees of Freedom</i>	<i>For Inferential Leverage</i>
Adding Data-Set Observations	Increases the N	Increases degrees of freedom	Greater degrees of freedom increase leverage; yet leverage may be reduced if the addition of new observations violates measurement and causal assumptions
Adding Causal-Process Observations	Usually does not affect the N	Usually does not affect degrees of freedom ^a	New information about causal patterns may increase leverage; and if observations are drawn from original set of cases, there is less risk of violating assumptions underlying measurement and causal inference
Adding Variables	Does not affect the N	Decreases degrees of freedom	Fewer degrees of freedom reduce leverage; yet leverage is increased if key missing variables are added

^a There is no effect, unless focusing on causal-process observations leads the analyst to modify either the model being estimated, or the data set.

case" (1979a: e.g., 62), producing a uniform collection of data-set observations based on qualitative data. Conversely, as Goldthorpe and others have argued (see above), causal-process observations can make a valuable contribution to mainstream quantitative research. The label "nested inference," noted above, is intended to highlight this two-way contribution.

Adding Observations and Adding Variables: Consequences for the N, Degrees of Freedom, and Inferential Leverage

The distinctions offered above may help refine the frequently repeated advice to add observations as a means of strengthening causal inference. We would frame this topic more generically as "adding data," which can include adding data-set observations, causal-process observations, and new variables. These three alternative ways of adding data have different consequences for the N, for degrees of freedom, and for inferential leverage (table 9.3).

Consequences for the N are summarized in the left-hand column of table 9.3. The N is the number of cases, which corresponds to the number of data-set observations, that is, the number of rows in a rectangular data set. As noted, this idea applies equally to quantitative and qualitative data. The key distinction here is that increasing the number of data-set observations increases the N—whereas adding causal-process observations often does

not affect the N . Given the extensive discussion of “increasing the number of observations,” this distinction is helpful. Finally, adding variables, which may incorporate many additional pieces of data into the analysis, adds columns to the rectangular data set but does not increase the N .

The second issue concerns the consequences of adding data for the degrees of freedom (see middle column in the table). Degrees of freedom merit attention, because to the extent they are greater, the researcher has more capacity to adjudicate among rival explanations, within the framework of analyzing data-set observations.³⁵ Other things being equal, the more data-set observations (i.e., the larger the N) vis-à-vis the number of parameters to be estimated (which usually corresponds to the number of explanatory variables), the greater the degrees of freedom. Adding causal-process observations does not usually increase the N or affect the degrees of freedom.³⁶ If the researcher adds data in the sense of adding variables, this typically reduces the degrees of freedom. This is because the N remains unchanged, while the number of parameters about which inferences are to be made has increased.

Another question concerns the overall consequence for inferential leverage of adding different forms of data (right column in table 9.3). Degrees of freedom is a useful concept, but it does not capture all relevant aspects of inferential leverage. For example, it is true that adding data-set observations—that is, adding cases—can often increase inferential leverage by increasing degrees of freedom. However, a loss of inferential leverage may occur if adding cases extends the analysis to new domains where prior conceptualizations are inappropriate, measurement procedures are invalid, or causal homogeneity is lacking.

Moving down the right column in the table, we see that if the researcher makes insightful use of causal-process observations, this can increase inferential leverage. Finally, adding variables decreases the degrees of freedom and can therefore decrease inferential leverage. However, if relevant missing variables are added to the model, inferential leverage thereby increases because missing-variable bias decreases.

As an example of how adding different forms of data affects inferential

35. It is important to note that degrees of freedom, and also inferential leverage in general, are not properties of the data, but rather of the researcher’s model in relation to the data. Adding a variable to an analysis decreases the degrees of freedom if the rest of the model is not changed. Yet, it could, for example, increase the degrees of freedom if it leads to a reconceptualization of the model as a sequence of causal steps in which the number of parameters estimated is smaller at each step.

36. However, the degrees of freedom could, once again, change if these causal-process observations lead the researcher to modify the statistical model being tested.

leverage, let us consider a comparative study with an N of twenty-four, focused on explaining change in electoral systems. One hypothesis is that such change occurs when (a) public protest over political corruption increases sharply, (b) electoral reform is seen as a salient response, and (c) legislators have the constitutional authority to rapidly introduce electoral reform (Shugart, Moreno, and Fajardo 2001: 3–5, 23–34). From this starting point, the researcher might add data-set observations to the study by finding additional episodes of potential electoral change. The N and the degrees of freedom are thereby increased; other things being equal, the scholar has gained inferential leverage. However, other things are not equal if concepts and indicators do not fit the new cases, or if causal homogeneity is violated. To the extent that these problems arise, leverage for causal inference may actually be reduced.

Alternatively, the researcher might add causal-process observations to strengthen causal inferences about the original four episodes of potential electoral reform. For example, the researcher might carefully examine critical moments in the crystallization or collapse of public protest, or turning points in the electoral reform process. Nonetheless, in terms of data-set observations, the N is still twenty-four. The degrees of freedom have not changed,³⁷ yet inferential leverage may have increased.

Finally, the investigator might add data by introducing new explanatory variables—for example, the structure of the party system—as part of the uniform array of scores on the dependent and independent variables. Clearly, the N has not increased, and, with more explanatory variables, the degrees of freedom will typically be reduced. On the other hand, if the original model was underspecified and the structure of the party system is, in fact, a key missing variable, then inferential leverage is strengthened by adding this variable, which may counteract the effect of the reduced degrees of freedom.

This example illustrates how adding data to an analysis can mean three different things, and that degrees of freedom, although a valuable concept, captures only one aspect of inferential leverage. This conclusion stands out clearly in table 9.3, where for all three rows the consequences for overall inferential leverage are different, and often more ambiguous, than they are for the degrees of freedom. In order to evaluate advice to “increase the number of observations” as a means of strengthening research design, we must adopt a multifaceted view of the types of data that may be added and of their varied contribution to improving inference.

Implications for Research Design

These arguments, as summarized in table 9.3, have implications for research design. KKV repeatedly makes a case for increasing the N , but we

37. Except under the conditions specified in note 36 above.

should recognize that researchers often have good reasons for focusing on a small *N*. Therefore, advice to increase the *N* may be misplaced. For instance, the researcher may have made an enormous investment in gaining expertise on a few cases. This expertise can provide the researcher with access to a broad array of causal-process observations, which in turn can sometimes yield greater leverage for valid inference than additional cases about which the investigator knows far less. Alternatively, this scholar may have serious doubts about whether the causal patterns in these cases will be found in other cases—that is, doubts about causal homogeneity and the generality of findings. In discussions of method and theory, the problem of generalization, and specifically of overextending findings, is both an old theme (Weber 1949: 72–76; Bendix 1963; Walker and Cohen 1985) and a recently renewed concern (Elster 1999: chap. 1). Given this potential problem, along with the issues of measurement validity that can arise in moving to new contexts, the analyst might be well advised to stick to a small *N*.

By contrast, adding causal-process observations does not pose this problem of overextending the analysis, because the focus typically remains on the original cases. Such research seeks to deepen the knowledge of causal processes and mechanisms in these cases, rather than extend the study to additional cases. The challenge a researcher faces when adding causal-process observations is to know which details to collect, when enough details have been collected to make an inference, and how to increase the likelihood that this inference is valid. The literature on case studies and within-case analysis would do well to address these issues in greater depth.

To conclude, although the advice to increase the number of data-set observations is sometimes valuable, it may simply be distracting for researchers who have deliberately focused on explaining a small number of important outcomes. These researchers may find that collecting relevant causal-process observations is more helpful. Further, for quantitative researchers, causal-process observations can be a valuable supplement to large-scale data sets.

Missing Data

A distinction should also be made about the implications of missing data for these two types of observations. With data-set observations, missing data can be a serious issue. Indeed, the idea that data-set observations involve a uniform array should be understood as encompassing the norm that the data set should preferably be complete, and that a problem of missing data requires close attention (Griliches 1986; Greene 2000: 259–63).

Almost by definition, the issue of missing data does not arise in the same way for causal-process observations. The inferential leverage derived from causal-process observations does not depend on having complete data

across a given range of cases and variables. Thus, one or a few causal-process observations may provide great leverage in making inferences. For example, Stokes's analysis of presidential policy switches, discussed above, derives analytic leverage from observations of the decision-making processes involved in only three of the thirty-eight cases that she considers. Her close analysis of these three cases obviously does not "prove" her hypothesis for all thirty-eight episodes, but it does increase the plausibility of her overall conclusions by offering telling evidence about three episodes. Likewise, data-set observations can potentially compensate for gaps or inadequacies in causal-process observations.

Standard Quantitative Tools versus Careful Analysis of Causal-Process Observations

The distinction between data-set observations and causal-process observations offers a new basis for thinking about the application of standard quantitative tools to different kinds of research. We have elaborate quantitative procedures for evaluating inferences made with data-set observations. By contrast, causal-process observations force us to make complex judgments about inference and probability without explicit guidance from quantitative tools. It is precisely the emphasis on standard quantitative tests that leads KKV to make what we view as a major mistake: subordinating causal-process observations to a conventional quantitative framework (see again 85–87, 226–28).

A small number of causal-process observations, that seek to uncover critical turning points or moments of decision making, can play a valuable role in causal inference. Making an inference from a smoking gun does not require a large N in any traditional sense. However, it does require careful thinking about the logic of inference and a rich knowledge of context, which may in turn depend on many additional causal-process observations. The several chapters in the present volume that discuss tools for qualitative analysis have suggested points of departure for reasoning about how these inferences take place.

CONCLUSION: DRAWING TOGETHER THE ARGUMENT

In chapters 8 and 9, we have expressed reservations about KKV's positions on causal inference, descriptive inference, and related methodological questions. KKV in effect treats causal inference as fairly straightforward, provided the researcher follows the quantitative template.³⁸ We would

38. See again the cautionary observation in chapter 1 above (21 n. 3 this volume).

instead argue that adequate causal inference is difficult. To the extent that KKV addresses challenges to causal inference, it treats these issues as in effect depending on the power of quantitative tests. Thus, the book focuses on increasing the number of observations, estimating uncertainty, and the closely related and misleading idea that—as KKV puts it—determinate research designs (i.e., designs with a sufficiently large N and a lack of perfect multicollinearity) are the “sine qua non” of causal inference.

This emphasis on determinate research designs obscures basic challenges in making what we prefer to call “interpretable” causal inferences: the challenges of ruling out an unknown number of alternative explanations and dealing with hard-to-test assumptions. Effective causal inference requires bringing to bear as many different kinds of evidence as possible, including evidence from qualitative research. Yet in KKV’s approach, the contribution of qualitative evidence is undervalued because it is inappropriately assessed in terms of the size of the N and quantitative tests, which misrepresents its distinctive contributions.

With regard to descriptive inference, KKV devotes a chapter to this topic. However, the book’s discussion focuses primarily on relatively straightforward questions, such as how to generalize from a sample to a population and how to productively organize and summarize descriptive detail. Yet descriptive inference raises broader, more complex issues that require far more attention. Causal inferences are only reasonable if measurement is valid. Measurement validity, in turn, depends on careful attention to conceptualization—a topic for which KKV’s advice points in the wrong direction—and on the plausibility of each decision taken in the measurement process. Issues of conceptualization and measurement are more fundamental than the conventional problem of generalizing from a sample to a population; indeed, such issues must be addressed even if researchers make no attempt to generalize their claims.

For many other methodological questions, we are again convinced that KKV adopts positions that are somewhat simplistic: for example, the book’s arguments about appropriate techniques for case selection and against testing deterministic causal models, along with the failure to recognize that techniques of within-case analysis yield a different kind of evidence than do conventional quantitative data. These are complex issues and must be addressed within a methodological framework that extends well beyond that of KKV and of mainstream quantitative methods.

In the present volume, we have sought to develop this broader framework and have argued that it yields a more positive perspective on qualitative tools for descriptive and causal inference. Part of this argument derives from what we have called the statistical rationale for qualitative research. Specifically, we have invoked the statistical idea that important gaps in causal inference based on the quantitative analysis of data-set observations

can be filled by evidence derived from qualitative, causal-process observations. Inference based on qualitative data routinely employs different assumptions than quantitative inference, and correspondingly it provides an alternative source of analytic leverage. Such leverage can serve to improve not only qualitative research, but also quantitative research.

Similarly, with regard to descriptive inference, we have argued that reasoning about measurement found in psychometrics and mathematical measurement theory points to concerns to which qualitative researchers are routinely more attentive—such as the foundational role of paired comparisons in the logic of measurement, as well as concern with issues of domain and context. The present volume has sought to show how these qualitative and statistical traditions can help lay a stronger methodological foundation for progress in the social sciences.

Running through this discussion have been the themes of diverse tools and shared standards. From one perspective, these ideas might seem contradictory: a strong set of shared standards might rule out all but a single, best package of tools. We are convinced that this contradiction does not arise in the social sciences for a simple reason. In light of the current state of methodological knowledge, scholars face many trade-offs in pursuing good descriptive and causal inference. Given these trade-offs, there is no such thing as a universally best set of tools. Rather, the existence of trade-offs requires a sustained recognition that diverse analytic tools are needed in social research.

BALANCING METHODOLOGICAL PRIORITIES: TECHNIFICATION AND THE QUEST FOR SHARED STANDARDS

In concluding this volume, we would like to reflect on the overall mix of concerns and priorities that are most productive in advancing both methodology and substantive research. We find ourselves in a period when increasingly technical approaches to methodology and theory have growing influence in the social sciences. Whether they involve new procedures for statistical estimation or new tools for deductive inference, these innovations unquestionably help us to understand political and social reality.

Yet this trend toward technification can impose substantial costs. It can lead to replacing a simple and appropriate tool with an unnecessarily complex one. It can sometimes distance analysts from the detailed knowledge of cases and contexts that is an invaluable underpinning for any inference, whether derived through complex research procedures or simpler tools. Technification can also devolve into a form of intellectual obscurantism in

which research ceases to be driven by important substantive questions and interesting intellectual agendas.

In some circumstances a sophisticated, technical solution is indeed more powerful. However, at other times it is better to adopt an alternative solution based on simpler tools. As qualitative methodologists routinely emphasize, these simpler tools can place scholars in closer contact with the cases being studied, sometimes enabling analysts to discover unanticipated causal patterns. Further, when highly technical tools are employed, they cannot be a substitute either for careful thinking about the process that produced the data, or for crafting good—and often elegantly simple—research designs that allow one to rule out alternative explanations. This careful thinking often relies on simple forms of data analysis—employing, perhaps, a scatterplot, or a two-by-two table—and on crafting a parsimonious model that undergirds the research design.³⁹

Scholars should recognize that simpler analytic tools can sometimes contribute more to achieving the shared standards of valid descriptive and causal inference and refining theory. We believe that the greatest promise for progress in social science lies in an eclectic view of methodology that recognizes the potential contributions of diverse tools to meeting these shared standards.

39. See Achen (2000, 2002) and also Diaconis (1998). For a broader statement on these tensions in the discipline of political science, see Keohane 2003.

II

CAUSAL INFERENCE: OLD DILEMMAS, NEW TOOLS

David Collier, Henry E. Brady, and Jason Seawright

The quest for better tools of causal inference¹ is an abiding concern of methodology. Looking beyond the debate with KKV, the preceding chapters have mapped out alternative paths and pitfalls in this quest. We have offered new warnings about the limitations of regression-based inference. With regard to qualitative tools, above all in chapter 9 we have sought to systematize their contribution, increase their coherence and transparency, and make them more amenable to being communicated among scholars and taught in the classroom.

QUALITATIVE TOOLS: PROCESS TRACING AND CAUSAL PROCESS OBSERVATIONS

Our treatment of causal inference in qualitative research joins two ideas: process tracing and causal process observations. Following Bennett (chap. 10, this volume), we understand process tracing as the examination of diagnostic pieces of evidence, commonly evaluated in a specific temporal sequence, with the goal of supporting or overturning alternative explanatory hypotheses. In the framework of process tracing, these diagnostic pieces of evidence are what

1. To reiterate a point made earlier, our own work has extended well beyond this focus to encompass extensive research on concept-formation and measurement. Our concern here, however, is with causal inference.

we have called *causal process observations* (CPOs).² Process tracing consists of procedures for singling out specific CPOs and evaluating their contribution to causal inference in a given analytic setting.³

In the first chapter of Part II, Andrew Bennett formulates a new typology of process tracing that places on two dimensions the alternative tests employed. The tests are distinguished according to whether passing a particular test is *necessary* for inferring causation, and whether it is *sufficient*. This provides a new framework for thinking about the tests originally proposed by Van Evera (1997: 31–32): the straw-in-the wind, hoop, smoking-gun, and doubly-decisive tests. Bennett illustrates his framework by applying the typology at the level of macro-politics, focusing on three well-known historical episodes in international relations.

Next, the statistician David A. Freedman (chap. 11) examines the role of process tracing at the level of micro-analysis, focusing on six major studies from the history of epidemiology—including John Snow’s classic work on cholera. According to Freedman’s view, in both epidemiology and the social sciences the analysis of qualitative evidence—specifically CPOs—is a basic type of scientific inquiry. Such evidence can play a valuable role in “refuting conventional ideas if they are wrong, developing new ideas that are better, and testing the new ideas as well as the old ones” (chap 11).

Freedman’s position thus challenges that of Piore (2006: 17), who asserts that information from case studies “cannot be treated directly as empirical evidence. . . .” It also quite different from the view of Fearon and Laitin (2008: 756), who maintain that it is quantitative rather than qualitative analysis that serves to establish a “relationship among variables.” More than a few quantitative researchers need to examine Freedman’s views and reconsider their skepticism about inference in qualitative research.

Freedman is strongly committed to the careful juxtaposition of CPOs and DSOs. He is skeptical about much quantitative research and believes that

2. We define causal-process observations (CPOs) as pieces of data that provide information about context, process, or mechanism and contribute distinctive leverage in causal inference. They are contrasted with data-set observations (DSOs), which correspond to the familiar rectangular data set of quantitative researchers. In quantitative research, the idea of an “observation” (as in DSO) has special status as a foundation for causal inference, and we deliberately incorporated this label in the idea of CPOs to underscore their relationship to causal inference. Obviously, we do not thereby mean that one directly *observes* causation. Rather, this involves *inference*, not direct *observation*. We are pleased that the idea of CPOs appears to have been applied successfully in other contexts. For example, Mahoney (2010: 123–31) uses it as one of the organizing principles in his essay of trends in methodology. See also Freedman (chap. 11, this volume).

3. Mahoney (2010: 124) offers this same account of the relationship between process tracing and CPOs.

optimally, qualitative and quantitative analysis should be juxtaposed and should work together. The examples from his chapter show again and again how this juxtaposition can occur.

Henry E. Brady in chapter 12, focusing on electoral behavior, shows how a sequence of CPOs can yield causal inference.⁴ This procedure gives crucial leverage in disputing the claim, advanced in a quantitative study by John Lott, that George Bush lost thousands of votes in the 2000 presidential election in Florida due to the media's early (and incorrect) call of the election outcome. Brady's working hypothesis is that the early call had little impact on the actual vote. He goes through a series of process-tracing steps to support his hypothesis, employing a sequence of vote counts and assumptions about voting behavior to identify the necessary conditions for Lott's hypothesis to be plausible. The necessary conditions are not met, and Brady concludes that Lott's hypothesis is not plausible.

Brady's chapter reminds us of the obvious but crucial point that process tracing and CPOs can involve numerical data. Thus, numerical data do not necessarily constitute DSOs—they do not necessarily take the form of a rectangular data set. Rather, the isolated (though carefully framed) pieces of quantitative information employed by Brady are CPOs.

As noted in the preface, the above readings on process tracing are supplemented by a set of exercises that are posted online.

QUANTITATIVE TOOLS

The other two chapters in Part II evaluate a spectrum of older and newer methods of quantitative investigation, with applications in both macro and micro studies. Jason Seawright (chap. 13) presents a "case study of failed causal inference," evaluating unsuccessful attempts to employ quantitative, cross-national regression analysis to address a classic topic in comparative social science: the impact of political regime type on economic growth.

Seawright's specific substantive focus is of great importance, and his analysis has broad implications for the vast literature that uses the quantitative, cross-national method to study dozens of major topics in political and social science. Indeed, in conjunction with his analysis of these pitfalls, he cites authors who call for an "obituary" for a significant part of the quantitative cross-national literature.

His arguments are also relevant for more general discussions of regression analysis. For example, many scholars believe they can always add "just one more control variable" to improve regression-based inference. Yet adding controls can potentially make inference worse, rather than better.

4. Brady's analysis was an appendix in the first edition.

Seawright goes on to show that refinements in regression—such as designs employing matching, regression discontinuity, and instrumental variables—are at best problematic in addressing the problems of regression-based inference that arise in research on his substantive topic. He concludes that scaling down to a finer-grained focus, which may include substantial use of qualitative analysis, is a sensible next step in response to these failures.

The final chapter by Thad Dunning (chap. 14) likewise formulates a novel typology, in this case for evaluating the family of techniques known as natural experiments—including regression discontinuity and instrumental variables designs. These techniques seek to overcome the failures of regression explored by Seawright and discussed above in the Introduction to the Second Edition. Dunning applies his typology to assess these designs on three dimensions: (1) plausibility of the “*as-if* random assignment” assumption central to the idea of natural experiments; (2) credibility of the statistical model; and (3) substantive relevance of the key explanatory variable. This third point is especially important because in some of these studies, the search for situations of *as-if* random assignment drastically limits substantive relevance. Indeed, this shortcoming is a serious weakness in these different forms of natural experiments.

By juxtaposing the three dimensions in his typology, Dunning provides an overall evaluation of whether particular natural experiments employ strong or weak research designs. He underscores the critical contribution of qualitative evidence, and while he views natural experiments as a promising tool, he argues that they should definitely not eclipse other methodologies.

To make their chapters accessible to a broad range of readers, Seawright and Dunning have presented technical issues in footnotes. For example, the Neyman-Rubin-Holland model of causal inference is crucial to Dunning’s analysis, and is entirely presented in a series of notes.

To summarize, we seek in these new chapters to advance the understanding of both qualitative and quantitative methods, showing—among other things—how the two can work together to yield stronger inference.

E. QUALITATIVE TOOLS FOR CAUSAL INFERENCE

10

Process Tracing and Causal Inference

Andrew Bennett

How should we judge competing explanatory claims in social science research? How can we make inferences about which alternative explanations are more convincing, in what ways, and to what degree? Case study methods—especially methods of within-case analysis such as process tracing—are an indispensable part of the answer to these questions (George and Bennett 2005: chap. 10). This chapter offers an overview of process tracing as a tool for causal inference, focusing on the study of international relations, an area rich with examples of this approach.¹ In contrast to the subsequent two chapters in this volume (chaps. 11 and 12), where Freedman and Brady analyze micro-level examples, the present chapter explores process tracing in macro studies.

This chapter uses three explanatory puzzles, about which scholars have advanced contending hypotheses, to illustrate how process tracing helps adjudicate among alternative explanations: (1) why and how the United

Maria Gould, Jody La Porte, and Miranda Yaver provided valuable comments on an earlier draft of this chapter.

1. Good examples include Drezner (1999), Eden (2004), George and Smoke (1974), Homer-Dixon (1999), Khong (1992), Knopf (1998), Larson (1997), Moravcsik (1998), Owen (1997), Rock (1989, 2000), Sagan (1993), Shafer (1988), Snyder (1984, 1991), Walt (1996), and Weber (1991). Brief descriptions of the research designs employed by Drezner, George and Smoke, Homer-Dixon, Khong, Knopf, Larson, Owen, Sagan, Shafer, Snyder, and Weber are provided by George and Bennett (2005: 118–119, 194–97, 302–325).

Kingdom and France resolved their competing imperial claims to the Upper Nile Valley without resorting to the use of force in the Fashoda crisis of 1898, an outcome that has been the subject of considerable research given its relevance to the inter-democratic peace hypothesis; (2) why in the middle of World War I, despite strong evidence that it was likely to be defeated, Germany expanded its war goals—for example, shifting to unrestricted submarine warfare—even though this risked (and in fact, resulted in) American entry into the conflict; and (3) why the Soviet Union did not intervene militarily in the Central European revolutions of 1989, in contrast to its military interventions in Hungary in 1956 and Czechoslovakia in 1968.

OVERVIEW OF PROCESS TRACING

Process tracing involves the examination of “diagnostic” pieces of evidence within a case² that contribute to supporting or overturning alternative explanatory hypotheses. A central concern is with sequences and mechanisms in the unfolding of hypothesized causal processes.³ The researcher looks for the observable implications of hypothesized explanations, often examining evidence at a finer level of detail or a lower level of analysis than that initially posited in the relevant theory. The goal is to establish whether the events or processes within the case fit those predicted by alternative explanations.

This mode of analysis is closely analogous to a detective attempting to solve a crime by looking at clues and suspects and piecing together a convincing explanation, based on fine-grained evidence that bears on potential suspects’ means, motives, and opportunity to have committed the crime in question. It is also analogous to a doctor trying to diagnose an illness by taking in the details of a patient’s case history and symptoms and applying diagnostic tests that can, for example, distinguish between a viral and a bacterial infection (Gill, Sabin, and Schmid 2005).

Process tracing, which focuses on the diagnostic intervening steps in a hypothesized causal process, can provide inferential leverage on two problems that are difficult to address through statistical analysis alone. The first is the challenge of establishing causal direction: if X and Y are correlated,

2. A case may be understood as a temporally and spatially bounded instance of a specified phenomenon. Although process tracing focuses on events within a case, it can play a role in comparisons of cases. An analyst can use process tracing, for example, to assess whether a variable whose value differs in two most similar cases is related to the difference in their outcomes.

3. Process tracing is also used as a method of discovering hypotheses, a contribution illustrated below in Freedman’s contribution (chap. 11). However, that facet is not addressed in the present chapter.

did X cause Y, or did Y cause X? Careful process tracing focused on the sequencing of who knew what, when, and what they did in response, can help address this question. It might, for example, establish whether an arms race caused a war, or whether the anticipation of war caused an arms race.

A second challenge is that of potential spuriousness: If X and Y are correlated, is this because X caused Y, or is it because some third variable caused both X and Y? Here, process tracing can help establish whether there is a causal chain of steps connecting X to Y, and whether there is such evidence for other variables that may have caused both X and Y. There is no guarantee that researchers will include in their analyses the variable(s) that actually caused Y, but process tracing backward from observed outcomes to potential causes—as well as forward from hypothesized causes to subsequent outcomes—allows researchers to uncover variables they have not previously considered. This is similar to how a detective can work forward from suspects and backwards from clues about a crime. It is likewise consistent with David Freedman's argument (chap. 11, this volume) that case expertise and substantive knowledge can play a key role in sorting out explanations—a claim that may for some readers appear counter-intuitive in light of Freedman's disciplinary background as a mathematical statistician.

Critics have raised two critiques of process tracing: the "infinite regress" problem and the "degrees of freedom" problem. On the former, King, Keohane, and Verba suggest that the exceedingly fine-grained level of detail involved in process tracing can potentially lead to an infinite regress of studying "causal steps between any two links in the chain of causal mechanisms" (1994: 86). Others have worried that qualitative research on a small number of cases with a large number of variables suffers from a degrees of freedom problem. This form of indeterminacy afflicts statistical studies, given that the number of cases in a data set must be far greater than the number of variables in a model to test that model through frequentist statistics.

The answer to both critiques is that not all data are created equal. With process tracing, not all information is of equal probative value in discriminating between alternative explanations, and a researcher does not need to examine every line of evidence in equal detail. It is possible for one piece of evidence to strongly affirm one explanation and/or disconfirm others, while at the same time numerous other pieces of evidence might not discriminate among explanations at all. What matters is not the amount of evidence, but its contribution to adjudicating among alternative hypotheses. Further, even a single case may include many salient pieces of evidence. The noted methodologist Donald Campbell recognized the value of process-focused tools of inference when he abandoned his earlier criticism of

case studies as lacking degrees of freedom, and argued in favor of a method similar to the process tracing under discussion here (Campbell 1975).

More concretely, process tracing involves several different kinds of empirical tests, focusing on evidence with different kinds of probative value. Van Evera (1997: 31–32) has distinguished four such tests that contribute in distinct ways to confirming and eliminating potential explanations. They are summarized briefly here, and will then be applied and illustrated throughout this chapter.

Hoop tests, which are central to the discussion below, can eliminate alternative hypotheses, but they do not provide direct supportive evidence for a hypothesis that is not eliminated. They provide a *necessary but not sufficient* criterion for accepting the explanation. The hypothesis must “jump through the hoop” just to remain under consideration, but success in passing a hoop test does not strongly affirm a hypothesis. Van Evera’s apt example of a hoop test is, “Was the accused in the state on the day of the murder?”

Smoking gun tests strongly support a given hypothesis, but failure to pass such a test does not eliminate the explanation. They provide a *sufficient but not necessary* criterion for confirmation. As van Evera notes, a smoking gun

Table 10.1. Process Tracing: Four Tests for Causation^a

		Sufficient To Establish Causation ^b	
		No	Yes
Necessary To Establish Causation	No	Straw in the Wind	Smoking Gun
		<i>Passing</i> affirms relevance of hypothesis but does not confirm it.	<i>Passing</i> confirms hypothesis.
		<i>Failing</i> suggests hypothesis may not be relevant, but does not eliminate it.	<i>Failing</i> does not eliminate it.
	Yes	Hoop	Doubly Decisive
	<i>Passing</i> affirms relevance of hypothesis but does not confirm it.	<i>Passing</i> confirms hypothesis and eliminates others.	
	<i>Failing</i> eliminates it.	<i>Failing</i> eliminates it.	

^a The typology creates a new, two-dimensional framing of the alternative tests originally formulated by Van Evera (1997: 31–32).

^b In this figure, “establishing causation,” as well as “confirming” or “eliminating” an hypothesis, obviously does not involve a *definitive* test. Rather, as with any causal inference, qualitative or quantitative, it is a *plausible* test in the framework of (a) this particular method of inference and (b) a specific data set.

in the suspect's hands right after a murder strongly implicates the suspect, but the absence of such a gun does not exonerate a suspect.

Straw in the wind tests provide useful information that may favor or call into question a given hypothesis, but such tests are not decisive by themselves. They provide *neither a necessary nor a sufficient* criterion for establishing a hypothesis or, correspondingly, for rejecting it.

Finally, *doubly decisive tests* confirm one hypothesis and eliminate others. They provide a *necessary and sufficient* criterion for accepting a hypothesis. Just one *doubly decisive* piece of evidence may suffice, whereas many *straw in the wind tests* may still be indeterminate vis-à-vis alternative explanations. Van Evera's example is a bank camera that catches the faces of robbers, thereby implicating those photographed and exonerating all others. He emphasizes that in the social sciences such tests are rare, yet a *hoop test* and a *smoking gun test* together accomplish the same analytic goal (1997: 32), a combination that is illustrated in the examples below.

In process tracing and in applying these tests, it is essential to cast the net widely in considering alternative explanations. Other standard injunctions advocate gathering diverse forms of data, being meticulous and even-handed in collecting and evaluating data, and anticipating and accounting for potential biases in the evidence (George and Bennett 2005, Bennett and Elman 2006). Further, as with all forms of causal inference, specific process tracing tests must be evaluated in relation to a wider body of evidence. These desiderata are especially important in process tracing on social and political phenomena for which participating actors have strong instrumental or ideational reasons for hiding or misrepresenting information about their behavior or motives.

Example: Why the Fashoda Crisis Did Not Result in War

Schultz provides excellent examples of the *hoop test* and *smoking gun test* in his analysis of the 1898 Fashoda crisis between Britain and France. This crisis arose over the confrontation between the two countries' expeditionary forces as they raced to lay claim to the Upper Nile Valley. War was averted when France backed down. With the emergence of the inter-democratic peace research program in the last several decades, this episode has assumed special interest as a near war between two democracies, leading scholars to closely scrutinize explanations of its non-occurrence.

Schultz lays out three alternative explanations that scholars have offered for why the crisis was resolved without a war. Neorealists argue that France backed down simply because Britain's military forces were far stronger, both in the region and globally (Layne 1994). Schultz rejects this explanation because it fails to survive a *hoop test*: it cannot explain why the crisis happened in the first place, why it lasted two months, and why it escalated

almost to the point of war, as it should have been obvious to France from the outset that Britain had military superiority (Schultz 2001: 177). A second argument, that democratic norms and institutions led to mutual restraint, also fails a *hoop test* in Schultz's view. Whereas traditional democratic peace theorists emphasize the restraining power of democratic norms and institutions, the British public and British leaders were belligerent throughout the crisis in their rhetoric and actions toward France (Schultz 2001: 180–183).

Schultz then turns to his own explanation: democratic institutions force democratic leaders to reveal private information about their intentions, making it difficult for them to bluff in some circumstances but also making threats to use force more credible in others. In this view, democratic institutions reinforce the credibility of coercive threats when domestic opposition parties and publics support these threats, but they undermine the credibility of threats when domestic groups publicly oppose the use of force.

Schultz supports this explanation with *smoking gun* evidence. The credibility of Britain's public commitment to take control of the region was resoundingly affirmed by the opposition Liberal Party leader Lord Rosebery (Schultz 2001: 188). Meanwhile, France's Foreign Minister, Theophile Delcasse, initially voiced an intransigent position, but his credibility was quickly undermined by public evidence that other key French political actors were apathetic toward, or even opposed to, a war over Fashoda (Schultz 2001: 193). Within a matter of days after such costly signaling by both sides revealed Britain's greater willingness and capability to fight for the Upper Nile, France began to back down, leading to a resolution of the crisis in Britain's favor. In sum, the close timing of these events, following in the sequence predicted by Schultz's theory, provides *smoking gun* evidence for his explanation; this, combined with the alternative explanations' failures in *hoop tests*, makes Schultz's explanation of the Fashoda case convincing.

Example: Expanding the Ends and Means of German Strategy in World War I

A second example shows how *hoop tests* and a *smoking gun test* help adjudicate among rival explanations for why Germany expanded both the ends and means of its wartime strategy in 1916–1917 even as it was becoming obvious that Germany was losing World War I. Goemans convincingly argues that four developments in 1916 made it increasingly evident to German leaders that they were unlikely to win the war: the German offensive at Verdun failed; Britain demonstrated its resolve—including its tolerance for casualties—in the battle of the Somme; Russia's Brusilov offensive showed it could still fight; and Romania entered the war against Germany

(Goemans 2000: 89–93). Meanwhile, President Wilson’s diplomatic note to Germany in April 1916 after the sinking of the unarmed SS *Sussex* made it clear that the United States was almost certain to enter the war against Germany if German U-Boats sank any more merchant ships, which inhibited Germany from attacking merchantmen for the rest of the year.

Despite these developments, in late 1916 Germany escalated its terms for concluding the war, expanding its claims on Polish territory and increasing the territorial or diplomatic concessions it demanded from France, Belgium, and Russia (Goemans 2000: 98–106). Moreover, Germany returned to unrestricted submarine warfare in early 1917, even though the predictable consequence was that the United States, in quick response, entered the war.

Why did Germany expand the ends and means of its war strategy even as its probability of victory declined? Goemans evaluates five rival explanations. A first alternative—that Germany should have behaved as a unitary actor and responded only to international considerations—fails a *hoop test*, based on thorough evidence that Germany’s goals in the war expanded even though German leaders themselves understood that their prospects for victory had diminished. A second argument, that Germany was irrevocably committed to hegemony throughout the war, is also undercut by evidence that German war aims increased over time. Goemans rejects a third argument—Germany’s authoritarian government made it a “bad learner” impervious to evidence that it was losing the war—with ample indications that German leaders understood very well by late 1916 that their chances for victory were poor. A fourth explanation, that the change in Germany’s military leadership led to expanded military goals, begs the question of why Germany replaced its military leaders in the midst of the war (Goemans 2000: 74–75, 93–105).

Goemans then evaluates his own hypothesis: when semi-authoritarian governments, like that of Germany during World War I, believe they are losing a war, they are likely to respond with war strategies that preserve at least a small probability of resounding victory, even if such strategies have a high likelihood of abject defeat. Goemans argues that for leaders in such governments, the consequences of negotiating an end to a war on modestly concessionary terms are little different from those of losing the war outright. In either case, semi-authoritarian leaders are likely to lose their power and property (and perhaps even their lives) to domestic opponents who blame them for having demanded immense sacrifices from their societies in a losing cause. Thus, when evidence mounts that a semi-authoritarian state is losing in a war, its leaders have an incentive to gamble for resurrection and adopt riskier strategies that offer at least some slim hope of victory, even though they also increase the odds of utter defeat.

Goemans provides a *smoking gun test* for this argument in the case of Ger-

many's escalating war aims. Among many other pieces of evidence, he quotes the German military leader Erich Ludendorff as arguing in a private letter that radical and unacceptable domestic political reforms would be required to stave off unrest if Germany were to negotiate a concessionary peace. Specifically, Ludendorff argued that the extension of equal voting rights in Prussia "would be worse than a lost war" (Goemans 2000: 114). This letter provides direct evidence of the German leadership's desperation to avoid losing the war because of the political consequences for German leaders should they be blamed for having lost the war, and it thereby constitutes a *smoking gun test* that substantially validates Goemans's main argument.

Example: The Peaceful End of the Cold War

The final example concerns use of the *hoop*, *smoking gun*, and *straw in the wind tests* to adjudicate among hypotheses about why the Soviet Union did not intervene militarily in the Eastern European revolutions of 1989.⁴ Three prominent accounts for the non-use of force, involving standard alternative explanatory perspectives in the international relations field, are: (1) a realist hypothesis, which emphasizes the changing material balance of power; (2) a domestic politics hypothesis, which focuses on the changing nature of the Soviet Union's ruling coalition; and (3) an ideational hypothesis centered on Soviet leaders' lessons from their recent experiences.

First, the most comprehensive realist/balance of power analysis of Soviet restraint in 1989 is offered by Brooks and Wohlforth (2000/2001; see also Wohlforth 1994/1995, Oye 1996). They argue that the decline in Soviet economic growth rates in the 1980s, combined with the Soviet Union's high defense spending and its "imperial overstretch" in Afghanistan, led to Soviet foreign policy retrenchment in the late 1980s. Soviet leaders were constrained from using force in 1989 because this would have imposed large direct economic and military costs, risked economic sanctions from the West, and forced the Soviet Union to assume the economic burden of the large debts that Eastern European regimes had incurred to the West. In this view, changes in Soviet leaders' ideas about foreign policy were largely determined by changes in their material capabilities.

Second, a domestic politics account has been well formulated by Snyder (1987/88). He argues that the long-term change in the Soviet economy from extensive development (focused on basic industrial goods) to intensive development (involving more sophisticated and information-intensive

4. I use this example in part because it involves my own research, making it easier to reconstruct the steps involved in the process tracing. See Bennett (1999, 2003, 2005).

goods and services) shifted the ruling Soviet coalition from a military/heavy-industry/party complex to a power bloc centered in light industry and the intelligentsia. This led the Soviet Union to favor improved ties to the West to gain access to technology and trade, and any Soviet use of force in Eastern Europe in 1989 would have damaged Soviet economic relations with the West.

The third line of argument maintains that Soviet leaders learned lessons from their unsuccessful military interventions in Afghanistan and elsewhere that led them to doubt the efficacy of using force to try to resolve political problems like the Eastern Europeans' demands for independence from the Soviet Union in 1989.⁵ The Soviet Union invaded Afghanistan in December 1979 and kept between 80,000 and 100,000 troops there for a decade, with over 14,000 Soviet soldiers killed and 53,000 injured. When even this effort and substantial economic aid failed to make the communist party of Afghanistan capable of defending itself, Soviet leaders withdrew their military forces in February 1989. The learning explanation argues that this experience made Soviet leaders unwilling to use force nine months later to keep in power Eastern European leaders who by that time faced strong public opposition.

While scholars agree that the variables highlighted by all of these hypotheses contributed to the non-use of force in 1989, there remains considerable disagreement on how these variables interacted and their relative causal weight. Brooks and Wohlforth, for example, disagree with the "standard view" that "even though decline did prompt change in Soviet foreign policy, the resulting shift could just as easily have been toward aggression or a new version of muddling through . . . and that other factors played a key role in resolving this uncertainty" (2002: 94). In contrast, I assert that this standard interpretation is persuasive and maintain that were it not for other factors, the economic decline of the Soviet Union relative to the West could indeed have led to renewed Soviet aggression or to more years of muddling through. Specifically, I argue that although changes in the material balance of power made Soviet leaders more open to new ideas, the particular lessons Soviet leaders drew from their uses of force in the 1970s and 1980s greatly influenced the timing and direction of changes in Soviet foreign policy.

What kinds of evidence can adjudicate among these hypotheses? In introducing a symposium on competing views on these hypotheses, Tannenwald (2005) poses three questions for judging them: (1) Did ideas correlate with the needs of the Soviet State, actors' personal material interests, or actors' personal experiences and the information to which they were

5. Bennett (1999, 2003, 2005). See also English (2000, 2002); Checkel (1997); Stein (1994).

exposed? (2) Did material change precede or follow ideational change? (3) Do material or ideational factors better explain which ideas won out? Each of these questions creates opportunities for process tracing tests.

Focusing on the first question, about the correlation of policy positions with material versus ideational variables, we find some evidence in favor of each explanation. Citing Soviet Defense Minister Yazov and others, Brooks and Wohlforth argue that Soviet conservatives and military leaders did not question Gorbachev's concessionary foreign policies because they understood that the Soviet Union was in dire economic straits and needed to reach out to the West. They also point to ample evidence that Gorbachev argued that Soviet economic decline created a need for better relations with the West (Brooks and Wohlforth 2000/2001). Their explanation thus satisfies a *hoop test*: given the salience of both economic issues and relations with the West, Brooks and Wohlforth's argument would be unsustainable without considerable evidence that Soviet leaders linked the two in their public and private statements.

However, Robert English suggests that the evidence we have employed in this *hoop test* is not definitive, and he points to other statements by Soviet conservatives indicating opposition to Gorbachev's foreign policies. He concludes that "whatever one believes about the old thinkers' acquiescence in Gorbachev's initiatives, it remains inconceivable that they would have launched similar initiatives without him" (English 2002: 78). In this view, much of the evidence linking material decline to Soviet retrenchment depends on Gorbachev's individual views and the political institutions that gave him power, rather than any direct and determinative tie between material decline and specific foreign policies.

Two other *hoop tests* yield more definitive evidence against Snyder's sectoral interest group hypothesis and in favor of the learning hypothesis. Consistent with Snyder's argument, Soviet military leaders at times argued against defense spending cuts, and the conservatives who attempted a coup against Gorbachev in 1990 represented the Stalinist coalition of the military and heavy industry. Soviet Conservatives, however, did *not* argue that force should have been used to prevent the dissolution of the Warsaw Pact in 1989, even after they had fallen from power in 1990 and had little to lose (Bennett 2005: 104). Indeed, military leaders were among the early skeptics regarding the use of force in Afghanistan, and many prominent officers with personal experience in Afghanistan resigned their commissions rather than participating in the 1994–1997 Russian intervention in Chechnya (Bennett 1999: 339–340). This suggests that the learning explanation has survived a difficult *hoop test* by correctly anticipating that those military officers who personally experienced failure in Afghanistan would be among the opponents rather than the supporters of using force in later circumstances.

Concerning Tannenwald's second question, about the timing of material and ideational change, Brooks and Wohlforth have not indicated precisely the time frame within which material decline would have allowed or compelled Soviet foreign policy change, stating only that material incentives shape actions over the "longer run" (2002: 97). This suggests that the timing of changes in Soviet policy in relation to that of changes in the material balance of power is at best a *straw in the wind test*. Brooks and Wohlforth's logic allows for the possibility that the Soviet Union could profitably have let go of its Eastern European empire in 1973. By that time, nuclear parity guaranteed the Soviet Union's security from external attack, and high energy prices meant that the Soviet Union could have earned more for its oil and natural gas from world markets than from Eastern Europe. Moreover, the sharpest decline in the Soviet economy came after 1987, by which time Gorbachev had already begun to signal to governments in Eastern Europe that he would not use force to rescue them from popular opposition (Brown 1996: 249). The timing of changes in Soviet policy therefore does not lend strong support for the "material decline" hypothesis.

The timing suggested by the ideational explanation coincides much more closely with actual changes in Soviet foreign policy. Despite slow Soviet economic growth, Soviet leaders were optimistic about the use of force in the developing world in the late 1970s due to the ease with which they inflicted a costly defeat on the United States in Vietnam, but they became far more pessimistic regarding the efficacy of force as their failure in Afghanistan deepened through the 1980s (Bennett 1999). Furthermore, changes in Soviet leaders' public statements generally preceded changes in Soviet foreign policy, suggesting that the driving factor was ideational change, rather than material interests justified by ad hoc and post hoc changes in stated ideas. In this regard, the ideational explanation survives a *hoop test*: if changes in Soviet leaders' ideas motivated changes in their policies, rather than being merely rationalizations for policy changes adopted for instrumental reasons, then changes in these ideas had to precede those in behavior (Bennett 1999: 351–2).

Tannenwald's third question, on why some ideas won out over others, is the one most effectively addressed by *hoop tests*. Here, although Snyder does not specifically apply his domestic politics argument to Soviet restraint in the use of force in 1989, his contention that the material interests of different sectors were the driving factor in Soviet policy appears to fail a *hoop test* (Snyder 1990). Outlining in early 1988 the (then) hypothetical future events that could in his view have caused a resurgence of the Stalinist coalition of the military and heavy industry, Snyder argued that the rise of anti-reform Soviet leaders would become much more likely if Gorbachev's reforms were discredited by poor economic performance and if the Soviet Union faced "a hostile international environment in which SDI [the Strate-

gic Defense Initiative] was being deployed, Eastern Europe was asserting its autonomy, and Soviet clients were losing their counterinsurgency wars in Afghanistan, Angola, and Ethiopia" (Snyder, 1987/88: 128). As it turned out, all these conditions were more than fulfilled within two years, except for the deployment of a working SDI system. Yet apart from the unsuccessful coup attempt of 1990, Soviet hardliners never came close to regaining power. Snyder's theory thus appears to have failed a *hoop test* when the developments he thought would bring the Stalinist coalition back to power indeed took place, but the Stalinists still did not prevail. Conversely, the learning explanation survives a *hoop test* on the basis of evidence that anti-interventionist ideas won out because they resonated with recent Soviet experiences, rather than because their advocates represented a materially powerful coalition.

Despite strong evidence that both material and ideational factors played a role in Soviet restraint in 1989, one variant of the material explanation appears to fail a *hoop test*. Two internal Soviet reports on the situation in Europe in early 1989, one by the International Department (ID) of the Soviet Communist Party and one by the Soviet Institute on the Economy of the World Socialist System (IEMSS in Russian), argued that a crackdown in Eastern Europe would have painful economic consequences for the Soviet Union, including sanctions from the West. The IEMSS report also noted the growing external debts of Soviet allies in Eastern Europe (Bennett 2005: 96–7). At the same time, these reports provide ample evidence for the learning explanation: the IEMSS report warns that a crackdown in Poland could lead to an "Afghanistan in the middle of Europe" (Bennett 2005: 101), and the ID report argues that "authoritarian methods and direct pressure are clearly obsolete . . . it is very unlikely we would be able to employ the methods of 1956 [the Soviet intervention in Hungary] and 1968 [the Soviet intervention in Czechoslovakia], both as a matter of principle, but also because of unacceptable consequences" (Bennett 2005: 97).

While both material and ideational considerations played a role, there is reason to believe that at least in one respect the former was not a factor in Gorbachev's thinking in the fall of 1989. In a meeting on October 31, 1989, just ten days before the Berlin Wall fell, Gorbachev was reportedly "astonished" at hearing from East German leader Egon Krenz that East Germany owed the West \$26.5 billion, almost half of which had been borrowed in 1989 (Zelikow and Rice 1995: 87). Thus, while Gorbachev was certainly concerned about Soviet economic performance, the claim that he was in part inhibited from using force in Eastern Europe because of the region's external debts appears to have failed a *hoop test* because almost up until the Berlin Wall fell, Gorbachev did not even know the extent of these debts.

In sum, the material decline explanation passes a *hoop test* by showing that a wide range of Soviet leaders acknowledged Soviet decline, and a *straw*

in the wind test on the timing of changes in Soviet foreign policy, but the variant of this explanation that stresses East German debts as a factor preventing the Soviet use of force in 1989 fails a *hoop test*. The learning explanation survives *hoop tests* in its expectations on which actors would espouse which foreign policy views, on the timing of changes in Soviet ideas and policies, and on why some ideas prevailed over others. The sectoral domestic politics explanation emerges as the weakest, having failed *hoop tests* on its predicted correlation of policy views and material interests and its expectations on which ideas would win out in which contexts.

CONCLUSION

Through process tracing, scholars can make valuable inferences if they have the right kind of evidence. "Right kind" means that some types of evidence have far more probative value than others. The evidence must strongly discriminate between alternative hypotheses in the ways discussed above. The idea of *hoop tests*, *smoking gun tests*, *doubly decisive tests*, and *straw in the wind tests* brings into focus some of the key ways in which this discrimination occurs. What matters is the relationship between the evidence and the hypotheses, not the number of pieces of evidence.

Process tracing is not a panacea for causal inference, as all methods of causal inference are potentially fallible. Researchers could fail to include an important causal variable in their analyses. Available evidence may not discriminate strongly between competing and incompatible explanations. Actors may go to great lengths to obscure their actions and motivations when these are politically sensitive, biasing available evidence. Yet with appropriate evidence, process tracing is a powerful means of discriminating among rival explanations of historical cases even when these explanations involve numerous variables.

11

On Types of Scientific Inquiry: The Role of Qualitative Reasoning

David A. Freedman

One type of scientific inquiry involves the analysis of large data sets, often using statistical models and formal tests of hypotheses. A moment's thought, however, shows that there must be other types of scientific inquiry. For instance, something has to be done to answer questions like the following. How should a study be designed? What sorts of data should be collected? What kind of a model is needed? Which hypotheses should be formulated in terms of the model and then tested against the data?

The answers to these questions frequently turn on observations, qualitative or quantitative, that give crucial insights into the causal processes of interest. Such observations generate a line of scientific inquiry, or markedly shift the direction of the inquiry by overturning prior hypotheses, or provide striking evidence to confirm hypotheses. They may well stand on their own rather than being subsumed under the systematic data collection and modeling activities mentioned above.

Such observations have come to be called "Causal Process Observations" (CPOs). These are contrasted with the "Data Set Observations" (DSOs) that are grist for statistical modeling (Collier, Brady, and Seawright, chap. 9, this volume). My object in this essay is to illustrate the role played by CPOs, and qualitative reasoning more generally, in a series of well-known episodes drawn from the history of medicine.

Why is the history of medicine relevant to us today? For one thing, medical researchers frequently confront observational data that present familiar challenges to causal inference. For another, distance lends perspective,

allowing gains and losses to be more sharply delineated. The examples show that an impressive degree of rigor can be obtained by combining qualitative reasoning, quantitative analysis, and experiments when those are feasible. The examples also show that great work can be done by spotting anomalies, and trying to understand them.

QUALITATIVE REASONING: CASE STUDIES FROM EPIDEMIOLOGY

Jenner and Vaccination

The setting is the English countryside in the 1790s. Cowpox, as will be clear from the name, is a disease of cows. The symptoms include sores on the teats. Those who milk the cows often became infected, with sores on their hands; by the standards of the time, the illness is rarely serious. In contrast, smallpox is one of the great killers of the 18th century.

In 1796, Edward Jenner took some matter from a cowpox sore on the hand of dairymaid Sarah Nelmes, and inserted it into the arm of an eight-year-old boy, "by means of two superficial incisions, barely penetrating the cutis, each about half an inch long." The boy was "perceptibly indisposed" on the ninth day, but recovered the following day. Six weeks later, Jenner inoculated him with matter taken from a smallpox pustule, "but no disease followed" (Jenner 1798, Case XVII).

Jenner published 23 case studies to demonstrate the safety and efficacy of "vaccination," as his procedure came to be called: *vacca* is the Latin term for cow, and *vaccinia* is another term for cowpox. Despite initial opposition, vaccination became standard practice within a few years, and Jenner achieved international fame. By 1978, smallpox had been eradicated.

What led Jenner to try his experiment? The 18th century view of disease was quite different from ours. The great Scottish doctor of the time, William Cullen, taught that most diseases were "caused by external influences—climate, foodstuffs, effluvia, humidity, and so on—and . . . the same external factors could cause different diseases in different individuals, depending on the state of the nervous system" (Porter 1997, 262).

Despite such misconceptions, it was known that smallpox could somehow be communicated from one person to another; moreover a person who contracted smallpox and survived was generally immune to the disease from that point on. As a preventive measure, patients could be deliberately infected (through scratches on the skin) with minute quantities of material taken from smallpox pustules, the idea being to induce a mild case of the disease that would confer immunity later.

This procedure was called "inoculation" or "variolation." It was not free of risk: serious disease was sometimes caused in the patient, and in people

who came into contact with the patient (smallpox is highly contagious). On the other hand, failure to inoculate could easily lead to death from smallpox.

By the early part of the 18th century, variolation had reached England. Jenner was a country doctor who performed variolations. He paid attention to two crucial facts—although these facts were not explicable in terms of the medical knowledge of his time. (i) People who had the cowpox never seemed to contract smallpox afterwards, whether they had been inoculated or not. (ii) Some of his patients who had been ill with cowpox in the past still wanted to be inoculated; such patients reacted very little to inoculation—

What renders the Cox-pox virus so extremely singular, is, that the person who has been thus affected is for ever after secure from the infection of the Small Pox; neither exposure to the variolous effluvia, nor the insertion of the matter into the skin, producing this distemper. (Jenner 1798, 6)

These two facts led him to a hypothesis: cowpox created immunity against smallpox. That is the hypothesis he tested, observationally and experimentally, as described above. In our terminology, Jenner vaccinated a boy (Case XVII) who showed no response to subsequent inoculation. Immunity to smallpox had been induced by the vaccination.

By “virus,” Jenner probably meant “contagious matter,” that being a standard usage in his time. Viruses in the modern sense were not to be discovered for another century. By a curious twist, smallpox and cowpox are viral diseases in our sense too.

Semmelweis and Puerperal Fever

The time is 1844 and the place is Vienna. The discovery of microbes as the cause of infectious disease will not be made for some decades. Ignac Semmelweis is an obstetrician in the First Division of the Lying-in Hospital, where medical students are trained. (Midwives are trained in the Second Division.) Pregnant women are admitted to one division or the other, according to the day of the week that they come to the hospital, in strict alternation. Mortality from “puerperal fever” is much higher in the First Division (Semmelweis 1981 [1860]: 356).

Eventually, Semmelweis discovers the cause. The medical students are doing autopsies, and then examining the “puerperae” (women who are giving birth, or who have just given birth). “Cadaveric particles” are thus transferred to the women, entering the bloodstream and causing infection. In 1847, Semmelweis institutes the practice of disinfection, and mortality plummets (Semmelweis 1981 [1860]: 393–4).

But how did Semmelweis make his discovery? To begin with, he has to reject conventional explanations, including “epidemic influences,” which meant something different then:

Epidemic influences . . . are to be understood [as] certain hitherto inexplicable, atmospheric, cosmic, telluric changes, which sometimes disseminate themselves over whole countrysides, and produce childbed fever in individuals predisposed thereto by the puerperal state. [“Telluric” means earthly.] Now, if the atmospheric-cosmic-telluric conditions of the City of Vienna are so disposed that they cause puerperal fever in individuals susceptible thereto as puerperae, how does it happen that these atmospheric-cosmic-telluric conditions over such a long period of years have carried off individuals disposed thereto as puerperae in the First Clinic, while they have so strikingly spared others also in Vienna, even in the same building in the Second Division and similarly vulnerable as puerperae?” (Semmelweis 1981 [1860]: 357).

The reasoning is qualitative; and similar qualitative arguments dispose of other theories—diet, ventilation, use of hospital linens, and so forth.

Now he has to discover the real cause. In 1847, his revered colleague Professor Kolletschka is accidentally cut with a knife used in a medico-legal autopsy. Kolletschka becomes ill, with symptoms remarkably similar to puerperal fever; then he dies. Again, qualitative analysis is crucial. Close attention to symptoms and their progression is used to identify Kolletschka’s illness with puerperal fever (Semmelweis 1981 [1860]: 391). Tracing of causal processes comes into play as well:

Day and night this picture of Kolletschka’s disease pursued me. . . . I was obliged to acknowledge the identity of the disease, from which Kolletschka died, with that disease of which I saw so many puerperae die. . . . I must acknowledge, if Kolletschka’s disease and the disease from which I saw so many puerperae die, are identical, then in the puerperae it must be produced by the self-same engendering cause, which produced it in Kolletschka. In Kolletschka, the specific agent was cadaveric particles, which were introduced into his vascular system [the bloodstream]. I must ask myself the question: Did the cadaveric particles make their way into the vascular systems of the individuals, whom I had seen die of an identical disease? This question I answer in the affirmative. (Semmelweis 1981 [1860]: 391–2)

The source of the infectious agent could also be a wound in a living person (Semmelweis 1981 [1860]: 396). Once the cause is discovered, the remedy is not far away: eliminate the infectious particles from the hands that will examine the puerperae. Washing with soap and water is insufficient, but disinfection with chlorine compounds is sufficient (Semmelweis 1981 [1860]: 392–96).

Semmelweis’ work was accepted by few of his contemporaries, due in

part to his troubled and disputatious personality, although his picture of the disease was essentially correct. Puerperal fever is a generalized infection, typically caused by bacteria in the group *Streptococcus pyogenes*. These bacteria enter the blood-stream through wounds suffered during childbirth (for instance, at the site where the placenta was attached). Puerperal fever can be—and today it generally is—avoided by proper hygiene.

Snow and Cholera

John Snow was a physician in Victorian London. In 1854, he demonstrated that cholera was an infectious disease, which could be prevented by cleaning up the water supply. The demonstration took advantage of a natural experiment. A large area of London was served by two water companies. The Southwark and Vauxhall company distributed contaminated water, and house-holds served by it had a death rate “between eight and nine times as great as in the houses supplied by the Lambeth company,” which supplied relatively pure water (Snow 1965 [1855]: 86, data in Table IX).

What led Snow to design the study and undertake the arduous task of data collection? To begin with, he had to reject the explanations of cholera epidemics that were conventional in his time. The predominant theory attributed cholera to “miasmas,” that is, noxious odors—especially odors generated by decaying organic material. Snow makes qualitative arguments against such explanations:

[Cholera] travels along the great tracks of human intercourse, never going faster than people travel, and generally much more slowly. In extending to a fresh island or continent, it always appears first at a sea-port. It never attacks the crews of ships going from a country free from cholera, to one where the disease is prevailing, till they have entered a port, or had intercourse with the shore. Its exact progress from town to town cannot always be traced; but it has never appeared except where there has been ample opportunity for it to be conveyed by human intercourse. (Snow 1965 [1855]: 2)

These phenomena are easily understood if cholera is an infectious disease, but hard to explain on the miasma theory. Similarly,

The first case of decided Asiatic cholera in London, in the autumn of 1848, was that of a seaman named John Harnold, who had newly arrived by the *Elbe* steamer from Hamburgh, where the disease was prevailing. . . . Now the next case of cholera, in London, occurred in the very room in which the above patient died. (Snow 1965 [1855]: 3)

The first case was infected in Hamburgh; the second case was infected by contact with dejecta from the first case, on the bedding or other furnishings

in that fatal room. The miasma theory, on the other hand, does not provide good explanations.

Careful observation of the disease led to the conclusion “that cholera invariably commences with the affection of the alimentary canal” (Snow 1965, 10). A living organism enters the body, as a contaminant of water or food, multiplies in the body, and creates the symptoms of the disease. Many copies of the organism are expelled from the body with the dejecta, contaminate water or food, then infect other victims. The task is now to prove this hypothesis.

According to Sir Benjamin Ward Richardson, who wrote the introduction to Snow’s book, the decisive proof came during the Broad Street epidemic of 1854:

[Snow] had fixed his attention on the Broad Street pump as the source and centre of the calamity. He advised the removal of the pump-handle as the grand prescription. The vestry [in charge of the pump] was incredulous, but had the good sense to carry out the advice. The pump-handle was removed and the plague was stayed. (Snow 1965 [1855]: xxxvi)

The pump-handle as the decisive test is a wonderful fable, which has beguiled many a commentator.

What are the facts? Contamination at the pump did cause the epidemic, Snow recommended closing the pump, his advice was followed, and the epidemic stopped. However, the epidemic was stopping anyway. Closing the pump had no discernible effect: the episode proves little. Snow explains this with great clarity (Snow 1965 [1855]: 40–55, see esp. Table I on p. 49 and the conclusory paragraph on pp. 51–2). Richardson’s account is therefore a classic instance of post hoc, ergo propter hoc.

The reality is more interesting than the fable. Snow was intimately familiar with the Broad Street area, because of his medical practice. He says,

As soon as I became acquainted with the situation and extent of this irruption of cholera, I suspected some contamination of the water of the much-frequented street-pump in Broad Street. . . . but on examining the water, on the evening of 3rd September, I found so little impurity in it of an organic nature, that I hesitated to come to a conclusion. (Snow 1965 [1855]: 38–39)

Snow had access to the death certificates at the General Register Office, and drew up a list of the cholera fatalities registered shortly before his inspection of the pump. He then made a house-to-house canvass (the death certificate shows the address of the deceased), and discovered that the cases clustered around the pump, confirming his suspicion. Later, he made a more complete tally of cholera deaths in the area. His “spot map” displays the locations of cholera fatalities during the epidemic, and the clustering is

apparent from the map (Snow 1965 [1855]: 44–45; Cholera Inquiry Committee 1855: 106–9).

However, there were a number of exceptions that had to be explained. For example, there was a brewery near the pump; none of the workers contracted the disease: why not? First, the workers drank beer; second, if water was desired, there was a pump on the premises (Snow 1965 [1855]: 10). For another example, a lady in Hampstead contracted cholera: why? As it turned out, she liked the taste of the water from the Broad Street pump, and had it brought to her house (Snow 1965 [1855]: 44). Snow gives many other such examples.

Snow's work on the Broad Street epidemic illustrates the power of case studies. His refutation of the usual explanations for cholera, and the development of his own explanation, are other indicators of the power of qualitative reasoning. The analysis of his natural experiment, referred to above, shows the power of simple quantitative methods and good research design. This was the great quantitative test of his theory that cholera was a waterborne infectious disease.

In designing the quantitative study, however, Snow made some key qualitative steps: (i) seeing that conventional theories were wrong, (ii) formulating the water hypothesis, and (iii) noticing that in 1852, the Lambeth company moved its intake pipe to obtain relatively pure water, while Southwark and Vauxhall continued to draw heavily contaminated water. It took real insight to see—*a priori* rather than *a posteriori*—that this difference between the companies allowed the crucial study to be done.

Snow's ideas gained some circulation, especially in England. However, widespread acceptance was achieved only when Robert Koch isolated the causal agent (*Vibrio cholerae*, the comma-shaped bacillus) during the Indian epidemic of 1883. Even then, there were dissenters, with catastrophic results in the Hamburg epidemic of 1892 (Evans 1987).

Inspired by Koch and Louis Pasteur, there was a great burst of activity in microbiology during the 1870s and 1880s. The idea that microscopic life-forms could arise by spontaneous generation was cast aside, and the germ theory of disease was given solid experimental proof. Besides the cholera vibrio, the bacteria responsible for anthrax (*Bacillus anthracis*) and for tuberculosis (*Mycobacterium tuberculosis*) were isolated, and a vaccine was developed against rabies. However, as we shall see in a moment, these triumphs made it harder to solve the riddle of beriberi. Beriberi is a deficiency disease, but the prestige of the new microbiology made investigators suspicious of any explanation that did not involve microorganisms.

Eijkman and Beriberi

Beriberi was endemic in Asia, from about 1750 until 1930 or so. Today, the cause is known. People need minute amounts (about one part per mil-

lion in the diet) of a vitamin called “thiamin.” Many Asians eat a diet based on rice, and white rice is preferred to brown.

Thiamin in rice is concentrated in the bran—the skin that gives rice its color. White rice is obtained by polishing away the skin, and with it most of the thiamin; what is left is further degraded by cooking. The diet is then deficient in thiamin, unless supplemented by other foods rich in that substance. Beriberi is the sequel.

In 1888, knowledge about vitamins and deficiency diseases lay decades in the future. That year, Christiaan Eijkman—after studying microbiology with Koch in Berlin—was appointed director of the Dutch Laboratory for Bacteriology and Pathology in the colony of Java, near the city now called Jakarta. His research plan was to show that beriberi was an infectious disease, with Koch’s methods for the proof.

Eijkman tried to infect rabbits and then monkeys with blood drawn from beriberi patients. This was unsuccessful. He then turned to chickens. He tried to infect some of the birds, leaving others as controls. After a time, many of his chickens came down with polyneuritis, which he judged to be very similar to beriberi in humans. (“Polyneuritis” means inflammation of multiple nerves.)

However, the treated chickens and the controls were equally affected. Perhaps the infection spread from the treated chickens to the controls? To minimize cross infection, he housed the treated chickens and the controls separately. That had no effect. Perhaps his whole establishment had become infected? To eliminate this possibility, he started work on another, remote experimental station—at which point, the chickens began recovering from the disease.

[Eijkman] wrote “something struck us that had escaped our attention so far.” The chickens had been fed a different diet during the five months in which the disease had been developing. In that period (July through November 1889), the man in charge of the chickens had persuaded the cook at the military hospital, without Eijkman being aware of it, to provide him with leftover cooked [white] rice from the previous day, for feeding to the birds. A new cook, who started duty on 21 November, had refused to continue the practice. Thirty years later, Eijkman was to say that “[the new cook] had seen no reason to give military rice to civilian hens.” (Carpenter 2000, 38)

In short, the chickens became ill when fed cooked, polished rice; they recovered when fed uncooked, unpolished rice. This was an accidental experiment, arranged by the cooks. One of Eijkman’s great insights was paying attention to the results, because the cooks’ experiment eventually changed the understanding of beriberi.

Eijkman’s colleague Adolphe Vorderman undertook an observational study of prisons, to confirm the relevance to humans. Where prisoners were

fed polished rice, beriberi was common; with a diet of unpolished rice, beriberi was uncommon. Beriberi is a deficiency disease, not an infectious disease.

The evidence may seem compelling, but that is because we know the answer. At the time, the picture was far from clear. Eijkman himself thought that white rice was poisonous, the bran containing the antidote. Later, he was to reverse himself: beriberi is an infectious disease, although a poor diet makes people (and chickens) more vulnerable to infection.

In 1896, Gerrit Grijns took over Eijkman's lab (Eijkman suffered from malaria, and had to return to Holland). Among other contributions, after a long series of careful experiments, Grijns concluded that beriberi was a deficiency disease, the missing element in the diet being concentrated in rice bran—and in other foods like mung beans.

In 1901, Grijn's colleague Hulshoff Pol ran a controlled experiment at a mental hospital, showing that mung beans prevented or cured beriberi. In three pavilions out of twelve, the patients were fed mung beans; in three pavilions, other green vegetables. In three pavilions, there was intensive disinfection, and three pavilions were used as controls. The incidence of beriberi was dramatically lower in the pavilions with mung beans.

Still, medical opinion remained divided. Some public health professionals accepted the deficiency hypothesis. Others continued to favor the germ theory, and still others thought the cause was an inanimate poison. It took another ten years or so to reach consensus that beriberi was a deficiency disease, which could be prevented by eating unpolished rice, or enriching the diet in other ways. From a public health perspective, the problem of beriberi might be solved, but the research effort turned to extracting the critical active ingredient in rice bran—no mean challenge, since there is about one teaspoon of thiamin in a ton of bran.

Around 1912, Casimir Funk coined the term "vitamines," later contracted to vitamins, as shorthand for "vital amines." The claim that he succeeded in purifying thiamin may be questionable. But he did guess that beriberi and pellagra were deficiency diseases, which could be prevented by supplying trace amounts of organic nutrients.

By 1926, B. C. P. Jansen and W. F. Donath had succeeded in extracting thiamin (vitamin B1) in pure crystal form. Ten years later, Robert R. Williams and his associates managed to synthesize the compound in the lab. In the 1930s, there were still beriberi cases in the East—and these could be cured by injecting a few milligrams of the new vitamin B1.

Goldberger and Pellagra

Pellagra was first observed in Europe in the eighteenth century by a Spanish physician, Gaspar Casal, who found that it was an important cause of ill-

health, disability, and premature death among the very poor inhabitants of the Asturias. In the ensuing years, numerous . . . authors described the same condition in northern Italian peasants, particularly those from the plain of Lombardy. By the beginning of the nineteenth century, pellagra had spread across Europe, like a belt, causing the progressive physical and mental deterioration of thousands of people in southwestern France, in Austria, in Rumania, and in the domains of the Turkish Empire. Outside Europe, pellagra was recognized in Egypt and South Africa, and by the first decade of the twentieth century it was rampant in the United States, especially in the south. . . . (Roe 1973: 1)

Pellagra seemed to hit some villages much harder than others. Even within affected villages, many households were spared, but some had pellagra cases year after year. Sanitary conditions in diseased households were primitive: flies were everywhere. One blood-sucking fly (*Simulium*) had the same geographical range as pellagra, at least in Europe; and the fly was most active in the spring, just when most pellagra cases developed. Many epidemiologists concluded the disease was infectious, and—like malaria or yellow fever—was transmitted from one person to another by insects.

Joseph Goldberger was an epidemiologist working for the U. S. Public Health Service. In 1914, he was assigned to work on pellagra. Despite the climate of opinion described above, he designed a series of observational studies and experiments showing that pellagra was caused by a bad diet, and is not infectious. The disease could be prevented or cured by foods rich in what Goldberger called the P-P (pellagra-preventive) factor.

By 1926, he and his associates had tentatively identified the P-P factor as part of the vitamin B complex. By 1937, C. A. Elvehjem and his associates had identified the P-P factor as niacin, also called vitamin B3 (this compound had been discovered by C. Huber around 1870, but its significance had not been recognized). Since 1940, most of the flour sold in the United States has been enriched with niacin, among other vitamins.

Niacin occurs naturally in meat, milk, eggs, some vegetables, and certain grains. Corn, however, contains relatively little niacin. In the pellagra areas, the poor ate corn—and not much else. Some villages and some households were poorer than others, and had even more restricted diets. That is why they were harder hit by the disease. The flies were a marker of poverty, not a cause of pellagra.

What prompted Goldberger to think that pellagra was a deficiency disease rather than an infectious disease? In hospitals and asylums, the inmates frequently developed pellagra, the attendants almost never—which is unlikely if the disease is infectious, because the inmates could infect the attendants. This observation, although far from definitive, set Goldberger on the path to discovering the cause of pellagra and methods for prevention or cure. The qualitative thinking precedes the quantitative investigation.

Pellaga is virtually unknown in the developed world today, although it remains prevalent in some particularly poor countries.

Fleming and Penicillin

Alexander Fleming was working at St. Mary's Hospital in London, under the direction of Sir Almroth Wright, studying the life cycle of staphylococcus (bacteria that grow in clusters, looking under the microscope like clusters of grapes). Fleming had a number of plates on which he was growing staphylococcus colonies. He left the plates in a corner of his office for some weeks while he was on holiday. When he returned, one of the plates had been contaminated by mold. So far, this is unremarkable. He noticed, however, "that around a large colony of a contaminating mould the staphylococcus colonies became transparent and were obviously undergoing lysis" (Fleming 1929: 226).

Bacteria "lyse" when their cell walls collapse. What caused the lysis? Rather than discarding the plate—the normal thing to do—Fleming thought that the lysis was worth investigating. He did so by growing the mold in broth, watching its behavior, and trying filtered broth on various kinds of bacteria. The mold, a species of *Penicillium*, generated a substance that "to avoid the repetition of the rather cumbersome phrase 'mould broth filtrate' [will be named] 'penicillin'" (Fleming 1929: 227). It was the penicillin that caused the bacteria to lyse. Fleming showed that penicillin destroyed—or at least inhibited the growth of—many kinds of bacteria besides staphylococcus.

Penicillin's therapeutic potential went unrealized until Howard Florey and his associates at Oxford took up the research in 1938 and found processes for purification and larger-scale production. Due to the exigencies of World War II, much of the work was done in the U. S., where a strain of *Penicillium* that gave high yields was found on a moldy cantaloupe at a market in Peoria. (Industrial-scale development was being done at a nearby Department of Agriculture laboratory under the direction of Kenneth Raper, and people were encouraged to bring in moldy fruit for analysis.)

Penicillin was widely used to treat battlefield injuries, largely preventing gangrene, for example. Along with the sulfa drugs (prontosil was discovered by Gerhard Domagk in 1932) and streptomycin (discovered by Selman Waksman in 1944), penicillin was among the first of the modern antibiotics.

CONCLUSIONS

In the health sciences, there have been enormous gains since the time of Jenner, many of which are due to statistics. Snow's analysis of his natural

experiment shows the power of quantitative methods and good research design. Semmelweis' argument depends on statistics; so too with Goldberger. On the other hand, as the examples demonstrate, substantial progress also derives from informal reasoning and qualitative insights. Recognizing anomalies is important; so is the ability to capitalize on accidents. Progress depends on refuting conventional ideas if they are wrong, developing new ideas that are better, and testing the new ideas as well as the old ones. The examples show that qualitative methods can play a key role in all three tasks.

In Fleming's lab, chance circumstances generated an anomalous observation. Fleming resolved the anomaly and discovered penicillin. Semmelweis used qualitative reasoning to reject older theories about the cause of puerperal fever, to develop a new theory from observations on a tragic accident, and to design an intervention that would prevent the disease. The other examples lead to similar conclusions.

What are the lessons for methodologists in the 21st century? Causal inference from observational data presents many difficulties, especially when underlying mechanisms are poorly understood. There is a natural desire to substitute intellectual capital for labor, and an equally-natural preference for system and rigor over methods that seem more haphazard. These are possible explanations for the current popularity of statistical models.

Indeed, far-reaching claims have been made for the superiority of a quantitative template that depends on modeling—by those who manage to ignore the far-reaching assumptions behind the models. However, the assumptions often turn out to be unsupported by the data (Duncan 1984a; Berk 2004; Freedman 2005; chaps. 1 and 9, this volume). If so, the rigor of advanced quantitative methods is a matter of appearance rather than substance.

The historical examples therefore have another important lesson to teach us. Scientific inquiry is a long and tortuous process, with many false starts and blind alleys. Combining qualitative insights and quantitative analysis—and a healthy dose of skepticism—may provide the most secure results.

FURTHER READING

Brady, Collier, and Seawright (chaps. 1 and 9, this volume) compare qualitative and quantitative methods for causal inference in the social sciences. As they point out,

it is difficult to make causal inferences from observational data, especially when research focuses on complex political processes. Behind the apparent

precision of quantitative findings lie many potential problems concerning equivalence of cases, conceptualization and measurement, assumptions about the data, and choices about model specification. (22 this volume)

These authors recommend using a diverse mix of qualitative and quantitative techniques in order to exploit the available information; no particular set of tools is universally best. Causal process observations (including anomalies and results of accidental experiments, even experiments with $N = 1$) can be extremely helpful, as they were in the epidemiological examples discussed here.

The role of anomalies in political science is also discussed by Rogowski (chap. 5, this volume). He suggests that scholars in that field may be excessively concerned with hypothesis testing based on statistical models. Scholars may underestimate the degree to which the discovery of anomalies can overturn prior hypotheses and open new avenues of investigation. Anomalies that matter have been discovered in case studies—even when the cases have been selected in ways that do considerable violence to large- N canons for case selection. He also suggests that failure to search for anomalies can lead to a kind of sterility in research programs.

Scientific progress often begins with inspired guesswork. On the other hand, if guesses cannot be verified, progress may be illusory. For example, Snow (1965 [1855]: 125–33) theorized that—by analogy with cholera—plague, yellow fever, dysentery, typhoid fever, and malaria (which he calls “ague” or “intermittent fever”) were waterborne infectious diseases. His supporting arguments were thin. As it turns out, these diseases are infectious; however, only dysentery and typhoid fever are waterborne.

Proof for dysentery and typhoid fever, and disproof for the other diseases, was not to come in Snow’s lifetime. Although William Budd (1873) made a strong case on typhoid fever, reputable authors of the late 19th century still denied that such diseases were infectious (Bristowe and Hutchinson 1876: 211, 629; Bristowe et al. 1879: 102–3). In the following decades, evidence from epidemiology and microbiology settled the issue.

Plague is mainly spread by fleas, although transmission by coughing is also possible. The causal agent is the bacterium *Yersinia pestis*. Yellow fever and malaria are spread by mosquitoes. Yellow fever is caused by a virus. Malaria is caused by several species of *Plasmodium*, one-celled organisms with a nucleus and an extravagantly complicated life-cycle spent partly in humans and partly in mosquitoes. The medieval Black Death is usually identified with modern plague, but this is still contested by some scholars (Nutton 2008).

Buck et al. (1989) reprints many of the classic papers in epidemiology; some classic errors are included too. Porter (1997) is a standard reference on history of medicine. Jenner’s papers are reprinted in Eliot (1910

[1897]). Bazin (2000) discusses the history of smallpox, Jenner's work, and later developments, including the eradication of smallpox; the last recorded cases were in 1977–78. There is a wealth of additional information on the disease and its history in Fenner et al. (1988).

Inoculation was recorded in England by 1721 (Bazin 2000: 13; Fenner et al. 1988: 214–6). However, the practice was described in the journals some years before that (Timonius and Woodward 1714). It was a common opinion in Jenner's time that cowpox created immunity to smallpox (Jenner 1801; Baron 1838: 122). Over the period 1798–1978, techniques for producing and administering the vaccine were elaborated. As life spans became longer, it became clear that—contrary to Jenner's teachings—the efficacy of vaccination gradually wore off. Revaccination was introduced. By 1939, the virus in the vaccines was a little different from naturally-occurring cowpox virus. The virus in the vaccines is called "vaccinia" (Bazin 2000: chap. 11; Fenner et al. 1988: chaps. 6–7, esp. 278).

Bulloch (1938) reviews the history of bacteriology. Bacteria were observed by Hooke and Leeuwenhoek before 1700. Otto Friderich Müller in Denmark developed a workable classification before 1800, improved about 50 years later by Ferdinand Cohn in Germany.

Some of Koch's work on anthrax was anticipated by Pierre François Rayer and Casimir-Joseph Davaine in France. Likewise, Pasteur's experiments disproving spontaneous generation built on previous work by others, including Lazzaro Spallanzani; contemporaneous research by John Tyndall should also be mentioned.

Freedman (2005: 6–9) reports on Snow and cholera. For detailed information on Snow's work, see Vinten-Johansen et al. (2003). Evans (1987) gives a historical analysis of the cholera years in Europe. Koch's discovery of the vibrio was anticipated by Filippo Pacini in 1854, but the implications of Pacini's work were not recognized by his contemporaries.

Henry Whitehead was a clergyman in the Soho area. He did not believe that the Broad Street pump—famous for the purity of its water—was responsible for the epidemic. He saw a gap in Snow's argument: the fatalities cluster around the pump, but what about the population in general? Whitehead made his own house-to-house canvass to determine attack rates among those who drank water from the pump and those who did not. Then he drew up a 2×2 table to summarize the results. The data convinced him that Snow was correct (Cholera Inquiry Committee 1855: 121–33). Snow made this kind of analysis only for his natural experiment.

William Farr, statistical superintendent of the General Register Office, was a leading medical statistician in Victorian England and a "sanitarian," committed to eliminating air pollution and its sources. He claimed that the force of mortality from cholera in an area was inversely related to its eleva-

tion. More specifically, if γ is the death rate rate from cholera in an area and x is its elevation, Farr proposed the equation

$$\gamma = \frac{a}{b + x}$$

The constants a and b were estimated from the data. For 1848–49, the fit was excellent.

Farr held the relationship to be causal, explained by atmospheric changes, including attenuation of noxious exhalations from the Thames, changes in vegetation, and changes in the soil. After the London epidemic of 1866, however, he came to accept substantial parts of Snow's theory—without abandoning his own views about miasmas and elevation (Humphreys 1885: 341–84; Eyles 1979: 114–22; Vinten-Johansen et al. 2003: 394).

For better or worse, Farr's belief in mathematical symbolism had considerable influence on the development of research methods in medicine and social science. Furthermore, the tension between the pursuit of social reform and the pursuit of truth, so evident in the work of the sanitarians, is still with us.

There are two informative web sites on Snow, Whitehead, and other major figures of the era (these sites were active as of January 8, 2010):

<http://www.ph.ucla.edu/epi/snow.html>

<http://johnsnow.matrix.msu.edu/index.php>

Loudon (2000) is highly recommended on puerperal fever; but also see Nuland (1979) for a more sympathetic account of Semmelweis' life. Hare (1970: chap. 7) discusses efforts to control puerperal fever in a London maternity hospital in the 1930s. The strain of *Staphylococcus pyogenes* causing the disease turned out to be a common inhabitant of the human nose and throat (Loudon 2000: 201–4).

A definitive source on beriberi, only paraphrased here, is Carpenter (2000). He gives a vivid picture of a major scientific advance, including discussion of work done before Eijkman arrived in Java.

The discussion of pellagra is based on Freedman, Pisani, and Purves (2007: 15–16). Goldberger's papers are collected in Terris (1964). Goldberger (1914) explains the reasoning that led him to the deficiency-disease hypothesis; Goldberger et al. (1926) identifies the P-P factor as part of the vitamin B complex. Carpenter (1981) reprints papers by many pellagra researchers, with invaluable commentary. He explains why in Mexico a corn-based diet does not lead to pellagra, discusses the role of tryptophan (an amino acid that can be converted to niacin in the body), and points

out the gaps in our knowledge of the disease and the reasons for its disappearance.

An excellent source on Fleming is Hare (1970), with Goldsmith (1946) adding useful background. Today, “penicillin” refers to the active ingredient in Fleming’s mold broth filtrate. What is the cell-killing mechanism? In brief, cell walls of most bacteria include a scaffolding constructed from sugars and amino acids. Components of the scaffolding have to be manufactured and assembled when the cells are dividing to form daughter cells. In many species of bacteria, penicillin interferes with the assembly process, eventually causing the cell wall to collapse (Walsh 2003).

Some species of bacteria manufacture an enzyme (“penicillinase”) that disables penicillin—before the penicillin can disable the cell. There are other bacterial defense systems too, which explain the limits to the efficacy of penicillin. Penicillin inhibits cell wall synthesis by a process that is reasonably well understood, but how does inhibition cause lysis? That is still something of a mystery, although much has been learned (Walsh 2003: 41; Bayles 2000; Giesbrecht et al. 1998).

Penicillin only causes lysis when bacteria are dividing. For this reason among others, a rather unusual combination of circumstances was needed to produce the effect that Fleming noticed on his Petri dish (Hare 1970: chap. 3). Was Fleming merely lucky? Pasteur’s epigram is worth remembering: “Dans les champs de l’observation, le hasard ne favorise que les esprits préparés.”¹

Almroth Wright, Fleming’s mentor, was one of the founders of modern immunology (Dunnill 2001). Among other accomplishments, he developed a vaccine that prevented typhoid fever. Wright was a close friend of George Bernard Shaw’s, and was the basis for one of the characters in *The Doctor’s Dilemma*.

1. This may be liberally translated as, “In the practice of observation, chance favors only the prepared mind.”

12

Data-Set Observations versus Causal-Process Observations: The 2000 U.S. Presidential Election

Henry E. Brady

Data-set observations (DSOs) and causal-process observations (CPOs) provide two alternative foundations for causal inference. DSOs are the familiar data set of quantitative scholars, and research based on such data involves standard regression techniques and numerous variants on regression. CPOs, by contrast, are diagnostic “nuggets” of data that make a strong contribution to causal inference. The search for CPOs is a form of detective work that we call process tracing (Bennett, chap. 10, this volume), which seeks to establish the “physical and social processes through which purported causes affect outcomes” (Bennett and George 1997b: 3).

This chapter critically evaluates a study based on DSOs, and compares its conclusions with those of an analysis of CPOs.¹ The substantive objective is to resolve one of the many controversies over the 2000 presidential election in Florida: the disputed outcome in the Florida Panhandle, which unlike the rest of the state is on Central Time. This led to a difference in timing that was at the center of the dispute.

1. For definitions of these two types of observations, see chapter 9 and the glossary in the present volume.

THE OPTION OF DSOs AND REGRESSION ANALYSIS

John R. Lott argues that in the 2000 U.S. presidential election, at least 10,000 votes were lost for George W. Bush in the ten panhandle counties of Florida.² The votes were lost because the networks declared Al Gore the winner in Florida after the polls had closed in eastern Florida but before the polls had closed in the panhandle counties, which are on Central Standard Time. Lott's conclusion was widely discussed in the aftermath of the 2000 election and led to a series of congressional hearings.

To get his result, Lott employed a "difference-in-differences" form of regression analysis, based on data-set observations.³ He obtained turnout data on all sixty-seven Florida counties for four presidential elections (1988, 1992, 1996, and 2000), and he estimated a time-series cross-sectional regression with fixed county and time effects and with a "dummy variable" for the ten panhandle counties. In effect, Lott looked at the difference between one set of counties that got a "treatment" in the year 2000 (the ten panhandle counties whose polls were still open when the election was "called") and those that did not (the remaining fifty-seven Florida counties in the eastern time zone), while controlling for differences reflected in the data from previous elections. Lott (2000) concluded that:

By prematurely declaring Gore the winner shortly before polls had closed in Florida's conservative western Panhandle, the media ended up suppressing the Republican vote. . . . An examination of past Republican presidential votes by county in Florida from 1988 to 2000 shows that while total votes declined, the Republican voting rate in the western panhandle was significantly suppressed relative to the non-Republican vote. The 4 percent greater reduction in Republican votes averages about 1,000 votes per county, [yielding] 10,000 Republican votes for all 10 counties in the western Panhandle. This holds true even after accounting for the average differences in voting rates across counties as well as the changes in voting rates from one election to another.

2. This discussion is based on three sources. The first is Lott's article in the November 14, 2000, *Philadelphia Inquirer* (Lott 2000) in which he provides a general description of his methodology and claims that 10,000 votes were lost. Second, Lott's econometric analysis is described in Mason, Frankovic, and Jamieson (2001: 77–78). Third, Congressman Billie Tauzin subsequently held hearings on the elections and collected different analyses and interpretations of the vote. Congressman Tauzin's office provided me with an annotated computer printout of Lott's analysis, which reflects a methodology identical to that described both in Lott's article and in Mason et al.

3. "Difference-in-differences" estimators are widely used in economics, and they are now a staple of introductory econometrics textbooks such as Stock and Watson (2003: 385–88) and Wooldridge (2009: chap. 13.2).

TURNING TO CAUSAL-PROCESS OBSERVATIONS

A researcher accustomed to the exclusive use of data-set observations might stop at this point, convinced that an adequate inference had been made. However, researchers oriented toward the use of causal-process observations would ask whether the result makes any sense. Is Lott's estimate reasonable, given the number of voters who had not yet voted when the media called the election for Gore? How many of these voters heard the call? Of these, how many decided not to vote? And of those who decided not to vote, how many would have voted for Bush? Researchers can obtain answers to these questions by consulting diverse data sources and constructing a more intricate characterization of events on election day.

An inquiry to the networks established that the media calls were made ten minutes before the panhandle polls closed at 7:00 p.m.—twelve hours after the opening time of 7:00 a.m. If we assume that voters go to the polls at an even rate throughout the day, then only 1/72nd (ten minutes over twelve hours) of the voters had not yet voted when the media call was made. Alternatively, an analysis of Census data from 1996 on time of voting suggests that no more than about one-twelfth of the voters in Florida come to the polls in the last hour. If we assume that voters go to the polls at an even rate in this last hour, then (once again) only 1/72nd (one-sixth of one hour times one-twelfth) of the voters had not yet voted when the media call was made. Of the 379,000 voters in the panhandle, about 20 percent were absentee voters—leaving about 303,000 voters who voted on election day. One seventy-second of this figure is, in round numbers, 4,200 voters. The major assumption in this calculation is that voters come to the polls uniformly during the day or during the last hour. Interviews with Florida election officials and a review of media reports suggest that, typically, no rush to the polls occurs at the end of the day in the panhandle.

Only 4,200 people could have been swayed by the media call of the election, if they heard it. How many heard it? Research on media exposure suggests that an audience of 20 percent of adults for all media outlets would be very large. To be very conservative, I will assume that 20 percent of the 4,200 voters who intended to vote in the last ten minutes, or 840 people, heard the early call—though this is undoubtedly an overestimate because not all media were reporting the elections. Moreover, many of these prospective voters were Democrats or Independents who would not have voted for George W. Bush. In the panhandle, the Bush vote was about two-thirds of the total. If we assume the same proportion among those who were still to vote, it yields a total of 560 Bush voters who might have been affected.

Of these 560 Bush voters who might have heard the media call, how many decided not to vote? A review of past work on the impact of early

calls (Jackson 1983) and a general knowledge of voting behavior suggests a figure of 10 percent for the fraction of voters who decided not to vote once they knew the call was made for the presidential election. After all, voters select other officials as well, and they vote for reasons other than the likelihood that their vote will be decisive. Ten percent of 560 yields fifty-six Bush voters who might have been deterred from voting.

This estimate of Bush's vote loss still probably exceeds the actual net effect. It seems just as likely that a Gore voter, rather than a Bush voter, might have decided not to vote. After all, for both candidates, the vote is no longer relevant to the presidential election once the call has been made. If 10 percent of the 280 Gore voters did not vote, then the net effect would be 28 Bush votes—56 Bush voters minus 28 Gore voters. This suggests a range of 28 to 56 Bush votes lost depending upon whether Gore voters were affected by the call. Even if we forget the offset for Gore voters and quadruple the estimate of 56 Bush voters who might have decided not to vote, the resulting upper-bound estimate of 224 voters is far short of the 10,000 that Lott claims.

My detective work leads to the inference that the approximate upper bound for Bush's vote loss was 224 and that the actual vote loss was probably closer to somewhere between 28 and 56 votes. Lott's figure of 10,000 makes no sense at all. This simple case-study analysis based upon information that goes beyond the turnout data used in the difference-in-differences model suggests a figure that is two orders of magnitude smaller than Lott's result.

Although this case study of late voting uses quantitative data, it employs inferential tools typically associated with qualitative research. It draws upon multiple sources of information, utilizing inferences based on common sense, to establish an argument. It tries to approach the problem in several different ways, cross-checking information at every turn, and asking if the posited causal effect is probable, or even possible, given what we know from many different sources. In short, it investigates causal processes in close detail, and it tries to get beyond the results of an elaborate quantitative analysis of data-set observations.

WHERE DID LOTT GO WRONG?

The difference-in-differences method is widely used in economics and other social science disciplines as a way to adjust observational data for confounding factors that can lead to incorrect inferences. In this case, the method assumes that turnout in 2000 can be predicted by turnout in past years after adjusting for idiosyncratic factors of two types: those factors that affect each county in the same way over the entire time period but vary

from county to county (county fixed effects), and those factors that affect all counties in a given year but vary over years (time fixed effects).

This method does badly when idiosyncratic factors vary both by county and over time. For example, in 2000, organized labor put significant effort into increasing turnout in Florida, and it seems likely that it put its effort into mobilizing Democratic voters. As a result, turnout would be increased, compared to prior years, in counties with more Democrats (namely those outside the panhandle). The difference-in-difference method would not control for this. In fact, it would presume that the higher turnout outside the panhandle in 2000 should be translated into higher turnout inside the panhandle as well. To the extent that this higher turnout was not realized, Lott's equation would pick it up as a negative coefficient on his dummy variable for the panhandle counties that he interpreted as the effect of the early media call. Instead, his coefficient might simply reflect labor's success in mobilizing voters outside the panhandle.

In addition, quantitative methods are most believable when researchers are conservative about their inferences. Instead of using the standard .05 level of significance, Lott chose to use a .10 level, and he chose to employ a one-sided test that made his t-statistic of 1.285 just significant at this 10 percent level. This lenient approach to hypothesis testing allowed him to claim that his regression detected a significant effect. However, if Lott had decided to provide a 10 percent one-sided confidence interval for his estimate instead of a point estimate of 10,000, his confidence interval would have gone from zero to 20,000, thus providing little confidence in his assertions.

Even if these problems in Lott's analysis were cleaned up by getting data on labor union activity and other factors, the analysis of such data would not necessarily supercede the inference based on causal-process observations. Even after putting aside the practical problems of collecting suitable data, it would be hard to collect data that could rule out all of the possible confounding effects. Consequently, rather than seeking additional data-set observations, in my judgment it would be more productive to do further in-depth analysis of causal-process observations drawn from these ten Florida panhandle counties, finding out what happened there, for example, by interviewing election officials and studying media reports.

CONCLUSION

Causal-process observations show that it was highly implausible for the media effect suggested by Lott's analysis to have occurred. From a technical perspective, CPOs might be seen as a less sophisticated tool of analysis, yet

they effectively demonstrate that Lott's quantitative conclusions based on regression analysis cannot be valid.

This chapter thus seeks to demonstrate the value of causal-process observations in what could be seen as a "least-likely case," that is, a data-rich domain of mass political behavior. Even in this domain, this strategy of causal assessment provides valuable inferential leverage that supplements, and in this instance contradicts, the conclusions based on the analysis of data-set observations. Indeed, the lesson for quantitative researchers is the necessity of paying attention to the causal processes underlying behavior. Otherwise, regression analysis is likely to jump off the rails.

Addendum: Teaching Process Tracing

David Collier

Quantitative researchers receive extensive training in the spectrum of statistical tools employed in their research. By contrast, notwithstanding extensive efforts to institutionalize training in qualitative methods,¹ techniques such as process tracing are not adequately taught. This deficit has motivated us to incorporate in this volume the three prior chapters on process tracing.

Looking beyond these chapters, those concerned with graduate training in methodology should devote more systematic attention to process tracing. Given this need, we have included with the four online chapters on the Rowman & Littlefield website a set of exercises for teaching these analytic tools. To reiterate, this online material can be accessed using the instructions on the copyright page of this book.

1. Notable among these efforts are the Institute for Qualitative and Multi-Method Research, held annually at Syracuse University, and the Qualitative and Multi-Method Organized Section of the American Political Science Association. In addition to sponsoring panels at the annual meetings, the section offers short courses on qualitative methods. These trends are discussed in Collier and Elman (2008).

F. QUANTITATIVE TOOLS FOR CAUSAL INFERENCE

13

Regression-Based Inference: A Case Study in Failed Causal Assessment

Jason Seawright

A recurring theme in this volume is the danger of exaggerated expectations about the contributions of conventional regression-based research. This chapter explores this problem through examining the use of regression in quantitative, cross-national studies focused on a classic question of great normative and theoretical importance: What is the relationship between democracy and economic growth? The goal is to treat this literature as a case study in failed causal inference and to draw appropriate methodological lessons.

I first examine the remarkably inconsistent findings on this topic in the literature that employs conventional cross-national regression analysis. The chapter then briefly considers whether refinements on conventional techniques solve these problems. I argue they do not. The final section offers scholars the option of “scaling down” their focus. They might consider moving away from cross-national regression analysis and focusing instead on the kind of quantitative, qualitative, and case-study evidence that provides insight into the causal mechanisms often posited as crucial to the connection between democracy and economic growth. The approach of scaling down abandons broad quantitative comparisons that have the apparent virtue of achieving generality that rises above the analysis of a small or medium number of cases. In fact, this apparent generality is too often illu-

The author would like to thank Tara Buss, Chris Chambers-Ju, Maria Gould, Annette Konoske-Graf, Taryn Nelson, Neal Richardson, and Miranda Yaver.

sory—the point of our case study of failed causal inference—and far more limited analytic goals may open a better path toward genuine insight. In large-N, cross-national literatures that are unable to meet statistical assumptions—as is true here—an appropriate intermediate goal is to explore whether basic sequences of intervening variables in the given substantive domain are empirically plausible in one or a few cases.

The failures of quantitative, cross-national analysis I explore are quite general. This form of analysis is a major research tool both in political science¹ and more widely in the social sciences. Over several decades, scholars have viewed this form of quantitative comparison as a powerful means for testing theories, accumulating knowledge, and moving toward findings of broad generality.² An immense number of articles,³ published in scholarly journals across several disciplines, have reported findings from this kind of study, and the wide spectrum of substantive topics addressed is striking.⁴ Further, while this case study of failed causal inference focuses specifically on research at the national level, this same set of methods is common in

1. To note one of many possible examples, in 2009–10 eight of the comparative politics articles in the *American Journal of Political Science* adopt this approach.

2. For programmatic statements on this approach that have appeared over many years, see Gillespie and Nesvold 1971; Hoole and Zinnes 1976; Jackman 1985; and Dietz, Frey and Kalof 1987. Beck and Katz (2006) briefly recapitulate the recommendations arising from a discussion of quantitative, cross-national methods that extended over many years. Laitin (2002) and Fearon and Laitin (2008) underscore the importance of this approach as a key building block in multi-method comparative research.

3. For economics alone, Lindauer and Pritchett (2002: 19) refer to “thousands” of cross-national regressions that have focused on explaining economic growth.

4. Any list of the substantive areas in which scholars have carried out quantitative, cross-national research would—across several disciplines—be long. In a great many of these areas, one finds over 100, and sometimes several hundred, published studies using this approach. Such a list would include research focused on such (partially overlapping) topics as political and economic institutions, party and electoral systems, the success or failure of left parties, economic cycles and elections, women’s representation; armed conflict and deterrence, military coups, democracy, transitions to democracy, direct democracy, persistence of different regime types, political and economic freedom; ethnic politics, ethnic conflict, religion and politics, culture and politics; bureaucracy and corruption; constitutions, legal institutions, courts, property rights, the rule of law; comparative and international political economy, causes and consequences of neoliberalism, explanations of economic growth (a vast literature that goes far beyond the focus of the present chapter), market performance, natural resource wealth, the resource curse; welfare state, health, education, human capital, quality of life; international conflict, correlates of war, democratic peace; world systems and dependency theory; and, cutting across many of these areas, international policy diffusion.

subnational research that compares provinces, states, districts, counties, and municipalities.⁵ The arguments presented in this chapter apply broadly to regression-based modeling focused on all of these levels of analysis. For better or for worse, this vibrant research tradition certainly merits the attention it receives here.

The concerns raised in this chapter are also important, given the current attention to nested research designs in which one or a small number of cases—which may be analyzed through case-study research—are embedded in a large-N, quantitative analysis. These studies—and the attraction of doing this kind of multi-method research (very prominently, for example, in doctoral dissertations)—have given a strong new impetus to quantitative, cross-national work.

Nothing in this chapter calls into question the overall agenda of multi-method analysis, but it does have implications for the cross-national component in many multi-method designs. In particular, as argued here, the results of regression analysis routinely are fragile and grounded in what are often hard-to-defend statistical assumptions. For this reason, scholars should react with some skepticism to the frequently-articulated position that regression analysis should do the heavy lifting in multi-method work, establishing the overall relationships among variables; and that qualitative work should serve primarily to provide narratives that make more plausible the causal connections as a post hoc validation of the inference or as part of a search for mechanisms (Fearon and Laitin 2008: 756-7, 761). Instead, qualitative and case-study research should be seen as making a fundamental contribution to these designs, with cross-national regression analysis serving as a potentially useful tool that is routinely in need of substantial bolstering from other kinds of evidence.

In sum, the method of analysis and form of data under discussion here are highly salient to today's discipline of political science.

THE BASIC PROBLEM

The relationship between democracy and socioeconomic development has been an analytic conundrum at least since Aristotle. This is a vital topic in contemporary scholarship, and the literature is both lively and voluminous. Against this backdrop, few questions are as compelling and relevant to the world of practical politics as the impact of political regime type on economic growth. Authoritarian leaders often justify their regimes with the

5. Examples of subnational, quantitative-comparative work are found, for instance, in nine articles published in the *American Journal of Political Science* in 2006-10.

claim that they produce superior economic performance (Haggard and Kaufman 1995: 45–108). By contrast, other scholars have argued that “the two freedoms”—economic and political—are by nature inseparable (Friedman 1962); political democracy and economic growth naturally accompany one another, so authoritarianism is inherently inimical to economic growth. Consequently, adjudicating among these sets of claims has relevance to real-world political debates, as well as to theoretical concerns.

The quantitative literature on this topic is notorious for its inconsistent findings (Przeworski and Limongi 1993; Sirowy and Inkeles 1991). For example, some scholars find an overall positive relationship between democracy and economic growth.⁶ Thus, Leblang argues that “the newest evidence allows us to conclude that democracy is a more important cause of economic growth than previously believed” (1997: 352). Others find a negative, linear relationship (e.g., Feng 1997; Gasiorowski 2000). For example, Gasiorowski states that “we can therefore conclude that more-democratic regimes have slower growth than less-democratic regimes . . .” (2000: 342). A growing number of scholars suggest that there is no relationship at all between democracy and economic growth.⁷ According to Przeworski et al., “total output grows at the same rate under the two regimes [democracy and authoritarianism], both in poor countries and in wealthier countries” (2000: 179).

Barro (1997) finds a curvilinear relationship in which countries with partially democratic regimes grow more quickly than those with fully democratic or fully undemocratic regimes. He concludes that “the pattern of results—a positive coefficient on the linear term and a negative coefficient on the square—means that growth in democracy is increasing at low levels of democracy, but the relation turns negative once a moderate amount of political freedom has been attained” (1997: 58). Still other scholars have found mixed effects of democracy on economic growth, with positive and negative causal pathways, or positive and negative time periods.⁸

These findings point to a spectrum of possibilities. Variation across levels of democracy may (1) cause economic growth, (2) prevent economic growth, (3) be irrelevant to economic growth, (4) have a curvilinear relationship with growth, or (5) have other forms of mixed effects on growth.

This chapter identifies three principal reasons for these markedly incon-

6. E.g., Leblang (1997); Minier (1998); Nelson and Singh (1998); Shen (2002); Kurzman, Werum, and Burkhart (2002).

7. E.g., Przeworski and Limongi (1993); De Haan and Siermann (1995); Alesina et al. (1996); Brunetti (1997); Durham (1999); Przeworski et al. (2000); Glaeser et al. (2004).

8. E.g., Helliwell (1994); Baum and Lake (2003); Kriekhaus (2004); and Pinto and Timmons (2005).

sistent findings. First, researchers have been unable to select a set of control variables that would allow their analyses to meet basic statistical assumptions. Second, it seems impossible to escape the problem of reciprocal causation between democracy and economic growth—i.e., to find exogenous variance in degree of democracy. Without such independent variation, researchers cannot adequately control for confounding causes of economic growth or deal empirically with other causal complexities. Third, variations in findings may result from inconsistencies in time periods and data sources (Krieckhaus 2004).

Until these three problems are resolved, the statistical models employed to make inferences about the effects of democracy on economic growth will be untrustworthy.

To state this problem more broadly—echoing a recurring theme in the present volume—researchers have not taken seriously enough the absolute dependence of statistical inference on underlying assumptions. Addressing this concern requires analysts to present compelling justifications for their assumptions, and it pushes them to show in great detail how and why their own statistical models and findings should be given more credence than those of other scholars.

DEMOCRACY AND ECONOMIC GROWTH: DIVERGENT FINDINGS

This section reviews the successes and shortcomings of conventional regression analysis in studying the relationship between democracy and economic growth, and focuses on what researchers would need to know for regression analysis to succeed. To simplify analysis of the key points of causal inference, I will focus on the measure of democracy developed by Przeworski et al., with occasional reference to the Polity measure when a graded indicator of democracy is useful to illustrate methodological issues, as in figure 13.1.

Figure 13.1 presents the bivariate relationship between the Polity measure of democracy and economic growth between 1960 and 1990. The horizontal axis reports the different categories of democracy from the Polity measure, while the vertical axis reports the mean growth rate within each category. As this figure shows, the bivariate relationship is inconsistent, to say the least. For some increments on the measure of democracy, higher levels are associated with higher economic growth rates, while for other increments the relationship is reversed.

Yet on the basis of the figure, it is still possible to believe that there *could be* a relationship. The overall slope is non-zero, and a bivariate regression would suggest a slight negative effect of democracy on economic growth.

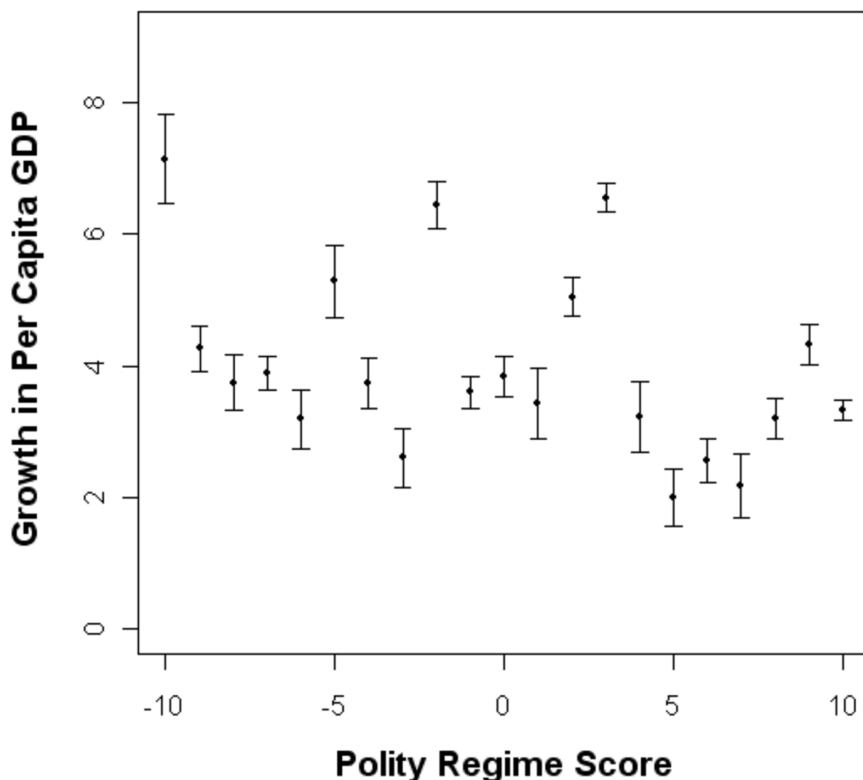


Figure 13.1. Mean Growth Rates by Level of Democracy

Horizontal axis: Polity 98, Measure of Democracy

Vertical axis: World Development Indicators (1998) Mean Income Growth

Time Period: 1960–1990

Explanatory Note: In this figure, the unit of observation is the country year (i.e., 218 countries times 31 years, or, after omitting observations with missing data, 4033 observations). Each of the dots represents the mean growth rate for all of the country years that correspond to a given polity score. The bars around the points are located two standard errors above and below the subgroup mean growth rate.

The various spikes and troughs in the graph could, perhaps, be the product of omitted variables, rather than evidence that there is no relationship. Given the overwhelming substantive interest driving research on this question, it is necessary to consider potential omitted variables, specifically confounders.⁹

9. Other issues are also important for evaluating a regression analysis, including error models, as well as additivity and linearity assumptions. Alternative estimators can sometimes reduce the importance of some of these assumptions, given a set of control variables that resolve all issues of confounding.

Confounders may be thought of as omitted variables that are correlated with both the dependent variable and one or more of the included independent variables, although this condition is technically neither necessary nor sufficient for a particular variable to be a confounder.¹⁰ A common approach to correcting for confounders is to add control variables to a regression equation. However, as the discussion will show, this approach can be problematic in practice.

Within the wide set of possible confounding factors, those referred to collectively as socioeconomic development are especially salient. For instance, one plausible confounder is a country-year's income level, or GDP per capita.¹¹ Macroeconomic theory argues that, other things being equal, a country with a higher income level will tend to grow more slowly than a country with a lower income level. The logic is as follows: other factors, such as capital stock and labor supply, determine a country's "target level" of income, or the level at which the country will be in an economic steady state (barring, of course, technological development or other changes that alter the "target"). After controlling for factors that influence the economy's steady state, a wealthier country has less scope for further growth before reaching the steady state (e.g. Barro and Sala-i-Martin 1995: 26–37). Therefore, a country's income level may influence its growth rate, and it is also hypothesized to be a cause of democracy (see discussion in Part 3 below). In other words, it has the classic characteristics of a confounding variable.

Table 13.1 reports the results of a regression of income growth rate on

Table 13.1. Economic Growth Regressed on Democracy and Lagged Income Level

OLS Regression Results	
Constant =	4.661 (0.145)
Democracy =	−0.00151 (0.029)
Lagged GDP (1000s) =	−0.132 (0.033)
R ² =	0.009
N =	3138

Note: Numbers in parentheses are standard errors.

10. An omitted variable is a factor that should be—but is not—included in a causal model, and this omission distorts the conclusions drawn about the relative importance of the variables that are included. For a further discussion and a more precise (but somewhat more difficult) definition, see Pratt and Schlaifer (1984).

11. A country's level of GDP per capita is, in a sense, a derivative of its economic growth rate, in that it is equal to the prior year's per capita GDP multiplied by 100% plus the prior year's growth rate. However, a country's growth in a given year is usually much smaller than the overall size of its economy, and growth rates differ a great deal from year to year. Therefore, level of per capita GDP and economic growth rate are in fact different variables.

lagged income level as reported by *World Development Indicators* (World Bank 1998) and the Polity democracy variable. We see here that, after controlling for lagged income level, the hypothesized relationship between democracy and economic growth is almost completely eliminated. The reason is that per capita GDP is highly correlated with democracy—with a coefficient of 0.565, which is statistically significant at a level smaller than 0.001—and its bivariate correlation with income growth is noticeable but far from overwhelming, with a coefficient of -0.074 at a significance level smaller than 0.001. In other words, the analysis reported in 13.1 suggests that income level, or GDP, is indeed a confounder for the relationship between democracy and economic growth. This simple finding provides the basic logic behind most research that reports no relationship between democracy and economic growth.

Yet this finding is not definitive. First, there may be serious issues of reciprocal causation, a common source of endogeneity.¹² Second, this result may still suffer from major problems of confounding. The control variable in this analysis taps only one of the factors involved in economic development: overall level of economic production. A more complete test would include additional control variables such as education level, labor costs, investment rates, and urbanization rates, all of which may affect both democracy and economic growth. Adding some or all of these controls would probably change the findings reported above.

However, many of these variables are hypothesized to have complex causal interrelationships with democracy. Thus, investment rates, average education levels, and labor costs may be both causes and consequences of democracy.¹³ Under these circumstances, failing to control for one of these variables exposes the analyst to the risk that the true relationship between democracy and growth rate will be confounded by the omitted variable. On the other hand, including these variables may also introduce bias into the model, creating an insurmountable paradox for the researcher.

Table 13.2 presents another regression illustrating a further cause for concern. Here, I add to the regression in table 13.1 a measure of education,

12. The term endogeneity indicates a correlation between an included independent variable and the error term of the regression equation. The error term represents the portion of the dependent variable not explained by the independent variables. Theoretically, if the specified model is correct, the error term will only capture randomness. A correlation between the error term and an included independent variable is a source of bias and can occur for a number of reasons, including the omission of an important independent variable, measurement error in the included independent variables, errors in sample selection, simultaneity (where variables are co-determined or reciprocally caused), as well as other factors.

13. E.g., Engerman, Mariscal, and Sokoloff (1998); Feng and Zak (1999); Rodrik (1999); Bourguignon and Verdier (2000).

Table 13.2. Economic Growth Regressed on Democracy, Lagged Income Level, and Primary Educational Attainment

OLS Regression Results	
Constant =	4.240 (0.253)
Democracy =	0.04371 (0.030)
Lagged GDP (1000s) =	-0.196 (0.033)
Primary School Attainment =	0.01349 (0.006)
R ² =	0.019
N =	2415

Note: Numbers in parentheses are standard errors.

drawn from Barro and Lee (1996). This measure taps another important facet of socioeconomic development, therefore producing a model that may more thoroughly account for confounders. Adding this additional control variable substantially changes the estimated effect of democracy on economic growth. After controlling for education and lagged income level, democracy is now estimated to have a moderate, positive relationship with economic growth. The result does not reach a standard threshold of statistical significance, but the change from the estimate in table 13.1 is remarkable.

It is essential to decide whether including a measure of educational attainment in the growth regression improves or harms the estimate of democracy's effect on economic growth. This, in turn, calls for choices about one's causal explanation: Does education cause democracy, and do either democracy or economic growth cause education? If any of these are true, the regression reported in table 13.2 may not, in fact, provide a better estimate of the effect of democracy on economic growth than the regression reported in table 13.1. If democracy causes education, for example, then the regression in table 13.2 misrepresents the effects of democracy on economic growth because it omits the indirect effect through education. While most scholars would not regard either of these models as sophisticated enough to capture the relevant causal structure, this issue is nonetheless worth considering carefully even for these very simple models. And of course, analysts must also confront these issues when constructing more complex models.

Unfortunately, the empirical evidence provides little leverage for answering these questions. The regression estimates give us only the correlations among democracy, education, and economic growth; the data do not tell us how to decompose these correlations into causal relations. For example, democracy and education may be correlated because democracy causes education, because education causes democracy, because a third variable causes both of them, or some combination of these. In other words, statistical evidence simply does not give us enough information to choose among the regression in table 13.1, the regression in table 13.2, and the bivariate relationship shown in figure 13.1.

The discussion presented up to this point has focused on standard regression analyses with only a few control variables. This simple approach has produced results that span the three categories of possible linear relationships between democracy and economic growth—democracy lowers economic growth rates, has no effect, and raises growth rates. The far more complex models employed in most published studies of democracy and economic growth are extensions of the three analyses just presented. Until we can decide on the basis of theory which of these three relatively simple analyses is best, it may make little sense to expend further energy developing elaborate statistical models of the relationship between democracy and economic growth, as these models would bring us no closer to understanding their true relationship.

A further, extremely important problem is that each of these simple models is compatible with more than one underlying causal structure. For example, the results presented in table 13.2 are themselves compatible with a wide range of alternative underlying causal structures. Figure 13.2 shows three alternative path diagrams for a few of the possible causal structures relating regime type, economic growth, average primary-school educational attainment (a potential control variable), and overall level of human and social capital (a latent variable,¹⁴ specifically an unmeasured confounder, represented by an oval).¹⁵

Each of these path diagrams represents a causal structure that could produce data consistent with the regression results in table 13.2, yet they have profoundly different implications for theory. The data analysis contributes nothing to distinguishing among these structures. Only in the world represented in figure 13.2a, where primary-school attainment is posited to be part of the mechanism relating the confounder to the outcome, can the regression of economic growth rates on regime type (controlling for primary-school attainment) point toward a helpful causal inference. In figure 13.2b, primary-school attainment is posited to be part of a mechanism relating regime type to economic growth (perhaps democracies invest more in education, which in turn affects overall economic production). Thus, controlling for it creates a gap between the empirical findings and the underlying causal structure by removing from the final result part of the real causal effect of regime on economic performance that operates through primary-school attainment. In figure 13.2c, primary-school attainment is

14. A latent variable is a theoretical construct that, based on theory, is believed to affect the dependent variable, but it is not directly measured in the dataset at hand.

15. The figures do not incorporate a second control variable, per capita GDP, which for present purposes is assumed to be relevant.

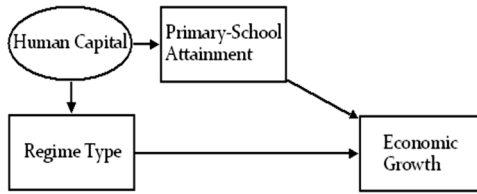


Fig. 13.2a. Schooling as intervening between human capital and growth

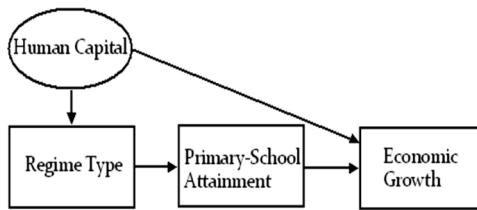


Fig. 13.2b. Schooling as intervening between democracy and growth

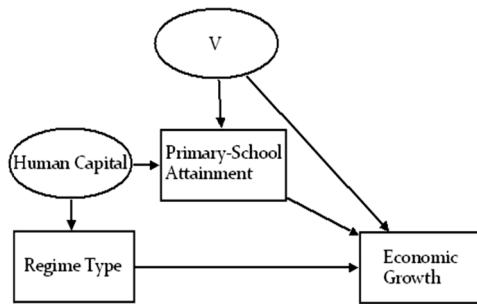


Fig. 13.2c. Schooling as intervening between human capital and growth induces a correlation between democracy and an unmeasured variable

Figure 13.2. Stable Empirical Results Are Compatible with Three Alternative Causal Models

Note: Rectangular boxes represent measured variables; ovals represent latent variables. Data on economic growth and primary-school enrollment are from World Bank (1998); regime data are from Marshall and Jagers (1998). The common regression results are: $Growth = 4.240 + 0.0437 * Regime - 0.000196 * GDP + 0.0135 * Schooling$.

what is sometimes called a “collider” variable.¹⁶ In this scenario, controlling for primary-school attainment induces a correlation between regime type and the unmeasured variable, referred to here as V ,¹⁷ which was previously unrelated to regime type; it is therefore not a threat to causal inference until primary-school attainment is added as a control variable, but it becomes a threat thereafter. Once again, adding the control variable moves the regression finding away from the true causal relationship of interest.

Regression analysis alone cannot distinguish among these three scenarios, even for the simple case of a two-variable relationship with only one control variable. Hence, it cannot by itself tell us whether controlling for primary-school attainment gets us closer to causal inference, moves us further away from successful inference, or has no relevance at all. Thus, when scholars lack strong prior causal knowledge of the relationships involving the hypothesized cause, potential control variables, and the dependent variable, it is difficult if not impossible to determine whether a specific control variable belongs in the regression model that will best capture the true causal structure.

However, even when we have enough causal knowledge to know that a given variable is a confounder and is neither a collider nor a part of the causal mechanism, it generally remains impossible to tell whether controlling for that variable improves the causal inference regarding the variable of primary interest. If the potential control variable is a genuine confounder but is not the *only* confounder, controlling for it can worsen causal inference in at least two ways (see Clarke 2005 for a statistical discussion).

First, the contribution to omitted variable bias of the potential control variable (when it is excluded from the model) may have the opposite sign of the net bias due to the other confounders. If so, then adding the potential control variable will move the causal inference further from the truth because it will remove one of two at least partially countervailing sources of bias. Second, adding the potential control variable may make the causal inference worse by providing a new, and sometimes stronger, statistical pathway—through which the remaining confounders can distort estimates within the regression equation. Neither of these two problems involves par-

16. A collider variable is a potential control variable that is caused both by the main causal variable of interest and also by an unmeasured variable with its own direct causal effect on the outcome. Conditioning on a collider creates a relationship between the main causal variable and the unmeasured variable, even if there was no such relationship beforehand. For two intuitive examples of how colliders cause trouble, see Cole et al. (2010).

17. In this particular example, no specific substantive variable is suggested as V because theory is not yet advanced enough to identify with certainty a variable that is both a cause of primary-school attainment and also completely unrelated to regime type.

ticularly unusual conditions, and in a simple simulation, Clarke (2005: 344-46) shows that the conditions that yield these problematic inferences are indeed roughly as common as the conditions in which adding a control variable improves causal inference.

Thus, the situation is grim for researchers seeking to make causal inferences from observational data by controlling for all relevant confounders. Unless they already know most causal facts about relationships involving the hypothesized cause and the outcome, it will usually be impossible to decide whether a potential control variable should be in the regression that will best capture the true causal structure. Furthermore, even if a given variable clearly should be there, as long as the variable in question is not the last or perhaps only confounder, it remains impossible to tell whether adding it improves or worsens the causal inference. In such circumstances, a causal inference always depends on additional sources of insight beyond multivariate regression-type analysis—for example, from qualitative evidence, experiments, or game theory.

CONTROLLING FOR SOCIOECONOMIC DEVELOPMENT: ALTERNATIVE STRATEGIES

It is thus difficult, at best, to select a set of control variables that would facilitate causal inference regarding the relationship between democracy and economic performance. How have researchers approached the challenge of balancing the inclusion of inappropriate control variables and the exclusion of necessary ones?

Most analysts address this dilemma by including a reasonably broad set of control variables, as can be seen in table 13.3. In this respect, the analysis of Przeworski et al. (2000) appears distinctive, only including controls for capital stock and labor productivity. The selection of these variables was theoretically justified as capturing the basic causes of growth identified in Solow's (1956) classic model. However, in the decades since its publication, analysts have identified several other factors that may affect economic growth, including the above hypotheses related to socioeconomic development. Przeworski et al. present little explanation for why they choose to disregard most of these hypotheses,¹⁸ and there is good reason to believe that including at least some of these other possible control variables might

18. On page 180, they defend the decision to disregard prior income level as a causal variable. However, they partially contradict this argument on page 196 by saying, "It is by now generally accepted that among the developed countries the rate of growth tends to decline in per capita income."

Table 13.3. Alternative Combinations of Control Variables

Variable	Feng (1997)	Casiorowski (2000)	Przeworski et al. (2000)	Barro (1997)	Kurzman et al. (2002)	Durham (1999)	Baum and Lake (2003)
Investment	X	X			X	X	X
Inflation	X	X		X			
Education/ Literacy	X	X		X	X	X	X
Trade	X					X	
GDP	X	X		X	X	X	X
Irregular Regime Change	X						
Major Regular Regime Change	X						
Minor Regular Regime Change	X						
Wage Growth			X				
Money Supply Growth			X				
Violent Unrest			X		X		
Peaceful Unrest			X				
Labor Supply			X				X
Capital Stock			X				
Population Growth				X	X	X	X
Population Size							
Govt. Consumption				X	X	X	
Life Expectancy				X	X		X
Terms of Trade				X			
Rule of Law				X			
Region Dummies						X	

have altered their conclusions about the relationship between democracy and economic growth.

Most other analysts use elaborate sets of control variables. Yet as table 13.3 demonstrates, the control variables are neither consistent nor cumulative across studies. Some analysts include political control variables, such as those reflecting regime changes that do not cross the boundary between democracy and authoritarianism or the occurrence of political protest episodes. Others exclude these variables, preferring to focus more narrowly on factors related directly to socioeconomic development.

How should we evaluate competing sets of control variables? One common answer is to turn to statistical measures of goodness of fit, in particular to the R^2 for each regressions. Researchers could infer that the regression with the highest R^2 best captures the relationship between democracy and growth. However, it is not clear that this approach would lead researchers to accept a regression model that actually captures the true relationship between democracy and economic growth over another that either contains too many or too few control variables (Kennedy 1998: 81–83). This problem is particularly severe in the current research context because democracy and other variables involved in socioeconomic development are known to be highly correlated, a situation that exacerbates the weaknesses of the R^2 as a test for correct specification.

More generally, we have no unambiguous basis for accepting any particular set of control variables as better suited to eliminating confounders. Any of the published models could be close representations of the actual causal relationships underlying economic growth—or, perhaps more likely, none of them capture the true relationships.

Furthermore, the remarkable diversity of approaches to introducing control variables represented in table 13.3 and throughout this literature more generally helps to explain the divergence in findings about democracy and growth. Because changing the set of included control variables can radically alter research findings, analysts' persistent use of non-comparable control variables is one reason why they disagree about the economic effects of democracy. Thus, the strategy of adding control variables to simple regression analyses has not been fruitful for discovering whether or not there is a causal connection between democracy and economic growth.

SCHOLARS' RESPONSES, AND POSSIBLE SOLUTIONS?

Two reactions have dominated responses to the contradictory findings resulting from these problems of specifying control variables. Some scholars take the inconsistency of empirical results as evidence for the claim that

regime type has little or no direct influence on economic growth. For example, Przeworski and Limongi state, "our own hunch is that politics does matter, but 'regimes' do not capture the relevant differences" (Przeworski and Limongi 1993: 65). In effect, this position seems to treat a lack of reliable empirical evidence as a weak form of positive evidence that there is no relationship between democracy and economic growth.

Other analysts note logical gaps in the conclusion that there is no relationship. These researchers claim that various studies have produced diverse and inconsistent conclusions because some or all of them were designed incorrectly. For example, Gasiorowski (2000: 323–25) lists several methodological problems that may afflict studies of the hypothesized relationship between democracy and economic growth, including purely cross-sectional research designs, poor model specification, exogeneity caused by omitted variable bias or reciprocal causation, heteroskedasticity of the error term,¹⁹ and causal heterogeneity. He then states that: all of the studies of which I am aware that examine how democracy affects growth or inflation suffer from one or more of these methodological problems. Although this does not necessarily invalidate these studies, it raises doubts about their findings and suggests that greater care must be taken in designing analyses of democratic performance (Gasiorowski 2000: 324).

This view suggests that establishing the empirical relationship between regime type and economic growth is a matter of asking the right question and answering it with the right evidence and tools. Different studies produce varying findings because *some* studies are theoretically or methodologically flawed. It may be true that there is no relationship between democracy and economic growth, as Przeworski and Limongi believe, or it may be that there is either a positive or a negative relationship between the two. However, advocates of this perspective assert that we *can* know the truth once we establish the appropriate set of set of statistical assumptions.

If better methods and assumptions will indeed resolve these analytic problems, then it is necessary to establish the standard that must be met. In order to learn about the true relationship between regime type and economic growth, scholars will have to employ statistical models that are rigorous, theoretically justified, and that meet the norms that econometricians and methodologists in political science have called the "assumption of correct specification," or the "specification assumption" (Hanushek and Jackson 1977: 79–86; Achen 1982: chap. 5; Leamer 1990; Darnell 1994: 369–73; Kennedy 1998: chap. 5; Greene 2000: 332–38; Collier, Brady, and Seawright, chap. 9, this volume). To meet this assumption, the major causes of the dependent variable must be included, and issues of endogeneity

19. Heteroskedasticity in regression analysis occurs when the variance of the error term is not constant across cases.

must be addressed through an appropriate statistical technique. As this chapter emphasizes throughout, adding the wrong control variables can hurt, rather than help. Thus, inappropriate controls must not be introduced, because they can weaken causal inference, rather than strengthen it. Yet making the right choices here is effectively impossible without certainty about the underlying data-generating process that one wishes to estimate. If, for a particular statistical model, these conditions are met, then the key causal inferences will, on average, be correct. Otherwise, these parameter estimates will suffer bias, generally of unknown direction and magnitude.

Some scholars act as if the above arguments imply that the best approach is to include measures of every conceivable factor in each analysis—for example, using all of the variables in table 13.3 as controls. This kind of “kitchen sink” approach would have the virtue of accurately representing the extreme uncertainty that analysts face in choosing a particular subset of control variables. Moreover, in some circumstances, such an analysis might fortuitously get us closer to an unbiased inference than the more cautious approach adopted by most researchers.

Yet the kitchen sink approach—quite apart from the obvious objection on the general grounds of parsimony—has two problems. First and most crucially, including certain kinds of variables may bias the estimates of the total causal effect. We noted above that adding the wrong controls can hurt rather than help. For example, if some included variables are intervening variables—that is, both consequences of democracy and causes of economic growth—then the portion of democracy’s effect on economic growth that is channeled through these variables will be subtracted from the estimate of democracy’s total effect (Pratt and Schleifer 1984). Subtracting these indirect effects can bias estimates of the total effect in any direction: if an indirect effect is positive, controlling for it will bias the estimate of the total effect in a negative direction, and vice versa. As discussed above, such problems can also arise even with control variables that belong in the model if there is at least one remaining confounder outside the specification.

A second problem with the kitchen sink approach is that issues of multicollinearity will almost certainly arise: different aspects of economic development will probably be correlated with each other to some degree, reducing the precision of estimation for each coefficient. Defending a multivariate specification in the face of these problems would require scholars to provide a detailed defense, based on empirical evidence and prior findings, of every decision and assumption leading to the statistical model that they employ. The specification assumption requires researchers to include *all* of the necessary control variables, so a convincing argument is required for every variable not included in the model. However, the problems of poor or nonexistent measures, multicollinearity, and complexities regard-

ing possible structures of causation mean that there are often serious obstacles to including variables in an analysis, and the researcher must demonstrate convincingly that each variable added to the model does not create more problems than it solves.

Overall, in most research contexts satisfying the specification assumption is impossible, both in principle and in practice. In principle, some variables may be both causes and consequences of the key independent variables,²⁰ and the standard presented above ironically would require that analysts simultaneously *exclude* and *include* such variables. In practice, analysts lack the theoretical knowledge necessary to make such a rigorous defense. Wide debate continues about the problematic implications of this lack of knowledge for successful causal inference (e.g., Leamer 1983; Bartels 1997; Freedman 1999; Clarke 2005).

ALTERNATIVE APPROACHES: REFINEMENTS ON REGRESSION, AND SCALING DOWN

We now consider two alternative approaches to overcoming this problem of failed causal inference: (a) introducing newer refinements on regression analysis; and (b) scaling down the analysis to pursue a more fine-grained examination of causal processes and mechanisms.

Newer Tools for Regression Analysis

The regression-based approach has had minimal success in teasing out the causal influence between regime type on economic growth. This work is bedeviled by many difficulties, prominent among them establishing which variables to include in a multivariate analysis. Some scholars might argue that these problems can be circumvented with newer research designs or statistical techniques, which may yield more adequate causal inferences. While this chapter can only touch briefly on these alternatives, a short discussion of three techniques will give a sense of why these new approaches do not easily resolve the concerns raised above. We comment here on matching designs, regression-discontinuity designs, and instrumental variables. (We have included an entry on each in the Glossary, and they are extensively discussed in chap. 14, this volume.)

20. This claim does not require analysts to accept an idea of simultaneous reciprocal causation, although such an idea is one way to reach a situation in which one variable acts as both a cause and a consequence of another. Other routes to this situation include one variable being a cause of another and also being caused by a third variable which is caused by the second variable (as in X causes Y, Y causes Z, and Z causes X), or sequential causation between one variable and another.

Substantial attention has recently been paid to *matching designs* as a tool for achieving better causal inference in observational studies, based on the introduction of statistical adjustments to control for known confounders—i.e., for variables associated with both the cause and effect.²¹ In the context of quantitative, cross-national research, can matching perform better than conventional regression in finding stable and credible evidence of a causal relationship between regime type and economic performance?

Unfortunately, there is little reason to expect matching to produce more reliable inferences. Fundamentally, both techniques seek to isolate causal effects by statistically controlling for confounders (Morgan and Winship 2007: 87–165), and basically the same difficult choices arise with the selection of control variables. Furthermore, the results are fundamentally similar; matching estimates of the effects of democracy on economic growth based on the specifications and data used above in illustrating challenges for regression produce the same pattern of instability. The two families of techniques do differ in the details of implementation. Matching controls for known confounders by selecting treatment and control cases that are as similar as possible on those variables, whereas regression seeks to achieve this goal by solving systems of simultaneous equations. Although matching may sometimes have advantages over regression in a given context, the core logic is fundamentally similar, and it is reasonable to adopt the same degree of skepticism about causal inferences based on either set of techniques.

Another option is to employ a *regression-discontinuity design* (Thistlethwaite and Campbell 1960) to get closer to causal inference. This design looks for sharp assignment rules that move cases to one group or another on the treatment of interest, based exclusively on whether they are above a well-defined threshold on another variable. That other variable cannot be either a measure of the hypothesized cause of interest or of the outcome. Cases with a score on the assignment variable that falls just below the threshold are compared with those that fall just above, on the assumption that these cases should otherwise be similar.

Can this approach help estimate the causal effect of regime type on economic performance? Probably not, because it seems unlikely that there is a well-institutionalized assignment rule that makes some countries democratic and others authoritarian based on their score on a variable unrelated to regime type or economic development. One example of a potential—but in fact, inappropriate—threshold might be noted: the famous finding that no democracy with a GDP higher than \$6,055 has ever failed (Przeworski et al. 2000: 98). Hence, one might consider comparing democracies that

21. For political science examples, see Imai (2005); Ho et al. (2007); see also discussion in chap. 9, this volume.

have GDPs just over \$6,055 with dictatorships that have GDPs just under the threshold in terms of their economic growth.

While this might be an interesting comparison, it would not be a regression-discontinuity design because, while there may be a jump in the probability of being a democracy at the \$6,055 threshold, there is no institutionalized assignment rule that forces countries to adopt political regimes based on that threshold. Further, the threshold is based on GDP, a variable closely related to the dependent variable in this research program. In sum, this approach does not appear promising for solving the failures of causal inference on which we are focusing.

In addition to matching and regression discontinuity designs, another option is to work with *instrumental variables* that seek to capture exogenous variation in democracy, i.e., variation that could not be related to any unobserved confounding variables or reciprocally influenced by the outcome. The instrument is used to predict democracy, and this predicted version of democracy (which is assumed to now be free from reciprocal causation or problems related to omitted variables) is in turn used as a predictor of economic growth.

For such an analysis to provide good estimates of the causal effect of democracy on economic growth, at least three conditions must be met. First, the instrument must be, to some extent, correlated with democracy; this condition can be evaluated using the data collected for the analysis.

Second, the instrument must be unrelated to any possible confounding variables. That is, something akin to randomization must take place that causes some countries to become democratic and others authoritarian. Such a situation may be hard to imagine, as world politics rarely produces anything like a "regime-type lottery." However, some reasonably analogous event might be possible in the right circumstances.

The third condition appears, for the democracy and growth relationship, to be substantially more problematic. For instrumental variables to produce a helpful estimate of the causal effect, the instrument must have no causal connection with the dependent variable other than via democracy. This condition apparently rules out long-term factors such as geography and colonial history; any major event from long ago will certainly affect the development of a society and its economy in more than one way, and thus cannot serve as an instrument. Yet in the present context, this third condition also rules out short-term instruments. Such instruments would necessarily be associated, for at least some countries, with changes in regime during or shortly before the period under analysis. The event of regime change in itself creates two causal paths by which such instruments can affect economic growth: through the level of democracy, which is the effect of interest, but also through regime transition, an often dramatic series of events that can certainly have its own economic consequences.

In sum, it appears that the relationship between democracy and economic growth offers distinctly unfriendly terrain for causal inference via instrumental variables.

Nonetheless, in contrast with matching or regression-discontinuity designs, this approach has seemed sufficiently promising to a number of researchers to yield a great deal of published work. For example, Przeworski et al. present “selection-corrected” results that depend on instrumental variables and other related tools (2000: 148, 150, 152, 157). However, this is only as useful as the instrument that Przeworski et al. exploit to find exogenous variation in democracy. Unfortunately, evaluating the proposed instrument they employ is not possible because it remains unidentified.²²

In another attempt to find exogenous variation in the independent variables, Barro (1997: 14) uses lagged values of the independent variables as instruments to try to find variation in the current values of the these variables that is exogenous from current economic processes. While this approach is promising in many circumstances, it is at least somewhat problematic in the current research context. After all, the processes of economic development—which, broadly speaking, are the most important hypothesized confounding variables in this relationship—tend to occur over the long term. Hence, simply lagging the indicators of development for one period may not be effective in identifying exogenous variation in democracy.

Feng’s (1997) attempt to find relevant exogenous variation is more substantively innovative. Positing that cultural differences across countries produce different levels of democracy that are not tightly connected to socioeconomic development, Feng uses dummy variables for countries where Islamic or Confucian culture has been influential as a predictor of exogenous variation in level of democracy (404–07). However, these particular variables may well be problematic because in the long-run, they may be causes of both socioeconomic development and democracy (Huntington 1991, 1996).

Feng’s effort builds on the supposition that cultural and leadership-based hypotheses about democracy may contribute useful instruments. Yet some

22. Rather than identifying the instrument in use, the authors report that “the factors that enter on the right-hand side of the performance equation are not statistically significant in the selection equation. Hence, throughout the book, we treat selection as exogenous” (Przeworski et al. 2000: 285). If selection were in fact exogenous, then correcting for it would have no effect, because all of the variation on the independent variables would already be exogenous (Heckman 1988: 7). However, the selection-corrected results that Przeworski et al. present are always at least subtly different from their standard regression results, so selection—or, more generally, endogeneity—is not necessarily unproblematic.

elements of leadership and culture may even worsen the problem of distinguishing between the effects of democracy and socioeconomic development. Development may, for example, increase the probability that a country will have educated leaders—which may, in turn, make the leaders more likely to behave democratically and to facilitate good economic outcomes.²³

Thus, variables from the domain of concern to Feng, and from other domains on which scholars have focused, do not yet offer a solid solution to the problem of finding exogenous variation in democracy. Unless analysts actually find such variation, instrumental variables will continue to offer no improvement over simple regression analysis in identifying the true causal relationships between democracy and economic growth. Furthermore, because the requirement that the instrument affect the economy only through democracy appears to rule out both long-term and short-term possible instruments, we may expect the search for satisfactory instruments will continue to be fruitless.

These arguments further build the case that, in the framework of large-scale cross-national comparisons, successful causal inference about the relationship between democracy and economic growth is difficult and—at least for the near future—unlikely. These techniques for causal inference which have recently generated a great deal of excitement—matching, regression-discontinuity designs, and instrumental variables—appear to add little in this context. How, then, might this literature proceed productively? Certainly, it should not simply be discontinued; the research question is, after all, of real normative significance.

Scaling Down to Mechanisms and Qualitative Evidence

The approach of quantitative, cross-national analysis has proved unsuccessful in evaluating the relationship between democracy and growth—just as these broad quantitative comparisons failed in many other substantive domains. Cross-national regression studies, as well as some key refinements on regression analysis, face serious obstacles, given the current state of knowledge. It appears that the causal questions of interest here cannot at present be resolved at this level of analysis.

If progress is to be made, it may be helpful for scholars to abandon these

23. Obviously, this hypothetical example adopts an optimistic view of the effects of education in the fields of economics and political science. It may well be the case that such an education enables leaders to be more effective at repression and corruption, in which case it may *inhibit* democracy and economic growth. The statistical point is equally valid in either case. For a slightly different discussion of possible inferential problems related to leadership, see Przeworski et al. (2000: 286).

sweeping comparisons and scale down their inferential goals. One option is to scale down analytically, following Goldthorpe's (2001) recommendation to test "generative," or causal, mechanisms. This avenue can be pursued through both quantitative and qualitative analysis. Indeed, Hedstrom (2008) emphasizes that mechanisms cannot be studied with statistical models, and Freedman (chap. 11, this volume) demonstrates the critical contribution of qualitative evidence in teasing out mechanisms.

Scaling down to mechanisms can draw on the theoretical literature on democracy and growth. This literature—which, after all, empirical research on democracy and growth is designed to test—offers plausible hypotheses about *why* and *how* democracy would cause growth, obviously involving issues of mechanisms. For example, Olson (1983) hypothesizes that democracy is harmful for growth because it encourages more interest groups to seek inefficient rents from the state, reducing the efficiency of the economy as a whole. On the other hand, Olson (1990) also proposes that democracy may be good for growth because it encourages national leaders to consider the economic well-being of a wider array of citizens and, consequently, to choose more efficient policies and tax rates.

A plausible causal mechanism for the hypothesis that democracy does not affect economic growth might rest on the assumption that all political leaders, whether democratic or authoritarian, are fairly likely to lose office during economic crises. If this is true, then all leaders will be motivated to pursue the best possible economic policy, and regime type may make no difference in economic growth. A wide range of other plausible mechanisms can of course be posited as relevant to any conceivable version of the relationship between democracy and economic growth.

The point is that each of these hypotheses has important empirical implications. These linkages consist of testable claims about relationships *other than* the overall relationship between democracy and growth. For example, with reference to Olson's theories discussed above, is it true that more interest groups seek rents under democracies than under other regimes? Is it true that an increasing number of rent-seeking interest groups in turn increases the overall volume of rents dispensed by the state? Is it true that a rise in the number of rent-seeking interest groups causes a decline in overall economic efficiency? Is it true that dictators care less about pursuing good economic policy than do democratic leaders?

Each of these questions is amenable to empirical exploration and any of them answered in the affirmative—based on specifications that prove to be robust to changes in statistical assumptions. Indeed, some recent work has begun to move in this direction, such as Baum and Lake's (2003) valuable analysis of the effects of democracy on economic growth through increased life expectancy and secondary education.

Unfortunately, this research faces two major inferential challenges of its

own. First, like work on the direct relationship between democracy and economic growth, it is based on observational studies. The results rely on untested assumptions, and at least some of these assumptions probably influence the conclusions—potentially in ways that are hard to detect. We may hope that inferences about causal mechanisms will be more robust than the broad comparative inferences about democracy and growth; nevertheless, even these more delimited inferences will require a certain suspension of disbelief. One strategy here—which may escape from some of these assumptions but certainly requires others—is to employ process-tracing tools (chaps. 10, 11, and 12, this volume).

A second difficulty is that neither quantitative nor qualitative confirmations of certain linkage mechanisms would aggregate in any straightforward way into an overall answer—at the broad level of generality that was the focus of the large-N quantitative cross-national literature—to the question of whether democracy causes growth. In fact, as Przeworski and Limongi (1993: 60) suggest, any number of arguments about linkages may be *simultaneously* true. Thus, while confirming the particular linkages entailed by a theory provides important evidence in favor of one aspect of that theory's validity, it does not necessarily confirm the theory's overall conclusions.

CONCLUSION

Research in many different substantive areas is likely to face the same problems of causal inference discussed in this chapter. Given the importance of understanding economic growth and its relationship with democracy, I have used this body of research to illustrate the wider shortcomings of quantitative, cross-national methods. Divergent conceptualizations of relevant control variables, inescapable problems of confounding and reciprocal causation, lack of equivalence in the scope of comparison, and problems with data have contributed to remarkably inconsistent findings. Different studies have concluded that democracy has a positive linear relationship, a negative linear relationship, no relationship, a curvilinear relationship, and other kinds of nonlinear relationships with economic growth. These remarkably divergent conclusions reflect deep theoretical and methodological problems in these studies.

Scholarly responses have ranged from claims that the inconsistent findings are evidence of no relationship, to discussions of problems in choosing control variables, to attempts at enhancing and moving beyond conventional regression modeling. Unfortunately, optimism about the leverage that can derive from the newer quantitative techniques discussed above appears to have been exaggerated, and in the substantive domain of cross-national analysis that has been our focus, it is hard to find opportunities

for applying tools such as regression-discontinuity designs. These refinements on regression have limited analytical power in this area of research.

In the introduction, we emphasized that this failure of causal inference extends well beyond the substantive domain under discussion here—arising in numerous other areas of quantitative cross-national analysis, spanning several disciplines. This outcome is poignantly lamented by two economists, Lindauer and Pritchett (2002: 18), who in recognition of this failure go so far as to call for an “obituary for growth regressions.”

How should analysts proceed in these challenging substantive domains? We have discussed the option of scaling down to generative or causal mechanisms. This focus does not sustain the level of generality that quantitative, cross-national analysis presumes to achieve—but as we have seen this is a false generality, given that the results are so unreliable.

In conjunction with scaling down, scholars should look systematically for the kinds of qualitative, historical, and case-based evidence that other chapters in this volume recommend as a supplement to—and sometimes a substitute for—quantitative analysis. These approaches offer the possibility of making progress toward establishing whether democracy has any causal effect on growth.

14

Design-Based Inference: Beyond the Pitfalls of Regression Analysis?

Thad Dunning

A perceptible shift of emphasis appears to be taking place in the study of quantitative political methodology. In recent decades, much research on empirical quantitative methods has been quite technical, focused—for example—on the mathematical nuances of estimating complicated linear and non-linear regression models.¹ Reviewing this trend, Achen (2002) notes that “steady gains in theoretical sophistication have combined with explosive increases in computing power to produce a profusion of new estimators for applied political researchers.”

Behind the growth of such methods lies the belief that estimation of these complex models allows for more valid causal inferences, perhaps compensating for less-than-ideal research designs. Indeed, one rationale for multiple regression and its extensions is that it allows for comparisons that approximate a true experiment. The pervasiveness of this idea is reflected in a standard introductory econometrics text: “the power of multiple regres-

I am grateful to Taylor Boas, Christopher Chambers-Ju, David Collier, William Hennessey, Daniel Hidalgo, Simeon Nichter, and Neal Richardson for helpful comments and suggestions.

1. Regression analysis involves “statistical models,” a key concept defined in the Glossary. A statistical model is a probability model that stipulates how data are generated. In regression analysis, the statistical model involves choices about which variables are to be included, along with assumptions about functional form, the distribution of (unobserved) error terms, and the relationship between error terms and observed variables.

sion analysis is that it allows us to do in non-experimental environments what natural scientists are able to do in a controlled laboratory setting: keep other factors fixed" (Wooldridge 2009: 77).

Yet this focus on complex statistical models and advanced techniques for estimating those models appears to be giving way to greater concern with more foundational issues of research design. Growing recognition of the frequently severe problems with regression-based inference, explored by Seawright (chap. 13, this volume), has intensified this trend. Leading methodologists have underscored the pitfalls of these techniques—including more technically-advanced models and estimators—which fall under the rubric of what Brady, Collier, and Seawright (chap. 1, this volume) call mainstream quantitative methods. Achen (2002), a prominent skeptic, proposes "A Rule of Three" (ART), arguing that multiple regression models should be limited to no more than three well-understood, well-theorized, and well-measured independent variables. This approach is a far cry from more conventional practice in quantitative research, in which the trend has been towards more complex statistical models in which the assumptions are difficult to explicate and defend—let alone validate. Trenchant critiques of the failures of applied regression modeling by statisticians such as David Freedman (1991, 1999, 2009) have likewise commanded growing attention.²

Of course, seminars on research design have long been a bedrock of graduate training in many graduate programs, and the importance of good design for causal inference has been emphasized by leading texts, such as King, Keohane, and Verba (1994; see also chap. 7, this volume). What distinguishes the current emphasis is the conviction that if research designs are flawed, statistical adjustment can do little to bolster causal inference. As Sekhon (2009: 487) puts it, "without an experiment, natural experiment, a discontinuity, or some other strong design, no amount of econometric or statistical modeling can make the move from correlation to causation persuasive."

Consequently, scholars have sharply increased their use of field and laboratory experiments (Druckman et al. 2006; Gerber and Green 2008; Morton and Williams 2008)—as well as observational studies such as natural experiments, which approximate the logic of true experiments (Dunning 2008a). At recent meetings of the Political Methodology Society, growing numbers of panels and papers have been devoted to questions of research design, while papers in the Society's journal, *Political Analysis*, show an increasing concern with this topic. Several working groups focused on this

2. After David Freedman's death in 2008, panels were held at the meetings of APSA (Toronto, Canada 2009) and the Society for Political Methodology (Yale, 2009) to discuss his influence on the social sciences.

methodology have also emerged in the discipline.³ While this shift in attention is perhaps not yet dramatic, it is both perceptible and growing.

This emphasis on research design points to the guiding question of this chapter: How far does strong research design take us beyond the pitfalls of conventional regression modeling? This focus in turn raises several other questions. To what extent can research design help us to make causal inferences? What are the strengths and limitations of different kinds of designs, including but not limited to field and natural experiments? What is the role of different conceptions of causation and alternative statistical models? Finally, what leverage do other modes of inference—for example, those involving qualitative methods—provide in discovering opportunities to construct such research designs and in complementing and bolstering their power?

This chapter explores these questions first by discussing the contrast between “design-based” and “model-based” inference. Of course, design-based inference routinely relies on statistical models, and model-based approaches routinely entail some sort of research design. In principle, then, a crucial difference concerns not the *presence* of statistical models, but rather their simplicity, transparency, and credibility.

In practice, unfortunately, this difference is not always apparent. While stronger research designs should permit data analysis with weaker assumptions, the conceptions of causation and statistical methods widely employed in what might appear to be design-based research are often virtually indistinguishable from more conventional model-based approaches. To realize more fully the potential of design-based methods, strong research designs should be analyzed as if they were true experiments—thereby allowing the use of the simpler statistical tools appropriate to them. Along with more complex statistical analyses that might be developed, researchers should employ simple tests—such as comparison of the mean scores of cases that fall in the different categories of the independent variable (that is to say, differences of mean outcomes across treatment and control groups). Calculations of standard errors should follow the best practices for true experiments, rather than resting on the assumptions behind standard regression models.

To explore answers to other questions about the strengths and limitations of design-based inference, I develop a typology based on three dimensions for evaluating research designs: (1) plausibility of *as-if* random assignment to the categories of the key independent variable; (2) credibility of the sta-

3. Examples include, *inter alia*, the Experiments in Governance and Politics (EGAP) network, an annual conference at the Center for Experimental Social Science at NYU, and multiple conferences and workshops organized at the Institution for Social and Policy Studies at Yale.

tistical model;⁴ and (3) wider substantive relevance of the principal explanatory variable. The three dimensions and the trade-offs among them are discussed against the backdrop of recognizing the critical importance of substantive, case-based knowledge in constructing and executing these research designs.

Each dimension corresponds to distinctive challenges that arise in drawing causal inferences about the social and political world, including problems of: (i) confounding; (ii) specifying the causal and/or stochastic process by which observable data are generated; and (iii) generalizing the effects of particular treatments or interventions to a wider set of social or political processes of analytic concern, and/or to populations other than that being studied.

To explore the importance of these dimensions, I locate several leading studies within the three-dimensional space established by this typology. I focus on research that claims to utilize natural experiments—both because such designs have increasingly been employed in political science⁵ (for reviews, see Gerber and Green 2008; Dunning 2008a) and because different natural experiments prove to be located in different positions within the cube generated by the typology. Along with its value in assessing natural experiments, the typology is likewise useful for situating any kind of research design, including experiments and conventional observational studies.

A final introductory point must be underscored. Many technical issues lie behind the ideas presented here—for example, the relation of these arguments to the Neyman-Rubin-Holland model of causal inference. To help ensure that the text is accessible to a wide range of readers, these arguments are presented in footnotes.

4. As with all “dimensions” that scholars construct, these three criteria for evaluation can obviously be disaggregated. Closely connected with the idea of the statistical model, one may also consider the “conception of causation” employed. For example, the Neyman model (1923), which is central to discussions of natural experiments, is based on a manipulationist and counterfactual conception of causality. Further, in Neyman’s approach (also known as the Neyman-Rubin-Holland model) we find a set of assumptions about causal process—for example, that one unit’s outcome when assigned to treatment or control is deterministic and does not depend on whether another unit is assigned to treatment or control. In the text below, the discussion of the second dimension of the typology—the credibility of the statistical model—will occasionally make reference to this related question of alternative conceptions of causation.

5. For reviews, see Gerber and Green (2008) and Dunning (2008a). Diamond and Robinson’s (2010) edited volume, *Natural Experiments in History*, includes studies across several disciplines and encompasses a much wider range of designs, including comparative case studies.

DESIGN-BASED AND MODEL-BASED INFERENCE

The distinction between design-based and model-based inference is central here (Dunning 2008b, Sekhon 2009). In one form of design-based inference, the dataset is generated through true experimental intervention that is planned and executed by the researcher. My concern, by contrast, is with a related research design. Here, the investigator searches for natural variation in social and political processes that produces certain forms of *as-if* random assignment that mimic a true experiment—hence the idea of a *natural experiment*. The goal is to mitigate standard concerns about confounding and omitted variable bias. Confounding factors—those associated with both a putative cause and a putative effect—typically bedevil causal inference in the social sciences. The objective here is to eliminate or mitigate confounders by taking advantage of “nature’s” *as-if* random assignment, using *a priori* reasoning and diverse forms of evidence to validate the claim that exposure to the putative cause is as good as random. Then, statistical adjustments for confounders—based either on control variables in a multivariate regression or analogous methods such as matching—may be unnecessary.⁶ Ideally, the researcher can make valid causal inferences by analyzing the simple mean or percentage difference between the treatment and control groups.⁷

In design-based inference involving natural experiments, this optimal situation may not be achieved. Good research design requires integrating and coordinating among the dimensions discussed above, and enhancing this integration on the basis of what may be seen as a fourth dimension or resource.

1. *As-if Random*. In designing a natural experiment, the researcher seeks instances of *as-if* random assignment of units (cases) to values of the key independent variable. One typically cannot prove that the allocation of units into “treatment” or “control” groups is truly random. Yet this assertion should be validated to the extent possible, through quantitative and qualitative evidence and through informed reasoning about the substantive domain under study.

2. *Judgments about the Data Analysis*. The investigator must make careful judgments about the degree to which assignment is indeed *as-if* random.

6. The strengths and limitations of various rationales for estimating regression models on natural experimental data, such as reducing the variance of treatment effect estimators, are discussed below.

7. The Neyman-Rubin-Holland model for causal inference provides the theoretical underpinnings for such simple comparisons, as discussed further in the next sections.

When a compelling case can be made, simple forms of data analysis are suitable—for example, the straightforward comparison of means or percentages just noted. If the *as-if* random character of assignment is not convincing, more elaborate statistical modeling may be necessary to correct for problems in the process of assignment. When *as-if* random assignment falls short, causal inferences may be even more vulnerable without such modeling. Yet, if models are used to adjust the data, the opportunity to sidestep many problems and assumptions associated with complex modeling may be lost.

3. *Wider Substantive Relevance.* In the search for situations of apparent *as-if* random assignment, the analyst must also be concerned with whether the explanatory variable thereby generated is in fact interesting and relevant to a wider set of substantive concerns. Clever studies in which this form of assignment is compelling, but have only limited substantive relevance, do not meet a high standard of research design.

4. *Subject-Matter Knowledge.* Judgments about coordinating among the first three dimensions should rely on deep knowledge of the subject matter and the context of research. It is an illusion to believe that mere technique is sufficient to design good natural experiments, just as it is an insufficient basis for regression analysis. Without a foundation of substantive expertise, a study will routinely make mistakes on the other three dimensions (see Freedman 2010, *passim*).

In sum, building strong research designs with natural experiments requires choices about these multiple objectives—compelling *as-if* random assignment, simplicity of data analysis, and wider relevance. These objectives may be in conflict, and strong research can be understood as the process of balancing astutely among them. Substantive expertise plays a vital role in striking the appropriate balance.

This design-based approach is contrasted with model-based inference, which relies on the statistical models that underlie different variants of regression analysis. Here, statistical adjustment for potential confounders is used to produce—always by assumption—the independence of treatment assignment and omitted (unobserved) causes of the outcomes being explained.⁸ Of course, conditional independence is difficult to achieve (Brady, chap. 3, this volume). The relevant confounding variables must be identified and measured, and the data must be analyzed within the strata defined by these variables. Without *as-if* random assignment, unobserved or unmeasured confounders may threaten valid causal inference.

Another problem with model-based approaches is that inferring causation from regression may require a theory of how the data are generated—

8. The meaning of independence and conditional independence is discussed below, and also in the Glossary.

i.e., a response schedule (Freedman 2009: 85–95, Heckman 2000). This theory is a hypothetical account of how one variable would respond if the scholar intervened and manipulated other variables. In observational studies, of course, the researcher never actually intervenes to change any variables, so this theory remains, to reiterate, hypothetical. Yet data produced by social and political processes can be used to estimate the expected magnitude of a change in one variable that would arise if one were to manipulate other variables—assuming, of course, that the researcher has a correct theory of the data-generating process. The problem is that these theories linking alternative values of the independent variable to the dependent variable sometimes lack credibility as descriptions of the true data-generating process.

Overall, as a heuristic distinction, the contrast between design-based and model-based inference is valuable, yet for several reasons this contrast is not absolute. First, strong research designs—including true experiments and natural experiments—also require statistical models. Before a causal hypothesis can be formulated and tested, a causal model must be defined, and the link from observable variables to the parameters of that model must be posited.⁹ Statistical tests, meanwhile, depend on the stochastic process that generates the data, and this process must also be formulated as a statistical model. The presence of a strong research design does not obviate the need to formulate a model of the data-generating process.

By the same token, model-based empirical inference requires some sort of research design. Indeed, questions about modeling assumptions and data-analytic techniques are analytically distinct from questions about design, as seen in recent debates about the conditions under which multiple regression models should be used to analyze experimental data (Freedman 2008a,b; Green 2009).

At least in theory, one major difference between design-based and model-based inference lies in the *types* of statistical models that undergird the analysis. However, in perusing the leading political science and economics journals, it is sometimes difficult to see a consistent difference. To be sure, empirical researchers increasingly have sought to use true experiments and natural experiments. In principle, such designs are often amenable to simple and transparent data analysis, grounded in credible hypotheses about the data-generating process.

In practice, large, complex regression models are often fitted to the data produced by these strong research designs. Researchers may have various objectives, some quite valid, in pursuing such analytic strategies. Yet these strategies can impose costs (often unacknowledged), both in terms of the

9. This is typically true even of so-called “non-parametric” models, in which (despite the name) there are typically parameters to be estimated from the data.

credibility of the underlying statistical models and the simplicity and transparency of the associated empirical techniques. The crux of the matter is suggested by this question: Why control for confounders if the research design ensures that confounders are statistically independent of treatment? Indeed, if assignment is truly *as-if* random, a simple comparison of average outcomes in treatment and control groups provides valid causal inference.¹⁰ Whether this objective is achieved will be a key criterion for evaluating the credibility of the research design.

NATURAL EXPERIMENTS

This section introduces what will be called “standard” natural experiments, followed by a discussion of two research designs that in effect build on this approach: regression-discontinuity designs and instrumental-variables designs. Finally, the contrast with matching designs is discussed.

Standard Natural Experiments

The importance of natural experiments lies in their contribution to addressing confounding, a pervasive problem in the social sciences. For instance, consider the obstacles to addressing the following assertion: College graduates earn more than individuals who do not go beyond high school. If this statement is interpreted causally, confounding may be a problem, in that the difference in income could in part be directly due to factors—such as intelligence and family background—that probably also make it more likely that people graduate from college.

Investigators may adjust for potential confounders in observational (non-experimental) data, for instance, by comparing college and high school graduates within strata defined by family backgrounds or measured levels of intelligence. At the core of mainstream qualitative methods (chap. 1, this volume) is the hope that such confounders can be identified, measured, and controlled. Yet it is not easy to control for them. Moreover, even within the strata defined by family background and intelligence, there may be other confounders (say, determination) that are associated with getting a college education and that also help to determine wages.

Randomization is one way to eliminate confounding (Fisher 1935; Duflo and Kremer 2006). In a randomized controlled experiment to estimate the returns to education, subjects could be randomly assigned to go to college (the treatment) or straight to work after high school (the control). Intelligence, family background, determination, and other possible confounders would be balanced across these two groups, up to random error, so post-

10. That is, a difference-of-means test validly estimates the average causal effect of treatment assignment.

intervention differences would be evidence for a causal effect of college education.¹¹ Of course, experimental research in such contexts would be expensive and impractical, as well as unethical.

Scholars therefore increasingly employ natural experiments—attempting to identify and analyze real world situations in which some process of *as-if* random assignment places cases in alternative categories of the key independent variable (Gerber and Green 2008, Sekhon 2009; Dunning 2008a). Because the *as-if* random assignment occurs as a feature of social and political processes, the researcher faces a major challenge in identifying situations in which this occurs. Hence, one often speaks not of “creating” a natural experiment, but of “exploiting” an opportunity for this kind of design in the analysis of observational data.

Recent studies have used this approach to study the relationship between income and political attitudes (Doherty, Green, and Gerber 2006), the effect of voting costs on turnout (Brady and McNulty 2004), the impact of electoral competition on ethnic identification (Posner 2004), and many other topics. Table 14.1 presents a non-exhaustive list of political science studies claiming to use this design-based approach to causal inference.¹²

Natural experiments share one crucial attribute with true experiments and partially share a second attribute (Freedman, Pisani, and Purves 2007: 3–8). First, outcomes are compared across subjects exposed to a treatment and those exposed to a control condition (or a different treatment), involving an independent variable that is often (though not always) a dichotomy. Second, in partial contrast with true experiments, subjects are usually assigned to the treatment not at random, but rather *as-if* at random.¹³ Given that the data come from naturally occurring phenomena that often entail social and political processes, the manipulation of the treatment is not under the control of the analyst; thus, the study is observational. However, a researcher carrying out this type of study can make a credible claim that the assignment of non-experimental subjects to treatment and control conditions is *as-if* random.¹⁴

11. The role of random error gets smaller as the treatment and control groups get larger; the point of statistical hypothesis testing is to distinguish chance variation from true treatment effects.

12. Table 14.1 includes the work of major scholars in this tradition, a great many of whom do outstanding research. The list is not intended to reflect a full spectrum of strong and weak natural experiments.

13. In some natural experiments, such as lottery studies (e.g. Doherty, Green, and Gerber 2006), a true randomizing device assigns units to treatments.

14. It is useful to distinguish natural experiments from the “quasi-experiments” discussed by Donald Campbell and colleagues (1963, 1968), in which *non-random* assignment to treatment is a key feature (see Achen 1986: 4). In the famous “interrupted time-series” discussed by Campbell and Ross (1968), Connecticut’s speeding law was passed after a year of unusually high traffic fatalities. Some of the subsequent reduction in traffic fatalities was due to regression to the mean, rather than to the effect of the law (Campbell and Stanley 1963).

Table 14.1. Examples of Natural Experiments, Including Regression Discontinuity (RD) and Instrumental Variable (IV) Designs^a

<i>Authors</i>	<i>Substantive focus</i>	<i>Source of alleged natural experiment</i>	<i>RD, IV, or standard natural experiment</i>	<i>Simple difference-of-means test</i>
Angrist and Lavy (1999)	Effect of class size on educational achievement	Discontinuities introduced by enrollment ceilings on class sizes	RD	No
Ansolabehere, Snyder, and Stewart (2000)	The personal vote and incumbency advantage	Electoral redistricting	Standard	Yes
Banerjee and Iyer (2005)	Effect of landlord power on development	Land tenure patterns instituted by British in colonial India	Standard and IV	No
Berger (2009)	Long-term effects of colonial taxation institutions	The division of northern and southern Nigeria at 7°10' N	Standard	No
Blattman (2008)	Consequences of child soldiering for political participation	<i>As-if</i> random abduction of children by the Lord's Resistance Army	Standard	No
Brady and McNulty (2004)	Voter turnout	Precinct consolidation in California gubernatorial recall election	Standard	Yes
Card and Krueger (1994)	The effects of minimum-wage laws on unemployment	Differential exposure to minimum-wage laws among fast-food restaurants on the New Jersey-Pennsylvania border	Standard (Difference-in-Differences)	Yes
Chattopadhyay and Dufló (2004)	Effects of electoral quotas for women in Rajasthan and West Bengal	Random assignment of quotas for village council presidencies	Standard	Yes
Cox, Rosenbluth, and Thies (2000)	Incentives of Japanese politicians to join factions	Cross-sectional and temporal variation in institutional rules in Japanese parliamentary houses	Standard	Yes
Doherty, Green, and Gerber (2006)	Effect of income on political attitudes	Random assignment of lottery winnings, among lottery players	Standard	No ^b
Dunning (2009)	Effects of caste-based quotas on ethnic identification and distributive politics	Regression-discontinuity based on rule rotating quotas across village councils in Karnataka	RD	Yes
Ferraz and Finan (2008)	Effect of corruption audits on electoral accountability	Release of randomized corruption audits in Brazil	Standard	Yes (with state fixed effects)
Galiani and Schargrodsky (2004); also Di Tella et al. (2007)	Effects of land titling for the poor on economic activity and attitudes	Judicial challenges to transfer of property titles to squatters	Standard	Yes (2004) No (2007)
Glazer and Robbins (1985)	Congressional responsiveness to constituencies	Electoral redistricting	Standard	No

Grofman, Brunell, and Koetzle (1998)	Midterm losses in the House and Senate	Party control of White House in previous elections	Standard	No
Grofman, Griffin, and Berry (1995)	Congressional responsiveness to constituencies	House members who move to the Senate	Standard	Yes
Hidalgo, Naidu, Nichter, and Richardson (Forthcoming)	Effects of economic conditions on land invasions in Brazil	Shocks to economic conditions due to rainfall patterns	IV	No ^b
Ho and Imai (2008)	Effect of ballot position on electoral outcomes	Randomized ballot order under alphabet lottery in California	Standard	Yes
Hyde (2007)	The effects of international election monitoring on electoral fraud	<i>As-if</i> random assignment of election monitors to polling stations in Armenia	Standard	Yes
Krasno and Green (2008)	Effect of televised presidential campaign ads on voter turnout	Geographic spillover of campaign ads in states with competitive elections to some but not all areas of neighboring states	Standard and RD	No ^b
Lee (2008)	The causal effect of incumbency on electoral advantage	Comparisons of near-winners and near-losers in U.S. congressional elections	RD	No
Lerman (2008)	Social and political effects of incarceration in high-security prison	Regression-discontinuity based on index used to assign prisoners to prisons in California	RD and IV	Yes
Lyall (2009)	Deterrent effect of bombings and swelling in Chechnya	<i>As-if</i> random allocation of bombs by drunk Russian soldiers	Standard	No ^c
Miguel (2004)	Nation building and public goods provision	Political border between Kenya and Tanzania	Standard	No
Miguel, Satyanath and Sergenti (2004)	Economic growth and civil conflict	Shocks to economic performance caused by rainfall	IV	No
Posner (2004)	Political salience of cultural cleavages	Political border between Zambia and Malawi	Standard	Yes
Snow on cholera (Freedman 1991, 2010)	Incidence of cholera in London	<i>As-if</i> random allocation of water to different houses	Standard	Yes ^d
Stasavage (2003)	Bureaucratic delegation, transparency, and accountability	Variation in central banking institutions	Standard	No ^b
Titunuk (2008)	Effects of term lengths on legislative behavior	Random assignment of U.S. state senate seats to two or four year terms after reapportionment	Standard	Yes

^a This non-exhaustive list includes published and unpublished studies in political science and cognate disciplines that either lay explicit claim to having exploited a “natural experiment” or adopt core elements of the approach.

^b The treatment conditions and/or instrumental variables are continuous in these studies, making calculation of differences-of-means less straightforward.

^c Matching—a form of control for observed confounders—was done prior to calculation of mean differences between treatment and control groups.

^d In Snow’s study, the highly transparent data analysis focused on differences in incidence of cholera among three types of households.

A classic, paradigmatic example of a natural experiment, introduced in discussions of social science methodology by Freedman (1991, chap. 11, this volume), comes from the health sciences. Here, the mid-19th century epidemiologist Snow (Snow 1936 [1855]) tests the hypothesis that cholera is waterborne. In addition to building on diverse forms of qualitative evidence, he employs a natural experiment to compare households that received water from two different companies. There were strong reasons to believe that the allocation of water had occurred *as-if* at random. Distribution from the two companies had not followed a systematic plan; adjoining households did not necessarily receive water from the same company; and there was every reason to think that the choice of a given household to reside in a particular dwelling was independent of any information about the corresponding water company. Just prior to a major cholera epidemic, one of the companies had moved its intake pipe away from an obviously contaminated water source, a change that could not have been anticipated by different households—thus sustaining the pattern of *as-if* random assignment to the water source.

To support his causal inference about the cause of cholera, Snow compares the incidence of cholera per 10,000 houses among those supplied by the suspect company, those supplied by the other company, and the rest of London. The data analysis is thus remarkably simple and transparent, *as-if* random assignment yields a strong likelihood that confounders are eliminated, and the study provides highly credible evidence that cholera is a waterborne disease.

An excellent social science example of a natural experiment is Galiani and Schargrodsky's (2004) study of how property rights and land titles influence the socio-economic development of poor communities. In 1981, urban squatters organized by the Catholic Church in Argentina occupied open land in the province of Buenos Aires, dividing the land into parcels that were allocated to individual families. A 1984 law, adopted after the return to democracy in 1983, expropriated this land with the intention of transferring titles to the squatters. However, some of the original landowners challenged the expropriation in court, leading to long delays in the transfer of titles to some of the squatters. By contrast, for other squatters, titles were granted immediately.

The legal action therefore created a (treatment) group of squatters to whom titles were granted promptly and a (control) group to whom titles were not granted. The authors find subsequent differences across the two groups in standard social development indicators: average housing investment, household structure, and educational attainment of children.¹⁵ They

15. On the other hand, they do not find a difference in access to credit markets, which contradicts De Soto's (1989, 2000) theory that the poor will use titled property to collateralize debt.

also find a positive effect of property rights on self-perceptions of individual efficacy. For instance, squatters who were granted land titles—for reasons over which they apparently had no control—disproportionately agreed with statements that people get ahead in life due to hard work (Di Tella, Galiani, and Schargrodsky 2007).

Is this a valid natural experiment? The key claim is that land titles were assigned to the squatters *as-if* at random, and the authors present various kinds of evidence to support this assertion. In 1981, for example, the eventual expropriation of land by the state and the transfer of titles to squatters could not have been predicted. Moreover, there would have been little basis for successful prediction by squatters or the Catholic Church organizers of which *particular* parcels would eventually have their titles transferred in 1984. Titled and untitled parcels sat side-by-side in the occupied area, and the parcels had similar characteristics, such as distance from polluted creeks. The authors also show that the squatters' characteristics such as age and sex were statistically unrelated to whether they received titles, as should be the case if titles were assigned at random. Finally, the government offered equivalent compensation—based on the size of the lot—to the original owners in both groups, suggesting that the value of the parcels does not explain which owners challenged expropriation and which did not. On the basis of extensive interviews and other qualitative fieldwork, the authors argue convincingly that idiosyncratic factors explain some owners' decisions to challenge expropriation, and that these factors were unrelated to the characteristics of squatters or their parcels.

Galiani and Schargrodsky thus present strong evidence for the equivalence of treated and untreated units. Along with qualitative evidence on the process by which the squatting took place, this evidence helps bolster the assertion that assignment is *as-if* random. Of course, assignment was not randomized, so the possibility of unobserved confounders cannot be entirely ruled out. Yet the argument for independence of assignment to treatment vis-à-vis the potential outcomes for the squatters appears compelling.¹⁶ Here, the natural experiment plays a crucial role. Without it, the intriguing findings about the self-reinforcing (not to mention self-deluding) beliefs of the squatters could have been explained as a result of unobserved characteristics of those squatters who did or did not successfully gain titles. It is the research design that makes the evidence for a causal effect of

16. Potential outcomes are those that would be observed if a subject were assigned to receive treatment (a land title) or assigned to the control group. These potential outcomes cannot simultaneously be observed for a single subject. The independence of treatment assignment and potential outcomes means that subjects with particularly high (or low) potential outcomes under the treatment condition are as likely to be assigned to treatment as to control.

land titling convincing. And as just noted, it is a study in which the investigators' case expertise appears to play a substantial role in crafting the research design.

Natural experiments in the social sciences involve a range of interventions. *As-if* random treatment assignment may stem from various sources, including a procedure specifically designed to randomize, such as a lottery; the non-systematic implementation of certain interventions; and the arbitrary division of units by jurisdictional borders. The plausibility that assignment is indeed *as-if* random—considered here to be one of the definitional criteria for this type of study—varies greatly in research that employs this design.

Regression-Discontinuity (RD) Designs

A regression-discontinuity design is a specific kind of natural experiment. Here, as part of a social or political process, individuals or other units are assigned to one or the other category of the independent variable (i.e., the treatment or control) according to whether they are above or below a given threshold.¹⁷ For individuals near the threshold, the process that determines location above or below the threshold is as good as random, ensuring that these individuals will be similar with respect to potential confounders. This in turn opens the possibility of a more compelling causal inference about the impact on the dependent variable. The contrast with the standard natural experiment is that *as-if* random assignment specifically involves the position of subjects in relation to this threshold.

For example, in their study of the National Merit Scholarship program, Thistlewaite and Campbell (1960) compare students who received public recognition of scholastic achievement—i.e., Certificates of Merit—with those who only received commendations, with the goal of inferring the impact on subsequent academic achievement. All students who achieved a test score above a threshold received certificates, while those who performed below the threshold received commendations—which confer less public recognition of scholastic achievement. In general, students who score high on such exams will be very different from those who score low. Thus, comparisons between all high scorers who received certificates, and

17. Put differently, in a regression-discontinuity (RD) design, treatment assignment is determined by the value of a covariate, sometimes called a forcing variable, and there is a sharp discontinuity in the probability of receiving treatment at a particular threshold value of this covariate (Campbell and Stanley 1963: 61–64; Rubin 1977).

all low scorers who did not, may be misleading for purposes of inferring the effect of receiving this public recognition.

However, given that students just above and below the threshold are not very different, and given the role of unpredictability and luck in exam performance, these two groups are likely to be similar on average—with the exception that students just above the threshold receive a certificate.¹⁸ Thus, assignment to receive a Certificate of Merit can be considered *as-if* random in the neighborhood of the threshold,¹⁹ and comparisons near the threshold allow an estimate of the effects of certificates, at least for the group of students whose scores were near the threshold.

Regression-discontinuity designs have recently become increasingly common. A well-known example, which illustrates both strengths and limitations, is Angrist and Lavy (1999), who analyze the effects of class size on educational achievement, obviously an issue with wide policy implications. They gain analytic leverage by building on a requirement in contemporary education in Israel—known as Maimonides' Rule, after the 12th century Rabbinic scholar—that requires secondary schools to have no more than 40 students per classroom. In a school in which the enrollment is near this threshold or its multiples—e.g., schools with around 40, 80, or 120 students—the addition of a few students to the school through increases in enrollment can cause a sharp reduction in class sizes, since more classes must be created to comply with the rule. Thus, the educational achievement of students in schools whose enrollments were just under the threshold size of 40 (or 80 or 120) can be compared to students in schools that had been just over the threshold and were reassigned to classrooms with a smaller number of students.

In Angrist and Lavy's study, as in the classic RD design of Thistlewaite and Campbell (1960), the effect of class size can be estimated in the neighborhood of the threshold. A key feature of the design is that students do not self-select into smaller classrooms, since the application of Maimonides' rule is triggered by increases in school-wide grade enrollment. The comparison of students in schools just under or just over the relevant

18. Oddly, Thistlewaite and Campbell (1960) remove from their study group Certificate of Merit (CM) winners who also won National Merit Scholarships (NMSs); only CM winners were eligible for NMSs, which are also based on grades. This would lead to bias, since the control group includes both students who *would have* won merit scholarships had they received CMs, and those who would not have; the treatment group includes only the latter type.

19. If the threshold is adjusted after the fact, this may not be the case; for example, officials could choose the threshold strategically to select particular candidates, who might differ from students in the control group on unobserved factors.

threshold is different from comparisons between, say, college and high school graduates. The design is interesting, and there is a plausible claim of *as-if* randomness in the neighborhood of the threshold.²⁰

Instrumental-Variables (IV) Designs

An instrumental-variables design relies on the idea of *as-if* random in yet another way. Consider the challenge of inferring the impact of a given independent variable on a particular dependent variable—where this inference is made more difficult, given the strong possibility that reciprocal causation or omitted variable bias may pose a problem for causal inference. The solution offered by the IV design is to find an additional variable—an instrument—that is correlated with the independent variable but could not be influenced by the dependent variable or correlated with its other causes. In effect, the instrumental variable is treated as if it “assigns” units to values of the independent variable in a way that is *as-if* random, even though no explicit randomization occurred. In instrumental-variables analysis, the predicted values of the independent variable based on the instrument are used in place of the original independent variable.

For example, Miguel, Satyanath, and Sergenti (2004) study the effect of economic growth on the probability of civil war in Africa, using annual change in rainfall as an instrumental variable. Reciprocal causation poses a major problem in this research—civil war causes economies to grow more slowly—and many difficult-to-measure omitted variables may affect both economic growth and the likelihood of civil war. However, year-to-year variation in rainfall is plausibly *as-if* random vis-a-vis these other social and political processes, and it is correlated with economic growth. In other words, year-on-year variation in rainfall “assigns” African countries to rates of economic growth, if only probabilistically, so the predicted value of growth based on changes in rainfall can be analyzed in place of actual economic growth rates. If rainfall is independent of all determinants of civil war other than economic growth, instrumental-variables analysis allows estimation of the effect of economic growth on conflict, at least for those countries whose growth performance is shaped by variation in rainfall.

20. A few other examples of RD designs in the social sciences include the studies by Lerman (2008), who exploits an index used in the California prison system to assign convicts to higher- and lower-security prisons to study the effect of high-security incarceration; Lee (2008), who estimates the returns to incumbency by comparing near-winners and near-losers of congressional elections (though see Sekhon and Titiunik 2009 for a critique); and Dunning (2009), who takes advantage of a rule that rotates electoral quotas for lower-caste presidents of village councils in the Indian state of Karnataka.

This example illustrates both the strengths and limitations of instrumental-variables analysis. Rainfall may or may not be independent of other sources of armed conflict, and it may or may not influence conflict only through its effect on growth (Sovey and Green 2009). Variation in rainfall may also influence growth only in particular sectors, such as agriculture, and the effect of agricultural growth on civil war may be quite different than the effects of growth in the urban sector (Dunning 2008c). Because using rainfall as an instrument for growth may capture relatively specific, rather than general, effects, caution should be advised when extrapolating results or making policy recommendations.²¹

Natural experiments often play a key role in generating instrumental variables.²² However, whether the ensuing analysis should be viewed as more design-based or more model-based depends on the techniques used to analyze the data. If multiple regression models are used, the assumptions behind the models are crucial, yet the assumptions may lack credibility—and they cannot be readily validated. Instrumental-variables analysis can therefore be positioned between the poles of design-based and model-based inference, depending on the application.

CONTRAST WITH MATCHING DESIGNS

This section contrasts natural experiments with the matching designs increasingly used in the social sciences. Matching, like the standard regression analysis of observational data, is a strategy of controlling for known confounders through statistical adjustment. In matching designs, assignment to treatment is neither random nor *as-if* random. Comparisons are made across units exposed to treatment and control conditions, while addressing observable confounders—that is, those we can observe and measure.

For example, Gilligan and Sergenti (2008) study the effects of UN peacekeeping missions in sustaining peace after civil war. These authors recognize that UN interventions are non-randomly assigned to countries

21. A similar example of an instrumental-variables design is found in Hidalgo et al. (forthcoming), who use rainfall as an instrument to study the impact of economic conditions on rural land invasions in Brazil. Acemoglu, Johnson, and Robinson (2001) is another prominent example of an IV design, in which colonial settler mortality rates are used as an instrument for current political institutions.

22. Instrumental variables are also used in true randomized experiments in which some subjects do not comply with treatment assignment. Here, treatment assignment serves as an instrumental variable for treatment receipt, allowing estimation of the effect of treatment on “compliers”—that is, subjects who follow the treatment regime to which they are assigned.

experiencing civil wars. In addition, differences between countries that receive missions and those that do not—rather than the presence or absence of UN missions per se—may explain post-war differences across these countries. Working with a sample of post-Cold-War conflicts, the authors use matching to adjust for nonrandom assignment. Cases where UN interventions took place are matched—i.e., paired—with those where they did not occur, applying the criterion of having similar scores on other measured variables such as the presence of non-UN missions, the degree of ethnic fractionalization, or the duration of previous wars. The assumption is that whether a county receives a UN mission, within the strata defined by these measured variables, is like a coin flip. This analogy is implied by the assumed conditional independence of treatment assignment and potential outcomes. The study yields the substantive finding that UN interventions are effective, at least in some areas.

In contrast to natural experiments—in which *as-if* random assignment allows the investigator to control for both observed and unobserved confounders—matching relies on the assumption that analysts can measure and control the relevant (known) confounders. Some analysts suggest that matching yields the equivalent of a study focused on twins, i.e., siblings, in which one unit gets the treatment at random and the other serves as the control (Dehejia and Wahba 1999; Dehejia 2005). Although matching seeks to approximate *as-if* random by conditioning on *observed* variables, the possibility cannot be excluded that *unobserved* variables distort the results.

In addition, if statistical models are used to do the matching, the assumptions behind the models may play a key role (Smith and Todd 2005; Arceneaux, Green, and Gerber 2006; Berk and Freedman 2003).²³ When all known confounders are dichotomous, the analyst may match cases that have exactly the same values on all variables, *except* the putative cause. However, this stratification strategy of “exact matching” requires substantial amounts of data, especially if many possible combinations of confounders are present. In many applications of matching—particularly when the confounding variables are continuous—regression models are used to do the matching. An example is propensity-score matching, in which the “propensity” to receive treatment typically is modeled as a function of known confounders.²⁴ Here, analysts compare units with “similar” propensity scores but different actual exposures to treatment, with a goal of estimating the causal effect of the treatment.²⁵

23. See also the special issue on the econometrics of matching in the *Review of Economics and Statistics*, February 2004, 86 (1).

24. More technically, the probability of receiving treatment is given by the logistic or normal cumulative distribution function, evaluated at a linear combination of parameters and covariates.

25. Much of the technical literature on matching focuses on how best to maximize the “similarity” or minimize the distance between matched units; some

Propensity-score matching and related techniques are best seen as examples of model-based approaches, in which analysts attempt to adjust for pre-intervention differences between groups by modeling the unknown data-generating processes. In the case of matching, analysts model the unknown process that generated the assignment of units to treatment and control conditions. To be sure, matching can have advantages relative to conventional linear regression analysis. For example, matching focuses analytic attention on simple contrasts between treatment and control conditions, and typical matching techniques ensure that values of measured confounders among the treated group are also found among the matched control group—a condition known as “common support”—so that treated units are not compared to apparently dissimilar control units.

Still, matching is fundamentally a conditioning strategy, and its success depends on the analyst’s ability to measure and control for confounders. With natural experiments, by contrast, the *as-if* random element in the research design generates balance between treated and control units on observed as well as (one hopes) unobserved variables. For this reason, matching designs should *not* be seen as part of the family of techniques being discussed here.

EVALUATING NATURAL EXPERIMENTS: THREE DIMENSIONS

The guiding question of this chapter asks: How much leverage does research design provide? The answer—to be developed throughout the chapter—points to considerable ground for optimism, yet also points to some important grounds for concern.

To address this question, it is helpful to discuss in more detail three dimensions along which natural experiments can be evaluated: (1) plausibility of *as-if* random assignment; (2) credibility of the statistical model, which as noted above is closely connected with the simplicity and transparency of the data analysis; and (3) substantive relevance of the intervention—i.e., whether and in what ways the specific contrast between treatment and control provides insight into a wider range of important issues and contexts. The fourth criterion, substantive expertise, is not presented as a separate dimension, but it is assumed to be fundamental as an underpinning for the other three. Carefully managing the relationships, and sometimes the trade-offs, between these dimensions is crucial to developing strong research designs.

approaches include nearest-neighbor matching, caliper matching, and Mahalanobis metric matching. See Sekhon (2009) for a review.

Plausibility of As-if Random Assignment

Natural experiments present an intermediate option between true experiments and the conventional strategy of controlling for measured confounders in observational data. In contrast to true experiments, there is no manipulation of treatment variables. Yet, unlike many observational studies, they employ a design-based method to control for both known and unknown confounders. The key claim—and the definitional criterion—for this type of study is that assignment is *as-if* random. As we have seen, this attribute has the great advantage of permitting the use of simple analytic tools—for example, percentage comparisons—in making causal inferences.

Given the importance of this claim to *as-if* randomness, we must carefully evaluate the extent to which assignment meets this criterion. Figure 14.1 evaluates several studies in terms of a continuum of plausibility, drawing on the examples presented in table 14.1. This discussion is not intended as a definitive evaluation of these studies, but rather has the heuristic goal of showing how useful it is to examine studies in terms of these dimensions.

Our paradigmatic example, Snow's (1965 [1855]) study of cholera, is not surprisingly located on the far right side of this continuum. Given that the presumption of *as-if* random is highly plausible, Galiani and Schargrodsky's (2004) study of squatters in Argentina is also a good example where *as-if* random is plausible. Here, *a priori* reasoning and substantial evidence suggest that assignment to land titles met this standard—thus, confounders did not influence the relationship between the possession of titles and outcomes such as housing investment and self-perception of efficacy. Chattopadhyay and Duflo (2004) study village council elections in which quotas for women presidents are assigned virtually at random (see also Dunning 2009), while in Doherty, Green, and Gerber's (2006) study of lottery players, lottery winnings are assigned at random, which may allow for inferences about the causal effects of winnings.²⁶

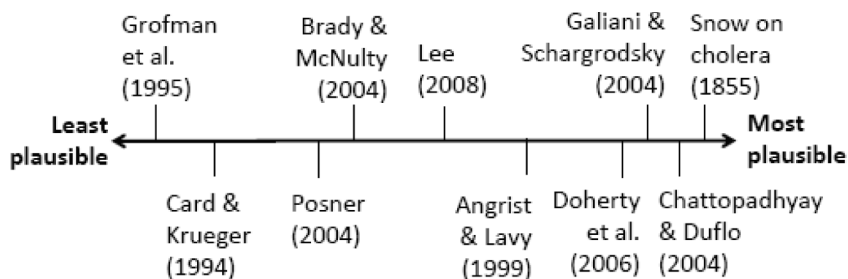


Figure 14.1. Plausibility of As-If Random Assignment

26. However, lottery winnings are only assigned at random conditional on the kind and number of lottery tickets bought; see Doherty, Green, and Gerber (2006) for details.

In parallel, Angrist and Lavy (1999) argue convincingly that according to Maimonides' Rule, students near the thresholds are assigned *as-if* at random to smaller or larger classes. In the close elections studied by Lee (2008), electoral offices may be assigned nearly at random, due to the elements of luck and unpredictability in fair elections with narrow margins. This allows for natural-experimental comparisons between near-winners and near-losers (though see Sekhon and Titiunik 2009 for a critique). In such studies, the claim of *as-if* random is plausible, which implies that post-intervention differences across treatment and control groups should not be due to confounding.

In other examples (figure 14.1), the plausibility of *as-if* random may vary considerably. Brady and McNulty (2004) study the effects on turnout of the consolidation of polling places during California's gubernatorial recall election of 2003. For some voters, the distances between their residences and their polling places had changed since the previous election; for others it remained the same. Here, the key question is whether assignment of voters to polling places in the 2003 election was *as-if* random with respect to other characteristics that affected their disposition to vote, and it appears that this standard may not have been fully met.²⁷ Posner (2004) argues that the border between Malawi and Zambia—the legacy of colonial-era borders—arbitrarily divided ethnic Chewas and Tumbukas. Of course, subsequent migration and other factors could have mitigated the *as-if* randomness of location on one side of the border or the other.

In another study, Card and Krueger (1994) analyzed similar fast-food restaurants on either side of the New Jersey-Pennsylvania border. Contrary to postulates from basic theories of labor economics, they found that an increase in the minimum wage in New Jersey did not increase—and perhaps even decreased—unemployment.²⁸ Yet do the owners of fast-food restaurants deliberately choose to locate on one or the other side of the border in ways that are related to wages and employment, thereby affecting the validity of inferences? A parallel concern might be that legislators choose minimum wage laws in ways that are correlated with characteristics of the units that will be exposed to this treatment.²⁹

27. Brady and McNulty (2004) raise the possibility that the county elections supervisor closed polling places in ways that were correlated with potential turnout, finding some evidence for a small lack of pre-treatment equivalence on variables such as age. Thus, the assumption of *as-if* random may not completely stand up either to Brady and McNulty's careful data analysis or to *a priori* reasoning (after all, election supervisors may try to maximize turnout).

28. In 1990, the New Jersey legislature passed a minimum wage increase from \$4.25 to \$5.05 an hour, to be implemented in 1992, while Pennsylvania's minimum wage remained unchanged.

29. Economic conditions deteriorated between 1990, when New Jersey's minimum wage law was passed, and 1992, when it was to be implemented. New Jersey legislators then passed a bill revoking the minimum wage increase, which the gover-

Finally, Grofman, Griffin, and Berry (1995) use roll-call data to study the voting behavior of congressional representatives who move from the U.S. House of Representatives to the Senate. These authors ask whether new senators—who represent larger and generally more heterogeneous jurisdictions (i.e., states rather than congressional districts)—modify their voting behavior in the direction of the state’s median voter.³⁰ Here, however, the treatment is the result of a representative’s decision to switch from one chamber of Congress to another. Issues of self-selection make it much more difficult to claim that assignment of representatives to the Senate is *as-if* random.³¹ Therefore, this study probably falls short of being a natural experiment in the framework of the present discussion.

A concluding point should be made about the array of studies in figure 14.1. Research that is closer to the less plausible pole more closely resembles a standard observational study, rather than a natural experiment. Such studies may well reach valid and compelling conclusions. The point is merely that in this context, researchers have to worry all the more about the standard inferential problems of observational studies.

How, then, can the assertion of *as-if* random at least partially be validated? This is an assumption, and it is never completely testable. Still, in an alleged natural experiment, this assertion should be supported both by the available empirical evidence—for example, by showing equivalence on the relevant measured antecedent variables³² across treatment and control groups—and by *a priori* knowledge and reasoning about the causal question and substantive domain under investigation. It is important to bear in mind that even when a researcher demonstrates perfect empirical balance on observed characteristics of subjects across treatment and control groups, in observational settings there typically is the strong possibility that unobserved differences across groups may account for differences in average outcomes. This is the Achilles’ heel of such studies as well as other forms of observational research, relative to randomized controlled experiments. The problem is worsened because many of the interventions that might provide

nor vetoed, allowing the wage increase to take effect (Deere, Murphy, and Welch 1995). Fast-food restaurants on the Pennsylvania side of the border were also exposed to worsened economic conditions, however.

30. Grofman, Griffin, and Berry (1995) find that there is little evidence of movement towards the median voter in the state.

31. As the authors themselves note, “extremely liberal Democratic candidates or extremely conservative Republican candidates, well suited to homogeneous congressional districts, should not be well suited to face the less ideologically skewed statewide electorate” (Grofman, Griffin, and Berry 1995: 514).

32. These variables are called “pre-treatment covariates” because their values are thought to have been determined before the treatment of interest took place. In particular, they are not themselves seen as outcomes of the treatment.

the basis for plausible natural experiments are the product of the interaction of actors in the social and political world. It can strain credulity to think that these interventions are independent of the characteristics of the actors involved, or that they do not encourage actors to “self-select” into treatment and control groups in ways that are correlated with the outcome in question. Still, strong regression-discontinuity designs, lottery studies, and other approaches can leverage *as-if* randomness to help eliminate the threat of confounding.³³

Credibility of Statistical Model

The source of much skepticism about widely-used regression techniques is that the statistical models employed require many assumptions—often both implausible and numerous—that undermine their credibility. By contrast, *as-if* randomness should ensure that assignment is statistically independent of other factors that influence outcomes, and in that case elaborate statistical models that lack credibility will not be required. The data analysis can be simple and transparent—as with the comparison of percentages or of means.³⁴

In the studies evaluated here, as becomes clear in comparing figure 14.2 with 14.1, this pattern is generally followed, though with some exceptions. The construction of figure 14.2 is parallel to figure 14.1, in that at the far left side the least credible statistical models correspond to those employed in model-based inference and mainstream quantitative methods. The most credible are those that use simple percentage or mean comparisons, placing them close to the experimental side of the spectrum.

Again, our paradigmatic example, Snow (1965 [1855]) on cholera, is

33. In a thoughtful essay, Stokes (2009) suggests that critiques of standard observational designs—by those who advocate wider use of experiments or natural experiments—reflect a kind of “radical skepticism” about the ability of theoretical reasoning to suggest which confounders should be controlled. Indeed, Stokes argues, if treatment effects are always heterogeneous across strata, and if the relevant strata are difficult for researchers to identify, then “radical skepticism” should undermine experimental and observational research to an equal degree. Her broader point is well-taken, yet it also does not appear to belie the usefulness of random assignment for estimating average causal effects, in settings where the average effect is of interest, and where random or *as-if* random assignment is feasible.

34. Such simple data-analytic procedures often rest on the Neyman-Rubin-Holland causal model (Neyman 1923, Holland 1986, Rubin 1978, Freedman 2006). Neyman’s model may be the right starting point for the analysis of data from many strong designs, including natural experiments. Below, I discuss other issues, such as the use of multivariate regression models to reduce the variance of treatment effect estimators.

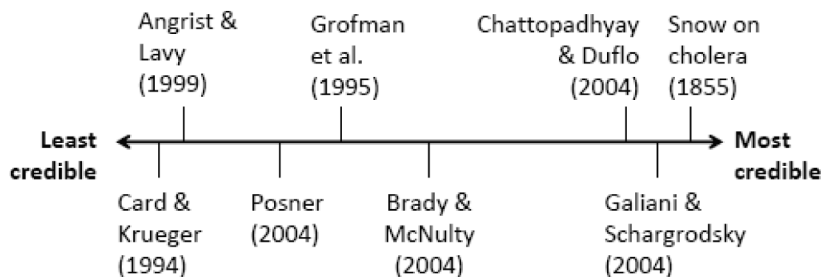


Figure 14.2. Credibility of Statistical Models

located on the far right side of the continuum. The data analysis is based simply on comparing the frequency of cholera deaths from the disease per 10,000 households, in houses served by two water companies (one with a contaminated supply).³⁵ This type of analysis is compelling as evidence of a causal effect because the presumption of *as-if* randomness is plausible. In two other studies, high credibility of the statistical model and plausibility of *as-if* random assignment also coincide. Thus, Galiani and Schargrodsky's (2004) analysis of squatters in Argentina and Chattopadhyay and Duflo's (2004) study of quotas for women council presidents in India both use simple difference-of-means tests—without control variables—to assess the causal effect of assignment. In figure 14.2, as in figure 14.1, these studies are both located on the right side. This may provide a further lesson about the elements of a successful natural experiment. When the research design is strong—in the sense that treatment is plausibly assigned *as-if* at random—the need to adjust for confounders is minimal. As Freedman (2009: 9) puts it, “It is the design of the study and the size of the effect that compel conviction”—because the often strong assumptions behind conventional regression models need not play a role in the analysis.

Unfortunately, credibility of the statistical model is not inherent in all studies that claim to use natural experiments. Consider the other examples among the 29 listed in table 14.1. The final column of the table indicates whether a simple, unadjusted difference-of-means test is used to evaluate the null hypothesis of no effect of treatment—which, where appropriate, constitutes a simple and highly credible form of statistical analysis.³⁶

35. Strictly speaking, Snow (1965 [1855]: 86, Table IX) compares death rates from cholera by source of water supply, but he does not attach a standard error to the difference (which is more than a factor of seven). Still, the credibility of the data analysis is very high.

36. An unadjusted difference-of-means test subtracts the mean outcome for the control group from the mean outcome for the treatment group and attaches a standard error to the difference. Note that in deciding whether such a test has been applied in Table 14.1, I adopt the most permissive coding possible. For example, if

Particularly given that the coding scheme employed is highly permissive in favor of scoring studies as “yes” in terms of employing difference-of-means tests (see again the preceding footnote),³⁷ it is striking in table 14.1 that over a dozen studies claiming to be natural experiments are coded as not using unadjusted differences-of-means tests.³⁸ With a more extensive list of studies that claim to be natural experiments, the proportion of simple differences-of-means tests might well fall even further.

Returning to figure 14.2 and comparing it to 14.1, note again that there is often convergence between the two figures. The discussion above noted that both the Galiani and Schargrodsky (2004) study of Argentine squatter settlements and the Chattopadhyay and Duflo (2004) electoral study are placed on the right side in both figures 14.1 and 14.2. For studies that were judged weaker on *as-if* random assignment and thus were placed on the left side of figure 14.1, the statistical analysis is correspondingly more complex, resulting in placement to the left in figure 14.2 as well. Brady and McNulty’s (2004) study of voting costs controls for possible confounders such as age; Card and Krueger (1994) also include control variables associated with exposure to minimum wage laws and with subsequent wages. In such studies, the use of multivariate regression models may reflect the possible violations of *as-if* random assignment—leading analysts to adjust for the confounders that they can measure.³⁹

an analyst reports results from a bivariate linear regression of the outcome on a constant and a dummy variable for treatment, *without control variables*, this is coded as a simple difference-of-means test (even though, as discussed below, estimated standard errors from such regressions can be misleading). More generally, the quality of the estimator of the standard errors—involving considerations such as whether the analyst took account of clustering in the *as-if* random assignment—is not considered here. All that is required for a coding of “yes” is that a difference-of-means test (or its bivariate regression analogue) be reported, along with any estimates of the coefficients of multivariate models or other, more complicated specifications.

37. See again footnote 34 and below on the rationale for difference-of-means tests.

38. Four of the studies in table 14.1 have continuous treatments or use instrumental variables, which complicates the calculation of a difference-of-means; these studies are marked with a “b.” Even excluding these studies, however, only 15 out of 25, or 60 percent of the studies, report unadjusted difference-of-means tests. Note that no special claim is made about the representativeness of the studies listed in table 14.1. Table 14.1 contains studies surveyed in Dunning (2008a), which appeared in a keyword search on “natural experiment” in JSTOR, and it is augmented to include several recent examples of successful natural experiments. These studies include some of the best natural experiments in the recent literature, analyzed by sophisticated scholars.

39. A special note should be added about the placement in Figure 14.2 of Posner’s (2004) study. This author presents a simple differences-of-means test; the key

By contrast, for other studies the position shifts notably between the two figures. Stronger designs *should* permit statistical tests that do not depend on elaborate assumptions. Yet in practice some studies in which assignment is plausibly *as-if* random nonetheless do not present unadjusted difference-of-means tests. This pattern is reflected in the contrasting positions of the Angrist and Lavy (1999) study in figure 14.1 and 14.2.⁴⁰ The contrast appears to reflect the authors' choice to report results only from estimation of multivariate models—perhaps because, as Angrist and Pischke (2009: 267) say, estimated coefficients from regressions without controls are statistically insignificant.⁴¹ On the other hand, comparing figures 14.1 and 14.2, Grofman, Griffin, and Berry is an example of a study that is evaluated as weak on the criterion of *as-if* random, yet it compares more favorably in the simplicity of the statistical model employed.⁴² Of course, such simplicity may not be justified, given the weakness of *as-if* random assignment: if unobserved confounders affect the decision of congressional representatives to run for the Senate, a simple differences-of-means test may not provide an unbiased estimator of the causal effect of treatment.

What is the major lesson here? In less-than-perfect natural experiments,

piece of evidence stems from a comparison of mean survey responses among respondents in Malawi and those just across the border in Zambia. There is a complication, however. There are essentially only two random assignments *at the level of the cluster*—living in Zambia or living in Malawi. From one perspective, this may lead to a considerable loss of precision in the estimates; at the level of the cluster, standard errors are undefined. Given this restriction, the data must be analyzed *as if* people were individually randomized rather than block randomized to these conditions—which may not necessarily be a credible statistical assumption.

40. The logic of the RD design used by Angrist and Lavy (1999) implies that treatment assignment is only *as-if* random near the threshold of the covariate determining assignment. Thus, the most defensible way to analyze data from an RD design is through a simple comparison of mean outcomes in the treatment and control groups, in the discontinuity sample of schools in the neighborhood of the relevant enrollments thresholds.

41. When estimating regression models, including control variables such as the percentage of disadvantaged students, Angrist and Lavy (1999) find that a seven-student reduction in class size raises math test scores by about 1.75 points or about one-fifth of a standard deviation. However, estimates with no controls turn out to be much smaller and are statistically insignificant, as are estimated differences-of-means in a sample of schools that lie close to the relevant regression-discontinuity thresholds (Angrist and Pischke 2009: 267). In other words, the published results in Angrist and Lavy (1999) rely on the inclusion of statistical controls in a multivariate regression model.

42. This raises the interesting question of how to analyze alleged natural experiments in which the treatment is not very plausibly *as-if* random. I focus on emphasizing the value of transparent and credible statistical analysis when the plausibility of *as-if* random assignment is high (i.e., in strong natural experiments).

in which the plausibility of *as-if* random is not strong, researchers may feel compelled to control for observed confounders. Indeed, given the absence of true randomization in many of these studies, it is not a bad idea to explore whether statistical adjustment—for example, the introduction of additional control variables in a multivariate regression—changes the estimated effects. When these changes are substantial, let the buyer beware (or perhaps more to the point, let the seller beware), because this may point to a lack of *as-if* random assignment.⁴³ In such cases, the use of statistical fixes should perhaps be viewed as an admission of less-than-ideal research designs.⁴⁴

43. One further caveat is in order. While the Neyman model that justifies simple differences-of-means tests for estimating causal effects is flexible and general (Freedman 2006), it assumes that potential outcomes for any unit are invariant to the treatment assignment of other units. This is the assumption of “no interference between units” (Cox 1958) or what Rubin (1978) called the “stable unit treatment value assumption” (SUTVA). This causal assumption does not always hold, even when the design apparently is strong: for example, Mauldon et al. (2000: 17) describe a welfare experiment in which subjects in the control group became aware of the treatment, involving rewards for educational achievement, and this may have altered their behavior. Thus, Collier, Sekhon, and Stark (2010: xv) seem to go too far when they say that “causal inference from randomized controlled experiments using the intention-to-treat principle is not controversial—provided the inference is based on the actual probability model implicit in the randomization.” Their caveat concerns inferences that depart from the appropriate statistical model implied by the randomization, but they do not address departures from the causal model on which the experimental analysis is based. Intention-to-treat analysis of an experiment such as Mauldon et al. (2000) certainly could be controversial, since the underlying causal parameter cannot appropriately be formulated in terms of the Neyman model. Of course, SUTVA-type restrictions are also built into the assumptions of canonical regression models—in which unit i 's outcomes are assumed to depend on unit i 's treatment assignment and covariate values, and not the treatment assignment and covariates of unit j .

44. Of course, researchers sometimes use multivariate regression to reduce the variability of treatment effect estimators (Cox 1958, Green 2009). Within strata defined by regression controls, the variance in both the treatment and control groups may be smaller, leading to more precise estimation of treatment effects within each stratum. However, whether variance is higher or lower after adjustment depends on the strength of the empirical relationship between pre-treatment covariates and the outcome (Freedman 2008a,b; Green 2009). Adjustment uses up degrees of freedom, which is one reason variance can be higher after adjustment. In such analysis, it is also important to note that nominal standard errors computed from the usual regression formulas do not apply, since they do not follow the design of the *as-if* randomization but rather typically assume independent and identically distributed draws from the error terms posited in a regression model. For example, the usual regression standard errors assume homoscedasticity, whereas an appropriately calculated standard error for a difference of means (see the next foot-

Of course, post-hoc statistical fixes can also lead to data mining, with only “significant” estimates of causal effects making their way into published reports (Freedman 1983). Because of such concerns, analysts should report unadjusted difference-of-means tests, in addition to any auxiliary analysis.⁴⁵ When an estimated causal effect is statistically insignificant in the absence of controls, this would clearly shape our interpretation of the effect being estimated.

Substantive Relevance of Intervention

A third dimension along which natural experiments should be classified is the substantive relevance of the intervention. Here I ask: To what extent does *as-if* random assignment shed light on the wider social-scientific, substantive, theoretical, and/or policy issues that motivate the study?

Answers to this question might be a cause for concern, for a number of reasons. For instance, the type of subjects or units exposed to the intervention might be more or less like the populations in which we are most interested. In lottery studies of electoral behavior, for example, levels of lottery winnings may be randomly assigned among lottery players, but we might doubt whether lottery players are like other populations (say, all voters). Next, the particular treatment might have idiosyncratic effects that are distinct from the effects of greatest interest. To continue the same example, levels of lottery winnings may or may not have similar effects on, say, political attitudes as income earned through work (Dunning 2008a, 2008b). Finally, natural-experimental interventions (like the interventions in some

note below) takes heteroscedasticity into account. Heteroscedasticity across the treatment and control groups is likely to arise, e.g., if treatment and control groups are of unequal size, or if treatment is effective for some subjects and not others.

45. How should the standard error for the difference of means be calculated? The sampling variance of the mean of a random sample can be estimated by the variance in the sample, divided by the number of sampled units (or the number minus one). The variance of a difference of means of two independent samples is the sum of the estimated variances of the mean in each sample. In natural experiments, the treatment and control groups can be viewed as random samples from the natural experimental population. Here we find dependence between the treatment and control groups, and we are drawing at random without replacement. Yet it is nonetheless generally valid to use variance calculations derived under the assumption of independent sampling (see Freedman, Pisani, and Purves 2007: 508-511, and A32-A34, note 11). Thus, the standard error for the difference of means can be estimated as the square root of the sum of the variances in the treatment and control groups. Statistical tests will typically rely on the central limit theorem; an alternative that can be useful when the number of units is small is to assume the strict null hypothesis of no unit-level effects and calculate p-values based on the permutation distributions of the test statistics (Fisher 1935).

true experiments) may “bundle” many distinct treatments or components of treatments. This may limit the extent to which this approach isolates the effect of the explanatory variable about which we care most, given particular substantive or social-scientific purposes. Such ideas are often discussed under the rubric of “external validity” (Campbell and Stanley 1963), but the issue of substantive relevance involves a broader question: i.e., whether the intervention—based on *as-if* random assignment deriving from social and political processes—in fact yields causal inferences about the real causal hypothesis of concern, and for the units we would really like to study.

Figure 14.3 arrays the same studies as figures 14.1 and 14.2 by the substantive relevance of the intervention. Once again, our paradigmatic example, Snow’s (1965 [1855]) study of cholera, is located at the far right side. His findings have remarkably wide substantive relevance—both for epidemiology and for public policy. Relatedly, research in epidemiology, as opposed to politics, has another key advantage. Given that causes of a certain disease may be the same across a wide range of contexts, findings routinely have broad substantive importance beyond the immediate context of the study.

In the study of politics and public policy, by contrast, what can plausibly be understood as substantive relevance will vary by context, so the degree of subjectivity involved in classifying individual studies is perhaps even greater here than with the previous two dimensions. Nonetheless, it is again useful to classify them, if only to highlight the substantial variation that can exist along this dimension among natural experiments. The studies in figure 14.3 vary, for instance, with respect to the types of units subject to a given intervention. These include voters in the Los Angeles area (Brady and McNulty 2004); fast-food restaurants near the Pennsylvania-New Jersey border (Card and Krueger 1994); children in Israeli schools that have certain enrollment levels (Angrist and Lavy 1999); politicians who move from the House to the Senate (Grofman, Griffin, and Berry 1995); village coun-

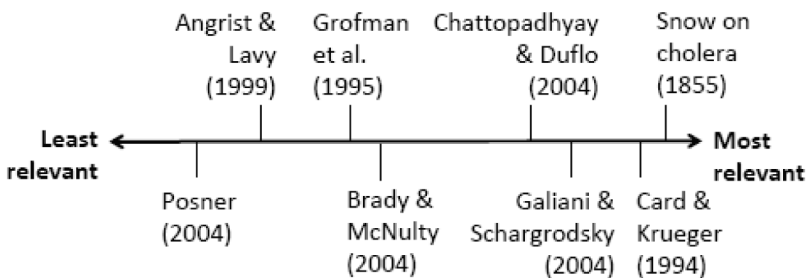


Figure 14.3. Substantive Relevance of Intervention

cils in two districts in two Indian states (Chattopadhyay and Duflo 2004); and ethnic Chewas and Tumbukas in villages near the Malawi-Zambia border (Posner 2004).

Whether the groups on which these studies focus are sufficiently representative of a broader population of interest seems to depend on the question being asked. Card and Krueger (1994), for instance, want to know whether minimum-wage laws increase unemployment in general, so any distinctive features of fast-food restaurants in Pennsylvania and New Jersey must be considered in light of this question. Brady and McNulty (2004) investigate how changes in the costs of voting shape turnout for voters in a specific electoral setting, the gubernatorial recall election in 2003, yet the impact of voting costs due to changes in polling locations may or may not be similar across different elections. Angrist and Lavy (1999) study a question of great public-policy importance—the effect of class size on educational attainment—in the particular context of Israeli schools, estimating the effect of class size for students at the relevant regression-discontinuity thresholds. In other settings—such as Grofman, Griffin, and Berry’s (1995) study of U.S. congressional representatives and senators⁴⁶—whether the group is representative of a broader population may not be of interest.

The search for real-world situations of *as-if* random assignment can narrow the analytic focus to possibly idiosyncratic contexts—as many have recently argued.⁴⁷ Of course, the extent to which this problem arises varies. In a natural experiment constructed from a regression-discontinuity design, causal estimates are valid for subjects located immediately on either side of the threshold—for example, students who score just above or below the threshold exam score; prisoners who are close to the threshold that triggers assignment to high-security prisons; and near-winners and near-losers in elections. The extent to which this limits the generality of conclusions depends on the kind of question being asked.

Moreover, there may be trade-offs in seeking a substantively relevant intervention. On the one hand, the relatively broad scope of the treatment is an attractive feature of many natural experiments, compared to some true experiments. After all, this approach can allow us to study phenomena—such as institutional innovations, polling place locations, and minimum wage laws—that routinely are not amenable to true experimental manipulation.⁴⁸ On the other hand, as discussed below, some broad and substantively-relevant interventions may not plausibly achieve *as-if* randomness.

46. The placement of the Posner study on figure 14.3 is discussed further below.

47. See Deaton (2009), Heckman and Urzua (2009), and the reply from Imbens (2009).

48. It is true, however, that some experimental researchers have become increasingly creative in developing ways to manipulate apparently non-manipulable treatments, thereby broadening the substantive contribution of that research tradition.

Another challenge relevant to substantive importance is “bundling,” a problem that arises when the treatment contains multiple explanatory factors, such that it is hard to tell which makes a difference. While broad interventions that expose the subjects of interest to an important intervention can appear to maximize theoretical relevance, the bundling in some such interventions can complicate interpretation of the treatment.

An illustration of this point is the study by Posner (2004), who asks why cultural differences between the Chewa and Tumbuka ethnic groups are politically salient in Malawi but not in Zambia.⁴⁹ According to Posner, long-standing differences between Chewas and Tumbukas located on either side of the border cannot explain the different inter-group relations in Malawi and in Zambia. Indeed, he argues that location in Zambia or Malawi is *as-if* random: “like many African borders, the one that separates Zambia and Malawi was drawn purely for [colonial] administrative purposes, with no attention to the distribution of groups on the ground” (Posner 2004: 530). Instead, factors that make the cultural cleavage between Chewas and Tumbukas politically salient in Malawi but not in Zambia presumably should have something to do with exposure to a treatment (broadly conceived) received on one side of the border but not on the other.

Yet such a study must face a key question which sometimes confronts randomized controlled experiments as well: What, exactly, is the treatment? To put this question in another way, which aspect of being in Zambia as opposed to Malawi causes the difference in political and cultural attitudes? Posner argues convincingly that inter-ethnic attitudes vary markedly on the two sides of the border because of the different sizes of these groups in each country, relative to the size of the national polities (see also Posner 2005). This difference in the relative sizes of groups changes the dynamics of electoral competition and makes Chewas and Tumbukas political allies in pop-

However, a trade-off may certainly arise between the scope of an intervention and manipulability by experimental researchers.

49. Separated by an administrative boundary originally drawn by Cecil Rhodes’ British South African Company and later reinforced by British colonialism, the Chewas and the Tumbukas on the Zambian side of the border are similar to their counterparts in Malawi, in terms of allegedly “objective” cultural differences such as language, appearance, and so on. However, Posner finds very different inter-group attitudes in the two countries. In Malawi, where each group has been associated with its own political party and voters rarely cross party lines, Chewa and Tumbuka survey respondents report an aversion to inter-group marriage and a disinclination to vote for a member of the other group for president. In Zambia, on the other hand, Chewas and Tumbukas would much more readily vote for a member of the other group for president, are more disposed to intergroup marriage, and “tend to view each other as ethnic brethren and political allies” (Posner 2004: 531).

ulous Zambia but adversaries in less populous Malawi.⁵⁰ Yet interventions of such a broad scope—with so many possible treatments bundled together—can make it difficult to identify what is plausibly doing the causal work, and the natural experiment itself provides little leverage over this question (see Dunning 2008a).⁵¹

Indeed, it seems that expanding the scope of the intervention can introduce a trade-off between two desired features of a study: (1) to make a claim about the effects of a large and important treatment, and (2) to do so in a way that pins down what aspect of the treatment is doing the causal work. Thus, while Posner's study asks a question of great substantive importance, the theoretical or substantive relevance of the treatment can be more challenging to pin down, as reflected in the study's placement in figure 14.3.⁵²

Comparing figure 14.3 to 14.1 and 14.2, we see some examples of studies in which the placement lines up nicely on all three dimensions. The study by Chattopadhyay and Duflo (2004)—as with the study by Snow—not only has plausible *as-if* randomness and a credible statistical analysis, but also speaks to the political effects of empowering women through electoral quotas. This topic's wide substantive relevance is evident, even when the particular substantive setting (village councils in India) might seem idiosyncratic. Similarly, Galiani and Scharfrodsky's study of land titling has wide substantive and policy relevance, given the sustained focus on the allegedly beneficial economic effects of property titles for the poor.

With other studies, by contrast, the placement in figure 14.3 stands in sharp contrast to that in 14.1. The study of Card and Krueger (1994), for example, while having less plausible *as-if* randomness and more complicated statistical analysis than other studies, incisively explores the effects of minimum wage level, which is of wide substantive and policy importance. This observation reinforces the point that different studies may manage the trade-off among these three dimensions in different ways, and which trade-offs are acceptable (or unavoidable) may depend on the question being asked. Again, reconciling such competing objectives and thereby realizing

50. In Zambia, Chewas and Tumbukas are mobilized as part of a coalition of Easterners; in much smaller Malawi, they are political rivals.

51. Clearly, the hypothesized "intervention" here is on a large scale. The counterfactual would involve, say, changing the size of Zambia while holding constant other factors that might affect the degree of animosity between Chewas and Tumbukas. This is not quite the same as changing the company from which one gets water in mid-nineteenth century London.

52. Many other studies use jurisdictional boundaries as sources of natural experiments; see, e.g., Banerjee and Iyer (2005), Berger (2009), Krasno and Green (2005), Laitin (1986), or Miguel (2004).

the full potential of design-based inference demands substantive knowledge and close attention to context.

CONCLUSION: SOURCES OF LEVERAGE IN RESEARCH DESIGN

This final section draws together the discussion, first by juxtaposing these three dimensions in an overall typology, and second by examining the role of qualitative evidence in good research design.

Typology: Relationship among the Dimensions

Following the numbering of the figures above, the typology in figure 14.4 brings together the three dimensions: (1) plausibility of *as-if* random assignment, (2) credibility of the statistical models, and (3) substantive relevance of the intervention. To reiterate, standing behind these should be the deep substantive knowledge that supports careful work on the three dimensions. Adding this fourth dimension the cube would make it at best unwieldy, and it is sometimes difficult to assess the investigators' level of

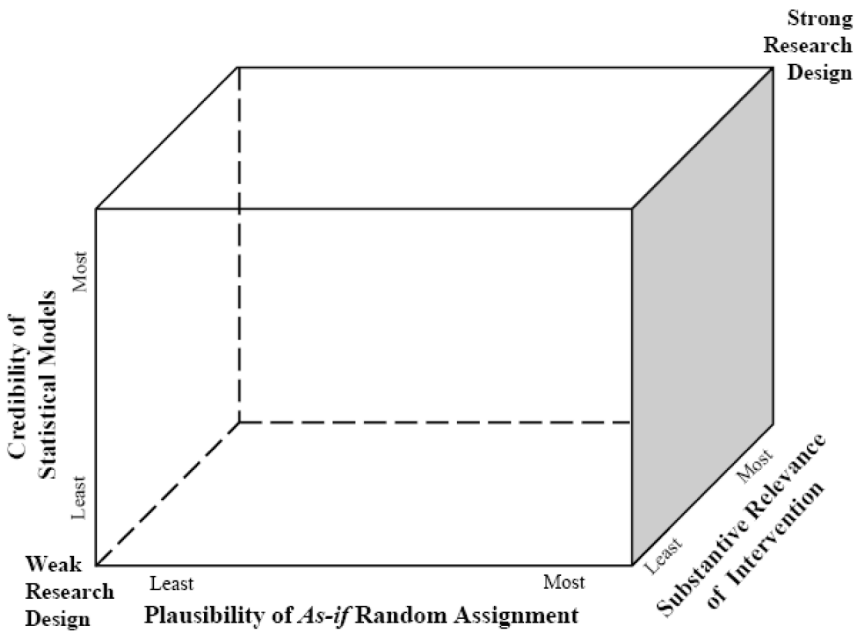


Figure 14.4. Typology of Natural Experiments

expertise simply on the basis of published articles. But from the standpoint both of evaluating natural experiments and making recommendations for conducting them, this fourth component is also critical.

Any natural experiment, and indeed any piece of research, can be placed in the three-dimensional space represented in the cube. The uppermost corner to the back and right corresponds unambiguously to a strong research design, and the bottom corner to the front and left is a weak research design—i.e., furthest from this ideal. The previous sections have made clear that these three dimensions are interconnected, and the cube is valuable for exploring these interconnections further.

As a base line, we can situate conventional, regression-based analysis of observational data within the cube. (1) These studies make no pretense of *as-if* random assignment, so they will be on the far left side. (2) Credibility of the statistical models varies considerably. Given the complex statistical modeling that is common in regression studies, it can readily be argued that the credibility and transparency of the statistical model is routinely low, placing these studies toward the bottom of the cube. (3) Finally, such regression studies may potentially allow greater scope in terms of the wider relevance of the analysis. For example, they can focus on macro-political outcomes of enormous importance, such as war, political regimes, and national political economy.⁵³ Hence, on this third dimension, they may contribute more than natural experiments. Of course, as critics such as Seawright (chap. 13, this volume) have suggested, the credibility of statistical models in these studies may be so low that the apparent contribution in terms of wider relevance may potentially be obviated.

To summarize the placement of regression-based, observational studies, they will be at the far left side of the cube and often in the lower part of the cube, reflecting the weaknesses just noted. However, they may be further toward the back, given their potential wider relevance compared to at least some natural experiments. The cube thus brings into focus what is basically conventional wisdom about these regression studies, and provides a useful point of departure for evaluating other research designs.

True experiments can also, at least approximately, be placed in the cube. (1) Genuine random assignment (and not merely *as-if* random) is presumably their defining characteristic, though in too many poorly designed experiments this is not achieved. Hence, taking the left-right dimension of the cube as a proxy for the plausibility of randomization in true experiments, many true experiments are not merely at the right side of the cube, but in a sense are well beyond it. For experiments with inadequate random-

53. Moreover, the estimation of complex models produces research that is not transparent to readers with a substantive interest in politics but less-than-expert technical knowledge.

ization, they will to varying degrees be more toward the left side. (2) The statistical models should in principle be credible and simple, though too often they are not—either because the investigator seeks to correct for a failure of random assignment, or because the temptation to employ elaborate statistical models is so engrained. (3) Depending on the ingenuity of the investigator, these studies potentially are of wide relevance, but again they may not be. Overall, experimental researchers can and should strive for the uppermost corner to the back and right in the cube—which is labeled as a “Strong Research Design”—but they potentially may fall short on any or all the dimensions.

Turning to natural experiments, I begin with our paradigmatic example, Snow’s (1965 [1855]) study of cholera. It is located at the upper-back, right-hand corner of the cube (Strong Research Design)—reflecting high plausibility of *as-if* randomization, strong credibility of the statistical model, and wide substantive importance. It is paradigmatic precisely because it is situated in this corner, and it is probably more successful on these dimensions than a great many true experiments. The natural experiments of Chattopadhyay and Duflo (2004), and well as Galiani and Schargrodsky (2004), are also located near this corner. Many other studies discussed above have weaknesses on one or more dimensions, which to varying degrees pushes them toward the lower-front, left-hand corner of the cube (Weak Research Design).

The cube is also helpful in reviewing the trade-offs discussed above. Achieving (1) plausible *as-if* randomness may come at the expense of (3) broad substantive relevance. Alternatively, (3) striving for broad substantive relevance may occur at the expense of (1) plausible *as-if* randomness, which may push the investigator toward (2) more complex and less credible statistical models.

Discussion of the cube likewise provides an opportunity to draw together the assessment of the studies in table 14.1 that employ regression discontinuity (RD) designs and instrumental variables designs (IV). Four of each type of study are included in the table. RD designs may (1) have plausible *as-if* randomness in the neighborhood of the threshold, and (2) data analysis may potentially be simple and transparent, as when mean outcomes are compared in the neighborhood of this threshold. Yet a trade-off can readily arise here. Data may be sparse near the threshold, which together with other factors may encourage analysts to fit complicated regression equations to the data, thus potentially jeopardizing the study in the credibility of the statistical models. As for (3) relevance, with an RD design causal effects are identified for subjects in the neighborhood of the key threshold of interest—but not necessarily for subjects whose values on the assignment variable place them far above or far below the key threshold. Whether a given RD study has broad substantive relevance (as in Angrist and Lavy

1999) or is somewhat more idiosyncratic may depend on the representativeness of subjects located near the relevant threshold.

For instance, to return to an earlier example of an RD design, perhaps recognition in the form of a Certificate of Merit is less important for exceptionally talented students than for much less talented students. For students at a middle level of talent and achievement, the salience of the national recognition may be harder to predict; perhaps it gives them an important boost and motivation, while failure to receive this recognition for students at middle level may weaken their motivation for further achievement. Thus, relevance might be undermined if the RD design produces a somewhat idiosyncratic finding that is only relevant to a specific subgroup—i.e., the group of students near the threshold score for Certificates.⁵⁴

For instrumental variables designs, substantive relevance may also be high. For example, the effect of economic growth on civil conflict in Africa studied by Miguel, Satyanath, and Sergenti (2004), is (3) a question of great policy importance. Yet perhaps precisely because scholars aim at broad substantive questions in constructing IV designs, these designs have significant limitations as well as strengths. The instrument (1) may or may not plausibly be *as-if* random. It may or may not influence the outcome exclusively through its effect on the main explanatory variable, and may or may not influence components of this variable which have idiosyncratic effects on the outcome of interest (Dunning 2008c). In practice, data analysis in many IV designs depends on (2) complicated statistical models, whose potentially questionable credibility may make these designs less compelling than other types of natural experiments.

Overall, the cube reminds us that good research routinely involves reconciling competing objectives (chap. 8, this volume). Strong research designs can help overcome issues of confounding that bedevil causal inference in many settings. Moreover, in some contexts natural experiments address questions of broad substantive relevance. Yet the extent to which they do so varies, and the contribution on each dimension must be weighed against the others in evaluating particular studies.

Contribution of Qualitative Evidence

The contribution of qualitative evidence must also be underscored. The qualitative methods discussed throughout this volume make a central contribution to constructing and executing natural experiments. I have emphasized that the substantive knowledge and detailed case expertise often

54. Whether the effect for this group of students is meaningful for inferences about other kinds of students may be a matter of opinion; see Deaton (2009) and Imbens (2009) for a related discussion.

associated with qualitative research is essential for working with the three dimensions of natural experiments discussed throughout this chapter (Dunning 2008a).

Returning one more time to our paradigmatic example—Snow’s study of cholera—Freedman makes clear (chap. 11, this volume) that qualitative evidence plays a central role. Indeed, Freedman labels the use of qualitative evidence as a “type of scientific inquiry,” which in this instance is used jointly with another type—the natural experiment.

Consider also Galiani and Schargrotsky’s study of squatters in Argentina. Here, strong case-based knowledge was necessary to recognize the potential to use a natural experiment in studying the effect of land titling—after all, squatters invade unoccupied urban land all the time. Yet it is undoubtedly rare that legal challenges to expropriation of the land divide squatters into two groups in a way that is plausibly *as-if* random. Many field interviews and deep substantive knowledge were required to probe the plausibility of *as-if* randomness—that is, to validate the research design. In many other examples, case-based knowledge was clearly crucial in recognizing and validating the alleged natural experiment. To mention just two, Angrist and Lavy (1999) not only knew about Maimonides Rule in Israel but also recognized its social-scientific potential, while Lerman (2008) gained insight into the assignment process of prisoners to high-security prisons through many qualitative interviews and sustained observation of the California prison system.

Hard-won qualitative evidence can also enrich analysts’ understanding and interpretation of the causal effect they estimate. What does property *mean* to squatters who receive titles to their land, and how can we explain the tendency of land titles to shape economic or political behavior, as well as attitudes towards the role of luck and effort in life? Qualitative assessment of selected individuals subject to *as-if* random assignment may permit a kind of “natural-experimental ethnography” (Paluck 2008; Dunning 2008b) that leads to a richer understanding of the mechanisms through which explanatory variables exert their effects.⁵⁵ Indeed, qualitative research, conducted in conjunction with quantitative analysis of natural experiments, may contribute substantial insight in the form of what Collier, Brady, and Seawright call “causal process observations” (chap. 9, this volume; see also Freedman, chap. 11, this volume).

Thus, natural experiments and other strong designs should in principle be strongly complementary to the kinds of qualitative methods emphasized elsewhere in this book. The case-based knowledge of many qualitatively-oriented researchers may allow them to recognize the possibility of

55. The term borrows from Sherman and Strang (2004), who describe “experimental ethnography.” See Paluck (2008).

conducting this type of research. Such scholars may be especially well-positioned to employ these strong designs as one methodological tool in an overall research program.

In conclusion, it seems that many modes of inquiry contribute to successful causal inference. Ultimately, the right mix of methods substantially depends on the research question involved. In every study, analysts are challenged to think critically about the match between the assumptions of models and the empirical reality they are studying. This is as much the case for true experiments and natural experiments as it is for conventional observational studies. Convergent lines of evidence, including various kinds of qualitative inquiry, should be developed and exploited (Freedman, chap. 11, this volume). There will always be a place for conventional regression modeling and matching designs based on observational data, because some interesting and important problems will not easily yield themselves to strong research designs. Yet where strong designs are available, the researcher should resist the impulse to fit conventional statistical models to the data from such designs—the assumptions behind which are not validated by the design. At a minimum, the assumptions behind the models and the designs should be defended. As with the many other analytic tasks discussed in this chapter, this defense is most effectively carried out using diverse forms of quantitative—and also qualitative—evidence.

RETURNING TO THE GUIDING QUESTION

I return now to the guiding question of this chapter: What leverage is provided by research design, and specifically by natural experiments, in overcoming the pitfalls of regression analysis? This chapter has explored many trade-offs and potential failures in natural experiments. Ideally, based on carefully crafted scholarship, these research designs can move toward the Strong Research Design corner in the typology. But they can equally well move toward the Weak Research Design corner, which should be a matter of concern.

This chapter has deliberately concentrated on meritorious examples of natural experiments—with the goal of drawing together and evaluating some of the most interesting work in this field. With weaker examples, the picture would be grimmer and the conclusions more pessimistic. In relation to the criterion of substantive relevance, there is a legitimate concern that too much scholarship might come to focus on discovering ingenious natural experiments, at the cost of larger substantive agendas.

Finally, like conventional regression analysis, this form of design-based inference depends critically on substantive expertise to guide the numerous choices in carrying out either approach. Natural experiments, like regres-

sion analysis, do not provide a technical quick fix to the challenges of causal inference. Yet the many examples discussed in this chapter also demonstrate that design-based natural experiments have numerous strengths, and this methodology certainly merits the growing attention it receives as a fundamental approach to research.

Overall, then, the answer to this chapter's guiding question—does strong research design take us beyond the pitfalls of conventional regression modeling?—is a cautious yes. Yet, design-based inference is not easy to do. There is no technical rule-of-thumb that allows analysts to develop strong research designs. Rather, design-based research is valuable to the extent it builds on real substantive knowledge and appropriate methodological craftsmanship, and a full awareness of the trade-offs inherent in this style of investigation.

Glossary

Jason Seawright and David Collier

This glossary defines methodological terms employed in this book. The core definition is presented in the initial paragraph of each entry, and additional paragraphs are included for terms that require more elaboration. Some definitions are drawn directly from the text.¹

For entries that extend beyond one paragraph, the initial paragraph is intended to provide a self-contained definition that may be sufficient for many readers. Cross-references to related terms are identified in boldface, with the exception of a few terms used so frequently that the repeated bolding would be distracting. Page references to the corresponding discussion in the text are noted in the index.

antecedent variable. A type of **independent variable** that stands causally prior to another **explanatory variable**, which may be called an **intervening variable**. A variable's categorization as antecedent or intervening is not a permanent status, but is understood in relation to a particular causal model. See **endogenous variable**, **exogenous variable**.

1. While in general we do not use bibliographic citations in the glossary, we do occasionally reference particular authors, for example, to highlight terms that appear in new chapters in the second edition. For definitions of methodological terms, Schwandt (1997) is a useful reference for the qualitative tradition, and Darnell (1994) and Kennedy (1998) are good sources for econometric terms. Van Evera (1997), Vogt (1999), and Gerring (2001, 2011 forthcoming) give many useful definitions relevant to both the qualitative and quantitative traditions. Some of our definitions parallel the usage in the text of King, Keohane, and Verba's (KKV) *Designing Social Inquiry*. For a discussion of their use of terms, see chapter 2 in the present volume.

as-if random assignment. In the family of techniques called natural experiments, the assumption that cases enter into the categories of the independent variable (“treatment” and “control”) through a process that is independent of any possible confounders. Researchers typically cannot prove that assignment is truly random. Yet this assertion should be validated to the extent possible, through quantitative and qualitative evidence and through informed reasoning about the substantive domain under study. (In the second edition, this design is extensively discussed by Dunning, chap. 14). See **regression-discontinuity design**, **instrumental-variables design**.

assumption. An underlying premise about the characteristics of a model being estimated, of the data being analyzed, and/or of the contexts from which the data are drawn. Although such premises are often difficult to test, they play a central role in descriptive and causal inference. To the degree that the assumptions made in a particular analysis are not met, inferences drawn from the analysis are questionable. Assumptions are sometimes misunderstood as relevant only to quantitative analysis, but in fact all forms of research depend on assumptions. See **causal homogeneity**, **conditional independence**, **constant causal effects**, **independence of observations**, **specification assumption**.

autocorrelation. A failure of the assumption of **independence of observations**, due to patterns of influence among observations that are either temporally or spatially proximate.

Bayesian inference. Procedures for statistical inference in which the researcher’s preexisting knowledge and beliefs are quantified as a prior probability that is used as a baseline to be adjusted on the basis of empirical evidence.

This approach contrasts with more traditional **significance tests**, which evaluate a **null hypothesis** (typically of “no relationship”) against an alternative hypothesis (typically of “some relationship”). Empirical data are then used to either reject or fail to reject the null hypothesis. While many scholars believe that the strict application of a Bayesian framework is inappropriate in much social science research, several ideas underlying Bayesian inference serve as a valuable point of reference.

bias. **Systematic error** in inference. With bias, successive **errors** cannot be expected to cancel each other out, and inferences will therefore be faulty, even with extremely large samples. Contrast with **random error**. See **selection bias**, **missing variable bias**.

Boolean algebra. A mathematical representation of formal logic. In Ragin's (1987) **qualitative comparative analysis**, Boolean algebra is used to formalize arguments about causal relations among dichotomous variables.

Campbell's checklist of threats to validity. An inventory of threats to validity in causal inference presented in Donald Campbell's classic work² on experimental and quasi-experimental research.

Campbell's perspective is especially relevant to inferences based on time-series data, and it represents a valuable supplement to the perspective on causal inference conventionally offered by regression analysis and **econometrics**. Examples of these threats to valid inference are history, maturation, instrumentation, selection, and mortality.

case-oriented research. Research in which the center of attention is the close analysis of one or a few cases.

This approach contrasts with **variable-oriented research** (Ragin 1987). Case-oriented researchers certainly think in terms of variables, but their attention is strongly focused on detailed contextual knowledge of specific cases and on how variables interact within the context of these cases. See **case; causation, multiple and conjunctural; comparative method**.

cases. The units of analysis in a given study. Cases are the political, social, institutional, or individual entities or phenomena about which information is collected and inferences are made. Examples of cases are nation-states, social movements, political parties, trade union members, and episodes of policy implementation.

In a **rectangular data set** the rows correspond to the cases, that is, to what we are calling **data-set observations**. In a given study, the scholar may shift to a different **level of analysis**, so the definition of a case may change. However, if the goal of this shift is to provide greater analytic leverage at the original level of analysis, as in **within-case analysis**, then the original definition of "case" still corresponds to the predominant focus of the analysis.

case selection. Identification of cases for analysis in a given study. This is a fundamental task in **research design**. See **sample, universe of cases**.

case-study. A **research design** focused on one ($N=1$) or a few cases, typically analyzing the case(s) in great detail through **cross-case** or **within-**

2. Campbell and Stanley (1963); Campbell and Ross (1968); Cook and Campbell (1979).

case analysis. Ragin's (1987) **case-oriented research**, with its emphasis on contextually specific patterns of causation, is one version of the case-study method. See **qualitative-quantitative distinction**.

causal effect. The impact of a given **explanatory variable** on a particular outcome. More specifically, other things being equal, the causal effect is the difference between the two values of the dependent variable that arise according to whether an independent variable assumes one of two specific values. **Causal inference** seeks to estimate such causal effects. This definition is understood as applying both to quantitative and qualitative analysis.

causal heterogeneity. Presence of contrasting causal patterns. Thus, it is the absence of **causal homogeneity**.

causal homogeneity. The assumption that, other things being equal, a given set of values for the **explanatory variables** always produces the same **expected value** for the dependent variable within a given set of cases. The causal homogeneity assumption is met if the scores on the dependent variable for all the cases included in the analysis are produced in accordance with one shared causal model. Thus, if all cases were, counterfactually, assigned the same values on the independent variables, they would have the same expected value on the dependent variable.

If this assumption is not met, yet a researcher analyzes the data as if it were met, the inferences will be misleading because they will average together different patterns of causation among subgroups of cases. In such situations, researchers may either divide the sample of cases and make inferences within each causal subset, or develop a more complex causal model that incorporates the differences between the subsets. Given these two possibilities, causal homogeneity may be seen as a property of the data *in relation to* a given causal model.

In the statistical literature on causation (e.g., Rubin 1974; Holland 1986), a stronger version of this assumption is presented, which is called **unit homogeneity**. According to this version, different units are presumed to be *fully identical* to each other in all relevant respects except for the main independent variable and, potentially, the dependent variable. Unit homogeneity is sufficient to allow causal inference without the assumption of **conditional independence**, but it is also unlikely that this strong homogeneity assumption holds in typical social science applications, even in experiments. This assumption is generally violated by the fact that no two individuals share identical life histories.

Although KKV uses the label **unit homogeneity** in discussing assumptions, its framework in fact relies on the idea of causal homogeneity. Hence, in discussing KKV's arguments, we use the label "causal homogeneity." See **constant causal effects, expected value.**

causal inference. The process of reaching conclusions about causation on the basis of observed data. See **descriptive inference, inference.**

causal inference, fundamental problem of. The major problem of causal inference, according to many philosophers of science. Given a **counterfactual definition of causation**, the problem is that—for a given case at a given point in time—the researcher can observe either the *presence* of the cause (and of its presumed effect), or the *absence* of the cause (and hence potentially the absence of its presumed effect), but not both. Therefore, the researcher can never make the comparisons that directly meet the criteria of the counterfactual definition, and must instead turn to imperfect real-world comparisons among cases. KKV (79–80) devotes central attention to this problem.

causal mechanism. A link or connection in a **causal process**. In the relationship between a given independent variable and a given dependent variable, a causal mechanism posits additional variables, sometimes called **intervening variables**, that yield insight into how the independent variable actually produces the outcome, including the sequence through which this occurs. Compared to the original causal relationship that the scholar is considering, the causal mechanism is often located at a more fine-grained level of analysis.

causal model. A framework of concepts and insights that provides a theoretical rationale for a set of hypothesized explanatory relationships. This term is most often used in referring to a "specified" form of a causal model that posits specific variables and particular relationships among those variables.

A causal model draws on, and is part of, a **theory**. Causal models are not necessarily expressed in equations, but they can be. Quantitative researchers routinely formalize such models, for instance, with regression equations. Qualitative researchers generally do not, though Ragin (1987, 2000) has used tools of Boolean algebra and fuzzy-set logic to formalize some kinds of qualitative analysis.

Several distinctions used to characterize particular types of variables or data sets (for example, independent versus dependent variable, higher versus lower degrees of freedom, or causal homogeneity versus heteroge-

neity) are only meaningful in relation to a particular causal model. A causal model may be derived from, or linked to, other forms of models, such as a game-theoretic model, but the concern in the present volume is with causal models as particular specifications of causal relations among variables. The ideas of **causal sequence**, **causal process**, and **causal mechanism** are elements of a causal model.

causal process. A sequence of events or steps through which causation occurs.

causal-process observations (CPOs). Pieces of data that provide information about context, process, or mechanism and contribute distinctive leverage in causal inference. They are contrasted with **data-set observations (DSOs)**, which correspond to the familiar rectangular data set of quantitative researchers. In quantitative research, the idea of an “observation” (as in DSO) has special status as a foundation for causal inference, and we deliberately incorporate this term in naming CPOs so as to place their contribution in causal inference on an equivalent footing. Obviously, we do not thereby mean that one directly *observes* causation. Rather, this involves *inference*, not *direct observation*. Process tracing is the overall research procedure, which identifies specific CPOs that yield valuable leverage in causal assessment.

causal sequence. Two or more steps in a causal chain that generally correspond to a chronological sequence. Similar to **causal process**, but with more emphasis on the idea of discrete causal steps. See **intervening variable**, **causal mechanism**.

causation, multiple and conjunctural. A causal pattern in which (a) alternative (i.e., multiple) combinations of factors can produce a given outcome, and (b) any one of these causal paths may involve the interaction (i.e., conjunction) of two or more explanatory factors. Ragin (1987) has formalized this perspective on causation with **Boolean algebra**.

cause. A factor that helps to bring about the occurrence of an outcome. Specific types of causes include deterministic, necessary, probabilistic, sufficient. See **causal homogeneity**, **causal inference**, **causal mechanism**, **causal model**, **causal process**, **causal sequence**.

censoring. See **truncation**.

classification. As a verb, the process of sorting cases into the categories of a conventional nominal or ordinal scale, or **typology**. As a noun, a con-

ceptual schema consisting of an organized set of analytic categories that may be used in making theoretical distinctions and categorizing cases.

comparative-historical analysis. Research combining: (1) a sustained comparative analysis of a well-defined set of national cases; (2) a focus on the unfolding of causal processes over time; and (3) the use of systematic comparison to generate and/or evaluate explanations of outcomes. Specific studies may be identified with this tradition even though they do not have all of these attributes.

comparative method. The systematic analysis of a relatively small number of cases. It involves a smaller N than most statistical studies, but a larger N than a case study. Tools associated with the comparative method include procedures for **concept formation**, standard practices for looking at **matching** and **contrasting cases**, and using theory to identify **most-likely** and **least-likely cases**.

complexification based on extreme cases. The tendency of research focused on cases with extreme values on the dependent variable to yield new, but potentially idiosyncratic, explanations. Such complexification may provide insight, yet may also distract from identifying causal patterns that are easier to detect in the full range of cases. This issue arises in discussions of **selection bias**.

concept. Various understood as an abstract idea that offers a point of view for understanding some aspect of our experience; an idea of a phenomenon formed by mentally combining its attributes; a mental image that, when operationalized, helps to organize the analysis of data. The word employed to label any particular concept is often called a **term**.

It is productive to distinguish the "classical" and the "frame" views of concepts. The classical view focuses on defining attributes and understands **concept formation** as centrally concerned with making careful choices about the **intension**, that is, the set of meanings associated with the concept itself; and the **extension**, that is, the range of cases seen as instances of the concept. By contrast, the frame perspective treats a concept as one component in a stylized scenario, or idealized cognitive model that constitutes a point of view for thinking about some domain within the real world. Here concept formation is centrally concerned with reasoning about the relationships among different components of this scenario or model and about their implications for the particular concept.

Though this distinction between the classical and the frame views of concepts is useful, these two perspectives in some respects overlap, and many

scholars hold elements of both. Qualitative and quantitative analysts may combine elements of the two perspectives in **concept formation**, in the **operationalization** of concepts, and in establishing **measurement validity**.

concept formation. The process of specifying and refining **concepts** employed in empirical research.³ Concepts may be provided by the observer (the *etic* approach), by the actors being studied (the *emic* approach), or by a combination of the two. Analyzing the concepts of the actors being studied involves **interpretation**.

conceptualization. See concept formation.

conceptual stretching. A form of measurement error that arises when scholars inappropriately apply established concepts and theories to new contexts. Prior assumptions about the meaning of some components of the concept, and about the interrelations among these components, are not met in these new contexts.

concreteness (as a property of theory). Precisely stated, and making specific predictions. Such a theory is, in principle, easier to **falsify**.

conditional independence. An assumption used to justify causal inferences based on observational data, that is, in the absence of a true experiment.

In an experiment, “independence” is achieved when the assignment of cases to the treatment and control groups is statistically unrelated to other characteristics of those cases that may influence the dependent variable in the study. **Random assignment** meets this criterion. With **observational data**, scholars seek to approximate independence by using tools such as **stratification** to control for, or “condition” on, relevant control variables—thereby achieving “conditional independence.” The assumption of conditional independence is similar in meaning to, although different in emphasis from, the **specification assumption**.

confounder. A theoretically relevant variable that, if added to a causal model, improves the causal inference. It is also called a **missing variable** or **omitted variable**. Adding **intervening variables** to a model may change the estimates of the direct effects of some explanatory variables, but not the estimates of total effects; intervening variables are not considered confounders.

3. Political theorists may of course engage in concept formation for other purposes, but our concern here is with the empirical application of concepts.

constant causal effects. The assumption that, other things being equal, a given increment in the explanatory variable always produces a fixed magnitude of change in the dependent variable. This standard is in effect equivalent to the assumption that the relationship between the independent and the dependent variable is linear, and that the independent variable does not appear in an **interaction term** with any other variables.

This differs from the **causal homogeneity** assumption, which requires that, other things being equal, all cases have the same **expected value** of the dependent variable for given values of the independent variables.

constructivism. A research tradition focused on how social and psychological processes influence the way people view, and in part create, reality. It is the study of how human beings, individually and collectively, constitute their world. Some usages also encompass the idea of a “reflexive” perspective, involving a concern on the part of researchers with the implications of their own social position for the focus and findings of their research. See **interpretation**.

context. The political, social, and historical setting within which the phenomenon under study is located. In **descriptive inference**, detailed knowledge of context may lead the scholar to recognize the need for **contextualized comparison**; in **causal inference**, such knowledge may lead to the refinement of the **causal model**. Human understanding inevitably draws selectively on the context, which is typically too complex to be entirely understood.

contextualized comparison. Measurement procedures that take into account differences in **context**. The goal is to establish the appropriateness of concepts and the equivalence of measurement across contexts. These procedures acknowledge that the interpretation of indicators, or even the indicators themselves, may need to vary across contexts if they are to validly measure a given concept.

contrasting cases. A set of cases that have very different scores on a variable of concern. For example, with a dichotomous variable, cases that have positive scores and cases that have negative scores; with a continuous variable, cases that have high scores and low scores. See **matching cases**.

contrast space. The analytic frame that establishes the range of a variable, based on identifying conceptually relevant **positive** and **negative cases**. The idea of contrast space is closely associated with the question, “as opposed to what?”

control. A key element in the evaluation of causal effects. One may distinguish between experimental control and statistical control.

In *experiments*, “control group” refers to the cases to which the experimental treatment is not applied. Comparing the treatment group with the control group is the basis for assessing the direction and magnitude of the causal effect.

In social science discussions of observational studies, a “control” is a variable that is introduced *statistically* (as opposed to experimentally) into the analysis with the goal of removing its effect on the relationships among two or more other variables. While the meaning of “control” might appear to be parallel in research based on experimental and observational data, in fact it is not. Statistical control with observational data is concerned with eliminating one or more rival explanations.⁴ By contrast, in experiments, rival explanations are eliminated not merely through the fact of having experimental and control groups, but rather through **random assignment** to these groups. Thus, in experiments, randomization is the equivalent of perfect statistical control in observational studies.

Further, in observational studies, “to control” for a variable (as a verb) means to statistically remove its effect from the relationship among two or more other variables.

controlled comparison, method of. Small-N analysis based on the careful matching of cases on selected variables. Depending on the variables selected for matching, this method may correspond either to the **method of agreement** or to the **method of difference**. This usage of “control” is related to, but different from, the ideas of experimental and statistical control.

correlation. A measure of the association between two or more variables.

counterfactual analysis. Reasoning about phenomena that did not occur. In causal assessment, this involves considering how outcomes would have changed if a prior event had not occurred, or had occurred in a different way. Also called a **thought experiment**.

counterfactual definition of causation. An influential understanding of **causation** as the difference between what actually happened and what

4. It is sometimes believed that adding more control variables always improves inference. In fact, the addition of a particular control may improve inference, may not affect it, or may make it worse. The key issue is whether adding the control brings the analysis closer to meeting the specification assumption.

would have happened if some prior circumstance(s) had been different in a particular way. Thus, the **causal effect** of a given explanatory factor on a particular outcome for a specific case at one point in time is defined on the basis of a comparison between the observed outcome and the hypothetical outcome that would have occurred in the same case at the same point in time if the explanatory factor had not been present. See **counterfactual analysis**; **causal inference**, **fundamental problem of**.

covariance structure models. Statistical models that explicitly incorporate assumptions about measurement and about causation. When applied to empirical data, these models yield inferences about unobserved parameters involving both the measurement relationships between observed variables and latent variables, and also the causal relationships among unmeasured **latent variables**.

These models combine aspects of regression analysis, factor analysis, and measurement modeling. Also called LISREL-type models, MIMC (multiple-indicator, multiple-causes) models, and structural equation models with latent variables.

critical juncture. A specific historical period in which particular political choices, or the emergence of a particular historical alternative, strongly dispose a given case to follow one path of change, and not others. The critical juncture can alternatively be viewed as involving a high degree of agency, or strong structural determinism. See **path dependence**.

cross-case analysis. The systematic comparison of **cases**. In discussions of small-N, case-study research, this term usefully points to the contrast vis-à-vis **within-case analysis**. Quantitative researchers would routinely assume that they do cross-case analysis. Cross-case analysis in both qualitative and quantitative research typically involves **data-set observations**.

cross-sectional analysis. Research that focuses on multiple cases at one point in time. Contrast with **longitudinal analysis**.

crucial case. A case that is seen as especially likely to make a valuable contribution to causal inference. For example, crucial cases may be strongly expected to confirm (**most-likely case**) or reject (**least-likely case**) a prior hypothesis. New causal insight may result if these expectations are *not* met. These ideas were developed by Harry Eckstein (1975).

data. Information collected by a researcher. In particular, data is typically information organized for analysis and used as a basis for inference. See **experimental data**, **observational data**.

data mining. In data analysis, the practice of trying out many different explanatory variables without theoretical justification, in the hope of finding one that explains an outcome. This is also called “data dredging,” “data snooping,” and “ransacking.” These terms often convey a negative evaluation of inductive research practices; the econometric term **specification search** is a more neutral label.

data, piece of. The value of a variable for a given case. Also called a datum or a score, and sometimes informally called an observation. See **data-set observation, causal-process observation.**

data point. In a two-dimensional scatterplot, the point that corresponds to the scores of the two variables for a particular case. A data point is an observation whose meaning crucially depends on simultaneously considering the scores for both the independent and the dependent variable. A data point can also be located in a multidimensional scatterplot, in which instance it corresponds to the scores for several variables.⁵ See **data-set observation.**

data set. A collection of scores for one or more variables across a given set of cases. Also called a **rectangular data set.**

data-set observation. All the scores in a given row, in the framework of a **rectangular data set.** It is thus the collection of scores for a given case on the dependent variable *and* all the independent variables. This includes **intervening** and **antecedent variables.** Put another way, it is “all the numbers for one case.”⁶ A data point in a two- or multidimensional scatterplot is a data-set observation.

Although this definition is presented in the language of quantitative research, it is fully as useful for qualitative researchers as for quantitative researchers. A piece of data that begins as an isolated **causal-process observation** can subsequently be incorporated into a rectangular data set. Thus, through the collection of additional data, it can become part of a data-set observation.

deductive analysis. In empirical social science, the use of theories and hypotheses to make empirical predictions, which are then routinely tested against data.

5. The term data point is also sometimes used informally to mean the score for a given variable on a given case.

6. For a nominal or ordinal variable, it is all the scores on the relevant categories for each case.

degrees of freedom. A basic tool in quantitative analysis, used in establishing whether an analyst has sufficient information to make a given inference. Usually, it is the number of independent observations used in making a causal inference, minus the number of **parameters** in the model being estimated. Thus, the greater the number of **data-set observations** vis-à-vis the number of parameters—of which there is usually one per explanatory variable—the greater the degrees of freedom, other things being equal.

Degrees of freedom is not a property of the causal model by itself, or of the data set by itself, but rather of the causal model in relation to the data set. In estimating more complex models that may include both causal and measurement components, degrees of freedom may also refer to the number of variances and covariances among observed variables in relation to the number of parameters being estimated. Increasing the degrees of freedom is generally seen as desirable and is a rationale for arguments in favor of increasing the N , because **inferential leverage** will be greater. See **determinate research design, identifiability**.

dependent variable. What the researcher seeks to explain. It is hypothesized to be caused by, or “dependent” on, one or more independent variables. It is also called an outcome variable.

description. A statement about what has occurred. Description differs from **explanation**, which in a commonsense understanding is concerned with why something occurred. The relationship between description and explanation is complex, yet this distinction remains fundamental in political and social research.

descriptive inference. The process of reaching descriptive conclusions on the basis of observed data. This may involve using what is inevitably partial or imperfect information about the real world to make inferences about a concept, or it may involve using such information to characterize a broader set of cases. We find KKV’s distinction between descriptive inference and **causal inference** to be valuable, and we follow it in the present volume.

In standard statistical usage, related terms are assigned somewhat different meanings. “Descriptive statistics” is concerned with numerically or graphically summarizing a data set. “Inferential statistics,” by contrast, is concerned with reaching conclusions about a larger population on the basis of a sample, or with estimating parameters in a model. Tests of statistical significance would be considered part of inferential statistics. Both descriptive statistics and inferential statistics are sets of tools that can

contribute to the goals of descriptive and **causal inference**, as conceptualized in the present volume.

determinate research design. A design with a sufficient number of **data-set observations** to estimate each **parameter** of interest, and to avoid situations of perfect **multicollinearity**. This is a key concept in KKV (116, 118–24). Chapter 12 in the present volume recommends the alternative concept of **interpretable**. Contrast with **indeterminate research design**. See **degrees of freedom**.

deterministic. A **measurement model** or a **causal model** that contains no random elements and is not **probabilistic**. In the case of a causal model, it posits an invariant relationship between cause and effect.

In common statistical usage, a deterministic model is, by assumption, deliberately designed without a random component or an **error term**. In the vocabulary of qualitative methodologists, by contrast, “deterministic causation” often refers to models of necessary and/or sufficient causation, which represent a subset of the causal models that are deterministic according to the statistical definition. Contrast with **stochastic**.

deterministic cause. See **deterministic**.

deviant case. A case that is an **outlier** with respect to a given empirical relationship. In standard regression analysis, a deviant case is a case with an exceptionally large value for the residual. Analysis of a deviant case may lead researchers to reconceptualize concepts, revise indicators, or rethink causal hypotheses.

dichotomy. A categorical variable that classifies cases into two groups. A dichotomy may be measured on a **nominal scale** (male/female) or on an **ordinal scale** (rich/poor). Behind dichotomies, of course, one routinely finds finer differentiation that would be associated with higher **levels of measurement**.

diffusion. A form of causation in which the value of a given variable in one case influences the value of that same variable in other cases. Diffusion can be a methodological problem in that the assumption of **independence of observations** may be violated. Diffusion is also treated as a substantive topic in its own right.

disturbance term. See **error term**.

doubly decisive test. One of the tests employed in **process tracing**. Confirms one hypothesis and eliminates others. It provides a necessary and

sufficient criterion for accepting a hypothesis. In the social sciences such tests are rare, yet a hoop test and a smoking gun test, used together, accomplish the same analytic goal.

econometrics. The methodological subfield within the discipline of economics, which has contributed major refinements to **regression analysis** and **time-series analysis**.

efficiency. The extent to which a given analytic procedure fully utilizes available evidence to maximize inferential leverage. The concept is used in the present volume in evaluating alternative procedures for assessing **necessary** and/or **sufficient causes**. In statistical usage, this term specifically refers to an estimator whose sampling distribution has a smaller variance than another estimator, or a test that has greater inferential power than another test.

elaboration model. Procedures for data analysis and causal inference that build up larger models from bivariate relationships by successively introducing control variables. The terms **intervening variable** and **antecedent variable** are identified with this approach, which is strongly associated with the work of the sociologist Paul Lazarsfeld. Compare with **stratification**.

empirical. Based on observation and evidence.

endogeneity. A problem that arises when one or more **endogenous variables** in a given causal model are treated as **exogenous**.

Endogeneity occurs when a researcher tests a causal model in which one of the explanatory variables is correlated with the **error term**. Specific examples of endogeneity include **missing variable bias** and reciprocal causation. If a variable in a causal model is endogenous and the analyst does not adopt an appropriate technique to correct this problem, the resulting causal inferences are invalid. Endogeneity is a failure to meet the **specification assumption**.

endogenous variable. A variable caused by other variables within a given causal model; or, a variable correlated with the **error term** (i.e., it could be caused by a missing variable). A variable that is *not* caused by other variables in the model is called an **exogenous variable**.

error. A discrepancy between the estimated value of a parameter and its "true" value; alternatively, in causal inference, a discrepancy between the

predicted and observed values of a given case on a given dependent variable. Error may be due to systematic mistakes in data collection or analysis, or to random factors. See **bias**, **random error**, **systematic error**, **uncertainty**.

error term. In a regression model, an unobserved variable that consists of the differences between the observed values of the dependent variable for each case and the theoretically expected values, given the scores on a set of independent variables.

The residuals in a regression equation, which consist of the difference between the observed values of the dependent variable and its *estimated* expected value, may be used to estimate the error term, but they are not themselves the error term. The difference between the true error term and these errors of prediction (which is sometimes called the disturbance term) in any particular regression analysis may be due to an incorrectly specified model, measurement error, or random factors. The variance of the residuals is a good estimate of the variance of the error term only if certain assumptions are met: for example the **specification assumption**, **causal homogeneity**, and the assumption that errors across cases are independent and identically distributed (IID).

estimation. The process of finding the most appropriate value for a **parameter** in a given model, based on the analysis of data. Estimation may be carried out using a statistical technique or a qualitative tool of descriptive or causal inference.

estimator. A procedure or formula used to find the most appropriate value for a **parameter** in a given statistical model, using the evidence provided by a particular set of cases. Formulas for calculating means, correlations, and slopes are estimators.

ethnographic research. Analysis based on sustained, direct observation of and interaction with the individuals or groups being studied, often involving participant observation.

exogenous variable. A variable not caused by other variables within a given causal model and not correlated with the error term. Whereas the pairing of exogenous and **endogenous** is fairly straightforward, the relationship of these terms to independent variable requires clarification. A strict understanding of "independent" could lead to the conclusion that an independent variable is necessarily an exogenous variable. However, the expression independent variable is commonly used more broadly for any explanatory variable, exogenous or endogenous.

expected value. The mean value of the theoretical sampling distribution of any statistic. Statistical reasoning is centrally concerned with the expected value, as opposed to any particular observed value. In statistical procedures that seek to predict the values of a dependent variable using one or more independent variables, the predictions are typically estimated expected values, conditional on the independent variables included in the analysis.

experiment. Research in which the investigator introduces a treatment or stimulus in order to evaluate its causal effect. Compared to an **observational study**, an experiment far more effectively eliminates rival explanations.

In general, the treatment is applied to one set of cases, but not to a **control** group, and the effects are then evaluated. In more complex research designs, more than two groups, with more than two levels of the treatment, may be employed. Assignment to the groups should be random, in order to isolate the causal effect of the treatment from the effects of other potential causes. See **natural experiment**, **quasi-experiment**.

experimental data. Data generated using a research design in which the investigator assigns particular values on one or more independent variables to the cases being studied. Contrast with **observational data**.

explanation. A statement about why an outcome has occurred. A given variable may be called an explanation, but the term is also applied to the larger framework of causal understanding within which a particular independent (i.e., explanatory) variable or variables are located.

Explanation differs from **description**, which in a commonsense understanding is concerned with what has occurred. The relationship between explanation and description is complex, yet this distinction remains fundamental in political and social research.

explanatory variable. See independent variable.

ex post facto hypothesis formation. Formation of new hypotheses after examination of the data.

Whereas in some traditions of research this is seen as a mistake, many qualitative researchers view the **iterated refinement of hypotheses** in light of the data to be essential. Within the quantitative tradition, the term **data mining** implies an inappropriate search for statistically significant relationships within a given data set, whereas **specification search** is intended to refer to a disciplined approach to this task.

extension. The range of cases to which a concept applies. This idea is particularly relevant to dichotomous concepts, for which the idea of empirical membership or nonmembership in the category suggested by the concept is especially meaningful. See **intension**.

external validity. The degree to which descriptive or causal inferences for a given set of cases can be generalized to other cases. It is also called generalizability. Contrast with **internal validity**.

falsifiable. The potential of a claim, **hypothesis**, or theory to be proven wrong.

field research. The collection of data from a real-life setting, as opposed to a library or laboratory. It commonly involves direct observation of, and sometimes interaction with, the political and social actors being studied. Collecting data through archival research would often be considered an aspect of field research. Field experiments, including experiments embedded within public opinion surveys, are a special type of field research that utilizes experimental intervention by the investigator.

goals. Objectives in the conduct of research. See **overarching** and **intermediate goals**.

goals, intermediate. Methodological norms for the application of research tools in pursuit of **overarching goals**. In carrying out description, intermediate goals include precision, reliability, and sensitivity to context. In causal assessment, alternative intermediate goals include generality, parsimony, and accuracy.

The pursuit of intermediate goals raises the issue of **trade-offs**, which may lead scholars to embrace some intermediate goals and reject others. In promoting the idea of **shared standards** as a basic theme in the present volume, our purpose is to encourage recognition that these varied choices at the level of intermediate goals may constitute legitimate, alternative means of achieving the overarching goals.

goals, overarching. Broad, shared goals that motivate diverse research practices. In the framework of the present volume, the overarching goals are to (1) strive for valid descriptive and causal inference, and (2) refine theory in the effort to improve these inferences and to strengthen our understanding of political and social reality.

Overarching goals are central to the idea of **shared standards** for evaluating research. We do not intend these goals to be construed narrowly, and

some scholars may use a different vocabulary in discussing these goals. For example, Ragin (chap. 3, online) suggests that “inference” can also be understood as “making sense of cases.” Of course, scholars make different choices about how they pursue overarching goals, and such choices are usefully understood at the level of **intermediate goals**.

guidelines. Norms for the conduct of research. The guidelines in chapter 2 of the present volume summarize KKV’s methodological advice.

hermeneutics. The epistemology and methodology of **interpretation**.

heteroskedasticity. The situation in which the error term in a regression model does not have a constant variance across all observations, conditional on the explanatory variables.

hoop test. One of the tests employed in **process tracing**. A hypothesis must “jump through the hoop” merely to remain under consideration, but success in passing a hoop test does not strongly affirm a hypothesis. It thus provides a necessary but not sufficient criterion for accepting the hypothesis.

hypothesis. A tentative answer to a **research question**. In causal analysis, a hypothesis is a conjecture about the relationship between one or more independent variables and a dependent variable. Typically, a hypothesis is connected to a larger conceptual framework/theory.

identifiability. A characteristic of a statistical model, in relation to a particular data set, that makes it possible to estimate the parameters.

A parameter is identifiable if different values for the parameter produce different distributions for some observable aspect of the data. In a regression model, two variables are not separately identifiable if there is **perfect multicollinearity** between them. A model is likewise not identifiable with too few **degrees of freedom**. The issue of identifiability is sometimes referred to as the **identification problem**. See **determinate research design**.

identification. The process of demonstrating that the researcher has sufficient information (typically involving the number of **data-set observations**) to produce estimates of the parameters in a given causal model.

identification problem. The dilemma that, in general, the researcher does *not* have sufficient information to fully identify a model without making

restrictive assumptions about some of the relationships among variables in the model. See **identification**.

independence of observations. The assumption that for each observation, a given outcome occurs independently (conditional on the included explanatory variables) from its occurrence or nonoccurrence in other observations.

To the extent that outcomes do not occur independently, for example, due to **diffusion** across observations, each new observation provides less new information for the purpose of causal inference. Interdependence among observations does not bias the causal inference, but it does bias tests of significance that depend on the N , in that such tests tend to overestimate the amount of new information provided by each observation. The issue of independence of observations is a completely different matter from the question of **conditional independence**.

For some readers, a familiar alternative label for this assumption, which is appropriate for discussing cross-sectional analysis, is “independence of cases.” However, this same assumption plays a major role in **time-series analysis**, in which the researcher analyzes multiple observations over time for each case. Hence, the broader idea of independence of these observations becomes a central issue, and it is useful to employ this more general label.

independent variable. A variable that influences, or is hypothesized to influence, another variable. This other variable is called the dependent, or outcome, variable. The term **explanatory variable** is often used interchangeably with independent variable.

Although “independent” might be understood to give this term the same meaning as **exogenous variable**, the term “independent variable” is routinely used more broadly to refer to all the explanatory variables in a model. Thus, in quantitative analysis, all the right-hand side variables in a regression equation are independent variables, including **intervening** and **antecedent variables**.

indeterminate research design. A design that lacks a sufficient number of **data-set observations** in relation to the number of **parameters** to be estimated, and/or may suffer from perfect **multicollinearity**. This is a key concept in KKV (118–24).

Within the framework of standard statistical techniques, an indeterminate research design can leave the analyst with insufficient information to adjudicate among rival explanations. However, these problems can

sometimes be overcome through techniques such as the analysis of **causal-process observations**. Contrast with **determinate research design**. See **degrees of freedom**, **interpretable**.

indicator. A procedure for measuring or **operationalizing** a concept. It may be a quantitative procedure that generates numerical scores, or an operational definition employed in qualitative research to classify cases.

inductive analysis. A method that employs data about specific cases to reach more general conclusions. Contrast with **deductive analysis**.

inference. The process of using data to draw broader conclusions about concepts and hypotheses that are the focus of research.

This definition is specifically intended for the present discussion of empirical research; in other contexts, including mathematics, formal logic, and game theory, scholars are concerned with logical inferences, rather than with inferences from data. Descriptive inference employs data to reach conclusions about what happened; causal inference employs data to reach conclusions about why it happened. See **nested inference**.

inferential leverage. The capacity to make valid inferences, given a particular measurement model or causal model and a specific data set. Some methodological tools serve to increase inferential leverage.

instrumental-variables (IV) design. A specific kind of **natural experiment**. Addresses the challenge of inferring the causal impact of a given independent variable by finding an additional variable—called an instrument—that is correlated with the independent variable but could not be influenced by the dependent variable and is not correlated with its other causes. The analysis builds on the assumption that the instrumental variable in effect “assigns” cases to values of the independent variable in a way that is *as-if* random vis-à-vis all potential confounders, even though no actual randomization typically occurs. In instrumental-variables analysis, the predicted values of the independent variable based on the instrument are used in place of the original independent variable. (In the second edition, this design is discussed by Seawright and Dunning, chaps. 13 and 14). See **regression-discontinuity design**, **matching design**.

intension. The core meaning or defining attributes of a concept. See **extension**.

interaction term. An element in a regression equation that reflects the joint, multiplicative effect of two or more independent variables on the depen-

dent variable. With an interaction term, the influence of each independent variable depends in part on the value of the other independent variable.

intermediate goals. See **goals, intermediate.**

internal validity. The degree to which descriptive or causal inferences from a given set of cases are correct for those cases. Contrast with **external validity.**

interpretable. A characterization of findings or inferences that can plausibly be defended. The interpretability of findings or inferences can be increased by many factors, including a large N, an ingenious comparative design, a rich knowledge of cases and context, well-executed conceptualization and measurement, and an insightful theoretical model.

The present volume recommends this concept as an alternative to KKV's idea of a **determinate research design.** This usage of the term interpretable involves different issues from the tradition of **interpretation.**

interpretation. A description or characterization of the meaning of human behavior from the standpoint of the individuals whose behavior is being observed. It is sometimes used interchangeably with **thick description** (following Geertz) and *Verstehen* (following Weber). See **constructivism.**

interpretivism. See **interpretation.**

interrupted time-series design. An **observational study** in which the researcher examines time-series data before and after a major event (for instance, a policy switch) that is hypothesized to affect the dependent variable. In some cases, this major event may be the principal explanatory variable; in other cases, it may be one of several explanatory variables. See **quasi-experiment.**

interval scale. See level of measurement.

intervening variable. A variable that stands causally between a given explanatory variable and the outcome being explained. The status of being an intervening variable should be understood in relation to a particular causal model. An **antecedent variable** (also called a background variable) stands prior to an intervening variable.

iterated refinement of hypotheses. Movement back and forth between hypotheses and data to refine hypotheses and take advantage of new

insights that can be gained from the data. See **data mining**, *ex post facto* hypothesis formation, **specification search**.

large N. A large number of cases. Contrast with **small N**.⁷ There is no well-established cut-point between a large and a small N, but it might be located somewhere between ten and twenty cases.

latent variable. An attribute or characteristic observed through indicators that measure it indirectly.

least-likely case. A case that is not expected to conform to the prediction of a particular theory.

A least-likely case often has extreme values on variables associated with rival hypotheses, such that we might expect these other variables to negate the causal effect predicted by the theory. If the case nonetheless conforms to the theory, this provides evidence against these rival hypotheses and, therefore, strong support for the theory. This contrasts with a **most-likely case**, which is strongly expected to conform to the prediction of the theory. See **critical case**.

level of analysis. The level of aggregation on which a given study is focused. This should be understood within the framework of a hierarchy of levels. Examples of levels in such a hierarchy are individual actors, subnational units (cities, states, or provinces), national organizations (nation-states, or components of nation-states such as national legislatures or national political regimes), and the international system (relations among nations and international institutions).

At any given level of analysis, research may focus on different **units of observation**. For example, at the level of contemporary nation-states, it can focus on individuals (e.g., on top decision makers within the state), on characteristics of national institutions, or on aggregated features of the national population.

level of measurement. The generic label for the logical relations entailed in nominal, ordinal, interval, and ratio scales (as well as various other scale types). Different types of scales constitute successive levels of measurement, in that they sequentially incorporate into the scale (in the case of the four types just noted) the ideas of equal/nonequal, order, unit of measurement, and a mathematically meaningful zero. See also **typology**.

7. Some confusion arises because large N and small N are hyphenated when they serve as a compound adjective, as in "large-N (or large-N) research"; but are not hyphenated when used as a noun, as in "they focused on a small N (or a large N)."

According to one major approach to measurement theory, measurement must ultimately be understood in terms of pairwise comparison among specific cases. Thus, a given level of measurement (or particular scale type) is based on: (1) a set of logical relations among cases located within a specified domain, logical relations which, in principle, must ultimately be validated by pairwise comparison of cases; and (2) the claim that these logical relations can validly be employed to compare those cases with respect to a given variable.

LISREL. A computer program (acronym for “Linear Structural Relations”) that estimates causal models which explicitly incorporate the researcher’s assumptions about measurement relations and causal relations. The more generic label is LISREL-type models or **covariance structure models**.

longitudinal analysis. Analysis of change over time, focused on one or more variables or cases. It is also called **time-series analysis**. See **cross-sectional analysis**.

mainstream quantitative methods. An approach to methodology strongly oriented toward regression analysis, econometric refinements on regression, and the search for statistical alternatives to regression models in contexts where specific regression assumptions are not met. (In the present volume, the Introduction to the Second Edition and chapter 13 by Seawright explore the wide-ranging criticism of mainstream quantitative methods that has emerged in recent years.)

matching cases. Cases that all have the same score on a particular dichotomous variable, or that all have similar scores on a continuous variable. See **contrasting cases**.

matching design. Like conventional regression analysis of observational data, a strategy of controlling statistically for known confounders. In the standard set-up of a dichotomous independent variable, assignment of cases to one category of that variable (i.e., to “treatment” or “control”) is not *as-if* random. Rather, cases are matched in pairs that are as similar as possible on observable confounders. Given that statistical models are routinely used in carrying out the matching, i.e., creating the pairs, this method faces many problems of conventional regression analysis and essentially does not escape the shortcomings of that method. (In the second edition, this design is discussed by Seawright and Dunning, chaps. 13 and 14). Contrast with **natural experiment**, **regression-discontinuity design**, **instrumental-variables design**.

measurement. The process of making empirical observations in relation to a given concept. This includes, in addition to quantitative measurement, the scoring of cases carried out by qualitative researchers on the basis of categorical variables. An **indicator** is a specific procedure for measurement.

measurement error. Failure to perfectly operationalize a concept, due to the use of indicators that lack **reliability** and/or **validity**. See **measurement**.

measurement model. A set of understandings or hypotheses concerning the relationship between one or more **concepts** and one or more **indicators** of those concepts. This relationship may or may not be formalized mathematically.

measurement theory. A body of literature, associated with **psychometrics** and mathematical measurement theory, which has developed logical foundations and empirical tools for measurement.

measurement validity. The extent to which the scores produced by a given measurement procedure meaningfully reflect the concept being measured.

One view is that measurement validity is concerned with nonrandom error (or **bias**), and that **reliability**, which concerns **random error**, is a separate issue. However, according to other definitions, reliability is a requisite for **validity**. Measurement validity is an issue in both quantitative measurement and qualitative classification.

method of agreement. A research design that compares cases which are matched (i.e., in *agreement*) on one of the main variables of concern (either an independent or a dependent variable), and which *differ* on other variables understood to be potential causes or effects of that variable. However, in current usage, this label is generally employed more specifically for designs in which cases are matched on the dependent variable and differ from one another on many explanatory variables. The method was proposed by J. S. Mill.⁸ Contrast with **method of difference**.

method of difference. A research design that compares cases which *differ* on one of the main variables of concern (either an independent or a

8. Mill (1974[1843]). It is well known that in Mill's view, the methods of agreement and difference are not applicable in the social sciences, yet they remain an important point of reference in social science methodology.

dependent variable), but that are *similar* on other variables understood to be potential causes or effects of that variable. However, in current usage, this label is generally employed more specifically for designs in which cases differ on the dependent variable and are matched (i.e., in agreement) on many explanatory variables. This method was proposed by J. S. Mill. Contrast with **method of agreement**.

The expression “most similar systems design,” introduced by Przeworski and Teune (1970), refers to essentially this same research design. With Przeworski and Teune’s label, the term *similar* refers to the matching of cases with respect to alternative explanations. With both approaches, the key step in causal inference is to find, along with the many explanatory variables on which the cases are matched, one on which they differ—which is thus congruent with the difference on the dependent variable.⁹ This congruence is then used as the basis for a causal inference.

Mill’s methods of agreement and difference. See **method of agreement**, **method of difference**.

missing variable. A theoretically relevant variable that, if added to a causal model, would change estimates of the effects of other explanatory variables. A model with no missing variables in this sense still may not explain all the variance in the dependent variable. Rather, other things being equal, the expected values of the causal estimates for a model with no missing variables will be nearly correct. Also called an **omitted variable** or **confounder**.

missing variable bias. Bias introduced in causal inference when a theoretically relevant explanatory variable is missing. As a consequence of miss-

9. In characterizing the most similar systems design, Przeworski and Teune state that “common systemic characteristics are conceived as ‘controlled for,’ whereas intersystemic differences are viewed as explanatory variables. The number of common characteristics sought is maximal and the number of not shared characteristics sought, minimal” (1970: 33). However, they go on to point out that “although the number of differences among similar countries is limited, it will almost invariably be sufficiently large to ‘overdetermine’ the dependent phenomenon” (34); they then characterize this design as based on “concomitant variation,” which is in fact another one of Mill’s methods. By contrast, Przeworski and Teune’s “most different systems” design (1970: chap. 2) begins with the cross-national analysis of individual-level data. If the researcher discovers that individual-level patterns are not homogeneous across national units, then the focus shifts to analyzing the differences among the national units (1970: 34–35). Thus, it is in fact not parallel to Mill’s method of agreement.

ing variable bias, the causal estimate for any given variable that is included may be too large, in which case the causal effect attributed to the included variable is at least partially spurious. Alternatively, the estimate may be too small, in which case the missing variable is a suppressor variable; or the estimate may have the wrong sign, in which case the missing variable is a distorter variable. See **missing variable**.

model. A framework of concepts, descriptive claims, and causal hypotheses, through which the analyst seeks to abstract understanding and knowledge from the complexities of the real world. A model is often seen as a more systematized version of a **theory**. See **causal model**, **measurement model**.

model, causal. See **causal model**.

most-likely case. A case that is strongly expected to conform to the prediction of a particular theory. If the case does not meet this expectation, there is a basis for revising or rejecting the theory. This contrasts with a **least-likely case**, which is strongly expected not to conform to the prediction of the theory. See **critical case**.

multicollinearity. A problem of statistical estimation and inference, in which high correlations among independent variables make it difficult to separate, and hence to estimate, their individual effects. Sometimes also called collinearity.

This problem is related to the issue of **degrees of freedom**, in that the larger the number of independent cases in relation to the number of parameters to be estimated, the easier it is to deal with multicollinearity. With perfect multicollinearity, there is a perfect linear relationship among two or more independent variables, and the coefficients cannot be separately estimated. See **determinate** and **indeterminate research design, identification**.

multi-method research. A study that combines two or more research tools. **Snow on cholera** juxtaposes qualitative analysis with a natural experiment. Alternatively, a scholar might bring together small-N comparative analysis and large-N regression analysis.

multiple conjunctural causation. See **causation, multiple and conjunctural**.

N. The number of cases in a given study. The N also corresponds to the number of rows in a **rectangular data set**, that is, to the number of **data-set observations**.

natural experiment. A research design based on observational data, in which *as-if* random assignment to the categories of an (often dichotomous) independent variable (i.e., to “treatment” and “control”) occurs as a natural result of unfolding social and political processes. The assumption that assignment is *as-if* random is the basis for presuming that causal inferences are not distorted by confounders. Given that the challenge is to discover situations in which such apparent *as-if* random assignment occurs, scholars often refer not to “creating” a natural experiment, but to “exploiting” a real-world opportunity for this kind of design. (In the second edition, an extended discussion and evaluation is provided by Dunning, chap. 14). See **regression-discontinuity design**, **instrumental-variables design**, **Snow on cholera**.

necessary cause. A cause whose presence is required for the outcome to occur. Correspondingly, its absence definitively prevents the outcome. It is also called a necessary condition. See **sufficient cause**.

negative cases. Theoretically or substantively relevant cases in which an outcome of concern does not occur. This label is sometimes used more broadly with a nondichotomous dependent variable in referring to cases in which, to a substantial degree, the outcome does not occur. See **contrast space**, **positive cases**.

nested inference. A causal inference that draws on both **data-set observations** and **causal-process observations**, sometimes at different levels of analysis. Such inference takes advantage of the distinctive contribution offered by each type of observation.¹⁰ See **triangulation**.

Neyman-Rubin-Holland model. A counterfactual theory of causation. According to this view, we cannot observe causation directly, but must make inferences about it in other ways, ideally with randomized experiments. Alternatively, and much more problematically, researchers may address causation in **observational studies**, using statistical tests and other analytic tools that approximate the procedures followed in experiments. (In the second edition, in Dunning’s extended treatment of natural experiments in chapter 14, at many points in the discussion the corresponding argument from the standpoint of the Neyman-Rubin-Holland model is presented in an endnote.)

According to the Neyman-Rubin-Holland model, the idea that “X causes Y” in any given unit of analysis raises the hypothetical question of how

10. This term is adapted from Coppedge (2001) and Lieberman (2003a).

the outcome on Y would have differed if X had been prevented from occurring in that unit. Given that it is impossible to observe both the occurrence and nonoccurrence of X for any given unit at one point in time, causal inference in effect involves comparing something that did occur with something that did not occur. This is the source of the **fundamental problem of causal inference**. While this is sometimes called the Rubin-Holland model, the central influence of Neyman makes it more appropriate to designate this as the Neyman-Rubin-Holland model (see, for example, Neyman 1923 [1990]; Rubin 1990).

Neyman, Rubin, and Holland embrace a “hypothetical manipulationist” view of causation, closely identified with the experimental tradition, in which a given factor can only be viewed as a potential cause if it can in principle be subjected to experimental manipulation. While respecting this view, and adopting other important components of the Neyman-Rubin-Holland framework, both KKV and the present volume see the strict hypothetical manipulationist position as sometimes being too limiting for the social sciences.

nominal scale. See **level of measurement**.

nonconforming cases. See **deviant cases**.

no-variance design. A research design with no variance (or little variance) on the main dependent variable. See **method of agreement**.

null hypothesis. A hypothesis against which the main hypothesis is tested. It is often, but not always, the hypothesis that there is no relationship.

observable implications. Empirical observations suggested by a given hypothesis. To the extent that such observations are found, this is routinely treated as evidence in support of the hypothesis.

observation. Information about the world that is collected in a given study. See **causal-process observation**, **data-set observation**.

observational data. Data in which the values of all variables are produced by real-world events and processes not subject to the direct control of the investigator. Contrast with **experimental data**.

observational study. A study based on **observational data**, in which the values of all variables are produced by real-world events and processes

not subject to the direct control of the investigator. Contrast with **experimental data**.¹¹

omitted variable. See **missing variable**.

omitted variable bias. See **missing variable bias**.

operationalization. The process of using **indicators** to measure concepts.

ordinal scale. See **level of measurement**.

outcome variable. The phenomenon that the researcher seeks to explain. It is hypothesized to be caused by one or more other variables. The term outcome variable is often used interchangeably with dependent variable. Independent variable (or explanatory variable) is the standard label for the hypothesized cause.

outlier. A **deviant case** in the relationship among two or more variables. It is sometimes also used to mean an extreme value on a given variable.

overarching goals. See **goals, overarching**.

parameter. A characteristic of a **causal model** that the researcher seeks to estimate. In **regression analysis**, the parameters that usually receive the most attention are the coefficients associated with each of the independent variables. Another major usage of the term parameter is to identify any feature of a population that the researcher seeks to estimate on the basis of a sample statistic.

parameter estimation. The use of available data to make inferences about a given characteristic or trait. In a typical regression analysis, parameter estimation involves finding values for the coefficients associated with each independent variable, as well as any other parameters included in the model.

Tools used in conjunction with parameter estimation allow researchers to carry out tests of statistical significance for specific parameters, as well

11. Rosenbaum (2002: 1–2) uses the term “observational study” much more narrowly. In his usage, it must involve a treatment, manipulation, or intervention that is applied to some cases and not to others. The distinction between an observational study in this sense and an experiment is simply that the experiment uses random assignment, while the observational study does not. To date, this usage has not become standard in the social sciences, and in the present volume we follow the more conventional usage.

as some tests that may help them improve or reject the model as a whole. Nevertheless, because statistical tools for parameter estimation rely on the assumption that the model is in fact correct, parameter estimation does not fully test the model.

parsimony. The use of few explanatory variables in a theory or **explanation**.

path dependence. A pattern of causation in which events or processes at one point in time strongly constrain subsequent events or processes. See **critical juncture**.

population. See universe of cases.

positive cases. Cases in which an outcome of concern does occur. This label is sometimes used more broadly with a nondichotomous dependent variable in referring to cases in which, to a substantial degree, the outcome occurs. See **contrast space**, **negative cases**.

power of a statistical test. The probability that a test will reject the **null hypothesis** when it is in fact false.

A test with greater power more effectively adjudicates between the null hypothesis and the hypothesis of interest. Increasing statistical power is one tool, although hardly the only tool, for strengthening causal inference. See **degrees of freedom**, **determinate research design**, **parameter estimation**, **significance test**.

probabilistic. Containing an element of randomness. Generally used interchangeably with **stochastic**. Contrast with **deterministic**.

probabilistic cause. A cause that makes a given outcome more likely (or less likely), but not inevitable. See **deterministic**.

probability theory. A body of mathematical theory concerned with analyzing the odds that uncertain events will occur.

process tracing. Examination of diagnostic pieces of evidence, commonly evaluated in a temporal and/or explanatory sequence, with the goal of supporting or overturning alternative causal hypotheses. These diagnostic pieces of evidence are called **causal process observations (CPOs)**, and process tracing provides criteria for evaluating their contribution to causal inference. (In the second edition, these tests are brought together

in Bennett's typology of process tracing in chapter 10).¹² See **doubly-decisive test**, **hoop test**, **smoking-gun test**, **straw in the wind test**.

psychometrics. The subfield of psychology concerned with **measurement theory** and tools for measurement.

The name of this subfield might lead some qualitative researchers in political science and sociology to conclude that its concerns are remote from their own. However, this subfield has been an area of considerable innovation in addressing the challenges of measuring difficult concepts and the idea that measurement is inherently context specific.

qualitative. See qualitative-quantitative distinction.

Qualitative Comparative Analysis (QCA). A systematization of small-N comparative analysis and **analytic induction** developed by Ragin (1987), based on Boolean algebra.

qualitative-quantitative distinction. A common heuristic distinction usefully understood in terms of four overlapping dimensions: level of measurement, size of the N, statistical tests, and thick versus thin analysis.

Although some studies are unambiguously qualitative or quantitative according to these criteria, mixed types are equally important, given the wide interest in combining tools of qualitative and quantitative analysis. However, the simple qualitative-quantitative dichotomy has productively structured much of the current debate.

- a. **Level of measurement.** Some scholars label **data** as qualitative if it is organized at a nominal level of measurement, and as quantitative if it is organized in terms of ordinal and higher levels of measurement. Alternatively, the threshold is sometimes placed between ordinal data and data that are at least at the interval level.
- b. **Size of the N.** The qualitative-quantitative distinction is sometimes identified with the contrast between small-N and large-N research, involving the number of observations analyzed by the investigator. It is certainly not meaningful to insist on a specific cut-point between these alternatives, but it might be placed somewhere between 10 and 20.
- c. **Statistical tests.** An analysis may be considered quantitative—even if it focuses on nominal scales—if it utilizes explicit statistical tests in reach-

12. The tests were originally formulated by Van Evera (1997: 31–32).

ing its descriptive and explanatory conclusions. By contrast, qualitative research employs a “verbal” style of analysis, often involving narrative treatment of the material. Adopting a verbal style of analysis does not mean that qualitative researchers work only with nominal variables; indeed, they employ variables at all levels of **measurement**. Moreover, they compare alternative indicators in the course of constructing composite measures and assessing **measurement validity**, and they may assess hypotheses through examining covariation among variables. Thus, they perform research operations that are in some respects analogous to standard statistical tests, yet they do not actually employ such tests.

- d. **Thick versus thin analysis.** Qualitative researchers are more inclined toward thick analysis that relies on detailed knowledge of specific cases. By contrast, quantitative researchers are more strongly oriented toward thin analysis, which relies on a more limited knowledge of each case and typically depends instead on a larger N for inferential leverage.¹³

quantitative. See **mainstream quantitative methods, qualitative-quantitative distinction.**

quantitative methods, mainstream. See **mainstream quantitative methods.**

quasi-experiment. An observational study that in some respects resembles an experiment. Specifically, the researcher observes one or more cases after (and often before) what may be thought of as a “treatment,” involving a change in an explanatory variable at a given point in time. This treatment can be a major policy change or some other large-scale political event, such as a revolution, or an individual choice, for example, a decision that a child will go to an integrated or segregated school. Thus, the treatment involves discrete, real-world events.

The idea of a quasi-experiment was initially popularized by Campbell, yet he later distanced himself from this design because it was too often misunderstood as overcoming the limitations of observational data (see 163 this volume). Specifically, the assumption of *as-if* random assign-

13. This distinction draws on Coppedge’s (1999) discussion of thick versus thin concepts. Neither our distinction nor that of Coppedge should be confused with Geertz’s (1973) distinction between “thick description,” which focuses on the meaning of human behavior to the actors involved, as opposed to “thin description,” which is not centrally concerned with this meaning. With the expression “thick analysis,” we mean research that focuses closely on the details of cases. These details may or may not encompass subjective meaning. In this sense, Geertz’s thick description is one tool for what we call thick analysis.

ment is not met in this design. Correspondingly, the expression quasi-experiment is most usefully understood as referring to an **interrupted time-series design**, and not as belonging in the family of techniques associated with natural experiments.

random assignment. See randomization.

random error. Error that is not attributable to any systematic relationship. Contrast with **systematic error**.

randomization. Assignment of values (e.g., treatment or control) on an independent variable to different cases according to an impartial chance procedure. See **experiment**.

random sample. A sample selected in such a manner that all cases from the relevant **universe of cases** have a known probability of being selected.

ratio scale. See **level of measurement**.

rectangular data set. An array or matrix of data in which the rows correspond to cases and the columns to variables. The variables in the columns include all dependent and independent variables. A rectangular data set may contain either quantitative or qualitative data. It is often called a **data set**.

regression analysis. An extension of correlation analysis, which makes predictions about the value of a dependent variable using data about one or more independent variables. A key **parameter** estimated in a regression analysis is the magnitude of change in the dependent variable associated with a unit change in an independent variable. This parameter is referred to as the slope or the regression coefficient. (In the present volume, the Introduction to the Second Edition and chapter 13 by Seawright explore the wide-ranging criticism of regression analysis that has emerged in recent years.)

regression-discontinuity design (RDD). A specific kind of **natural experiment**. Addresses the challenge of inferring the causal impact of a given independent variable, and is employed in situations where, as part of a social or political process, cases are assigned to a category of a dichotomous independent variable (i.e., to “treatment” or “control”) according to whether they are just above or below a given threshold. For cases near the threshold, the process that determines placement vis-a-vis the threshold is *as-if* random, ensuring that these individuals will be very similar

with respect to potential confounders. This in turn opens the possibility of a more compelling causal inference. The contrast with the standard natural experiment is that *as-if* random assignment specifically involves the position of cases in relation to this threshold.

reliability. The stability of an indicator over (potentially hypothetical) replications of the measurement procedure. Reliability involves the magnitude of **random error**. Repeated application of a reliable measure to a subject who has not changed regarding the trait being measured produces results that cluster in a narrow range. See **measurement validity**.

replication. An attempt to reproduce the findings of a given study. Two different research practices are both called replication: a narrow version, which involves reanalyzing the original data, and a broader version based on collecting and analyzing new data.

research cycle. The sequence of steps typically undertaken in research. These commonly include defining the **research problem**, **specifying the theory**, selecting cases, carrying out descriptive and causal inference, and sometimes the **iterated refinement of hypotheses**, based on movement back and forth between data and hypotheses. The later steps in this cycle routinely provide insight that may lead the researcher to revise the earlier steps, and in practice, researchers may move in many different ways among these steps.

research design. A plan for carrying out a given study, commonly involving a sequence of research steps such as those listed under **research cycle**.

research problem. See **research question**.

research program. A coordinated effort to address a given set of research questions.¹⁴ Whereas a **research design** is a plan for carrying out a specific study, a research program encompasses a number of studies and the work of many scholars.

research question. The theoretical or empirical puzzle that motivates a given study. It is also called a **research problem**.

Rubin-Holland model. See **Neyman-Rubin-Holland model**.

sample. The set of cases on which the analysis is focused, and which are often selected from a larger **universe of cases**. Selecting cases is a funda-

14. The term is thus often used in a broad sense, and not with the relatively specific meaning intended by Lakatos (1970).

mental task of **research design**, and scholars in different research traditions have approached this task in a variety of ways. See **random sample**.

sampling error. Random error in inferences from a **sample** to a **universe of cases**. This error occurs because the sample, although randomly drawn, is imperfectly representative of the universe. Sampling error is sometimes contrasted with **sampling bias**, which involves **systematic error**. Sampling error can affect the **validity** or **reliability** of descriptive and causal inference. See **sample**.

scientific. A normative view of the theoretical, methodological, and empirical goals of research.

Alternative definitions of “scientific” express different normative views. For example, KKV (8–9) presents a four-part definition of “scientific research” that is fundamental to the book’s framework: Scientific research is based on **inference**, it makes its procedures public, it views conclusions as inherently uncertain, and its findings are judged in light of the method employed. KKV’s understanding of scientific method in qualitative research is closely tied to basic ideas of mainstream quantitative methods, and correspondingly has been subject to widespread criticism. (In the present volume, see the Introduction to the Second Edition, and also chapter 13 by Seawright.)

By contrast, Freedman (chap. 11, this volume) argues that qualitative methods are a type of scientific inquiry in their own right. Still other definitions place central emphasis on the importance for science of building theory. Scientific research is thus a prominent example of a contested concept.

scope conditions. Criteria that specify the appropriate range of cases (i.e., the **universe of cases**) to which a theory applies.

score. The value assumed by a variable for a given case. This includes not only quantitative scores, but also the results of qualitative classification. A score is sometimes informally called an **observation**.

selecting on the dependent variable. Any pattern of case selection that overrepresents cases at one end of the dependent variable. That is to say, the researcher tends to select cases that consistently have higher, or lower, values on the dependent variable. The form of selecting on the dependent variable that receives most attention in the present volume is **truncation**.

Selecting on the dependent variable is routinely viewed as a source of **selection bias**. In regression analysis, truncation does indeed produce such bias. However, some modes of selecting on the dependent variable do not yield selection bias. For example, if the analyst selects on the dependent variable indirectly by choosing cases that have high scores on a key independent variable, this will yield cases with high scores on the dependent variable, but will not produce selection bias—because it does not constrain the error term.

“Selecting on the dependent variable” sometimes has an alternative meaning, in that it is used to designate the deliberate selection of cases that reflect the full range of that variable. In this instance, the mode of selection may not be correlated with the dependent variable.

selection bias. Systematic error that arises either when cases are selected according to an unrepresentative sampling rule, or when some (often unknown) nonrandom process assigns causes to cases. Such bias can result from selection procedures employed by the investigator, from self-selection of individuals or other units of analysis into the sample, or from self-selection of the cases under study into the categories of a major independent variable. In this last situation, causes may in effect be assigned to cases in a way that reinforces preexisting differences among the cases. Under any of these conditions, tests of explanatory hypotheses routinely suffer from systematic error.

The source of selection bias of primary concern in the present volume is deliberate **truncation** by the investigator, which yields bias due to the interplay among three elements. Thus, truncation on (1) the dependent variable produces selection bias by creating a **correlation** between (2) the independent variable and (3) the **error term**. This correlation yields bias because it flattens the slope of the regression line in the truncated sample. Alternative sources of selection bias are real-world political or social processes that “select” cases into the sample or into key analytic categories in ways that confound the impact of a hypothesized cause with the selection mechanism. These processes may include self-selection by the individuals being studied.

Selection bias is generally treated as an issue in regression analysis. **Within-case analysis** in the qualitative tradition, which employs different tools of causal inference, may not be subject to this form of bias.

shared standards. Commonly accepted methodological norms for the conduct of research. The **overarching goals** of valid descriptive and causal inference and of building theory are central to the idea of shared standards.

The present volume argues that scholars face basic **trade-offs** in selecting research **tools** and also in choosing **intermediate goals**. The idea of shared standards centrally involves the search for common criteria in evaluating and managing these trade-offs.

significance test. A tool for addressing the concern that an observed relationship could be due to **sampling error** or other hypothesized forms of random error. It thus provides a set of rules for deciding when empirical evidence suggests a relationship that is not simply due to chance.

In contemporary social science, significance tests are often treated much more broadly as a general-purpose test for the **validity** and **reliability** of causal inferences, a practice that extends these tests beyond the uses for which they were designed and raises serious concerns among some statisticians.

small N. A small number of cases. Contrast with **large N**.

smoking gun test. One of the tests employed in **process tracing**. Analogous to finding a murder suspect holding a smoking gun. Strongly supports a given hypothesis, but failure to pass such a test does not eliminate the hypothesis—just as the absence of a smoking gun does not exonerate a suspected murderer. It provides a sufficient but not necessary criterion for confirmation.

Snow on cholera. A classic study in epidemiology in which, in mid-19th century London, John Snow carried out **multi-method research** that astutely combined qualitative data and a **natural experiment**. In seeking tests for the hypothesis that cholera was a water-borne disease, Snow discovered an area of London where water was supplied to households by two different water companies, one providing water contaminated with London sewage and the other not contaminated. Water was distributed to specific households based on criteria that could not have been associated with confounders, and thus was *as-if* random. The striking difference in cholera rates between the two groups of households yielded strong evidence that the disease was water borne. (In the second edition, see the discussions by Freedman and Dunning, chaps. 11 and 14).

specification. The construction or revision of a **causal model**.¹⁵ Specification is the process of establishing the variables to be included, the func-

15. The process of specification is also important in noncausal statistical models, such as forecasting models.

tional form of the model, and the assumptions relevant to making inferences with the model. See **specification assumption**, **specification search**, **underspecified model**.

specification assumption. An assumption used to justify causal inferences based on observational data, that is, in the absence of a true experiment.¹⁶ If the specification assumption is met, researchers can expect to achieve estimates that are unbiased.

Two major threats to this assumption are: (1) excluding a variable that should be included in the analysis, which can produce **omitted variable bias**; and (2) including an **endogenous variable** without using an analytic technique that successfully corrects for the endogeneity, so that endogeneity bias is likely.

Meeting the specification assumption is a requirement for valid causal inference, but it is not by itself sufficient. Scholars must also know enough about the structure of the **error term** to judge the amount of independent information contributed by each observation. Further, scholars must present evidence that makes it appropriate to treat the statistical inference as causal. The clearest and most common example of how this step may be taken is found in studies that employ **natural experiments**, where evidence is used to show that variation in the hypothesized cause is due to exogenous manipulation. The specification assumption encompasses several issues of causal inference that are also addressed through the assumption of **conditional independence**.

specification search. An iterated process of fitting a model to data. The literature on specification searches has sought to develop a disciplined approach to this task that considers where such a search should start, where it should stop, and how to report the steps in between. By contrast,

16. To define the specification assumption formally, in a context where the true causal relation is $Y = X\beta + W\gamma + \epsilon$ and where the analyst wishes to estimate a regression model that posits the relationship $Y = Xb + e$, the specification assumption requires that $E(e|X) = 0$. By comparison with the true causal model, we see that $e = W\gamma + \epsilon$. Therefore, in order to meet the specification assumption, each explanatory variable in X must be statistically unrelated to W and ϵ . A variable that is statistically unrelated to W and ϵ is exogenous, whereas one that is related to any variable in W or to ϵ is endogenous. It should be clear from this discussion that the specification assumption involves many issues beyond those assessed through residual plots and other standard tools of regression diagnostics. To clarify the notation, Y , ϵ , and e are vectors with one value per case, β , γ , and b are vectors with one value per relevant variable, and W and X are matrices with one column per relevant variable and one row per case.

data mining often implies carrying out this task in an undisciplined manner that inappropriately increases the likelihood of finding a model that fits the data.

specifying the theory. Clarifying theoretical arguments to the point where they can generate specific hypotheses. This is one step in a **research cycle**.

spurious correlation. A relationship in which two or more variables are statistically related (i.e., correlated), but are not causally linked. Rather, the statistical relationship occurs because a third variable causes both of them. See **confounder**, **missing variable bias**.

standardized slope. A regression coefficient that has been adjusted to make it comparable with the coefficients for other independent variables with different ranges and variances. Thus, all variables are standardized to have a mean of zero and a variance of one. Contrast with **unstandardized slope**.

standards, shared. See **shared standards**.

statistical control. See **control**.

statistical model. A set of equations that relate observable data to underlying parameters. The values of these parameters are intended to reflect descriptive and causal patterns the real world. Constructing a statistical model entails choices about which variables to include, the posited relationships among these variables including functional form, temporal sequencing, issues of causal heterogeneity, choices about error terms, and ideas concerning counterfactual outcomes under interventions. All of these choices depend on assumptions, intuitions, and prior knowledge—including insights derived from qualitative evidence.

statistical power. See **power of statistical tests**.

statistical theory. A broad framework for reasoning about evidence and inference, employing mathematical probability theory to address tasks such as measurement, selecting estimators for causal inference, and inference from samples to populations.

The present volume devotes central attention to the distinction between important ideas drawn from statistical theory and **mainstream quantitative methods**. A well-established tradition of thinking in statistical theory, dating back to the emergence of statistics as an academic discipline,

expresses serious doubts about the applicability of the assumptions behind regression analysis and related tools to **observational data** in the social sciences.¹⁷ Correspondingly, this statistical tradition sometimes advocates techniques that allow researchers to draw more delimited inferences that depend on fewer untested assumptions about the data. By contrast, mainstream quantitative methodologists sometimes strongly advocate regression-based tools.

Statistical theory is understood here as a multidisciplinary body of work that encompasses, in addition to research by statisticians, other lines of research in **econometrics**, **psychometrics**, and **measurement theory**, as well as some methodological contributions by scholars in disciplines such as political science and sociology.

Although work in statistical theory is sometimes thought of as distinctively linked to quantitative analysis, it may also offer a rationale for some practices of qualitative investigation. For example, this statistical tradition provides part of the justification for **causal-process observations**.¹⁸

stochastic. A model or process containing an element of randomness or error. It is used interchangeably with **probabilistic**. Contrast with **deterministic**.

stratification. An approach to causal inference that controls for alternative explanations by using categorical measures of independent variables to create subgroups of the data that effectively hold these rival explanatory factors constant. Causal inferences are then made within each subgroup.

17. This statistical tradition grows out of debates among statisticians on causal inference in experiments and observational studies. It may be dated to Karl Pearson's 1896 critique of G. Udny Yule's causal assessment, based on a regression analysis of observational data, of the relation between welfare policy and poverty in Britain (Stigler 1986: 351–53, 358). For a recent statement about this tradition, see Freedman (1999).

18. The distinction between statistical theory and mainstream quantitative methods is not intended to imply that these are sharply bounded categories. Many scholars are located between these alternatives, and all work by any given scholar will not always fall in the same category. Indeed, it is likely that some statistical theorists become mainstream quantitative methodologists when they turn to applied work. Further, analytic tools that are sometimes called "quantitative tests" may also be called "statistical tests," and this choice about labeling should not be seen as reflecting a position vis-à-vis the larger distinction between mainstream quantitative methods and statistical theory.

This involves multivariate cross tabulation, and is a standard form of hypothesis testing in experiments.¹⁹ Compare with **elaboration model**.

straw in the wind test. One of the tests employed in **process tracing**. Provides useful information about “which way the wind is blowing” in favoring one hypothesis and calling others into question, but is not decisive by itself. It offers neither a necessary nor a sufficient criterion for accepting a hypothesis or, correspondingly, for rejecting it.

subtype. A concept or category derived from a broader concept, with the goal of introducing finer differentiation. Subtypes are often formed by adding an adjective to the noun that designates the original concept, as in “parliamentary democracy.”

sufficient cause. A cause whose presence inevitably produces an outcome. This is also called a sufficient condition. See **necessary cause**.

systematic error. Error whose direction and magnitude can in principle be predicted, as opposed to **random error**. With systematic error, the expected value of a given statistic is **biased**, because the errors do not cancel one another out.

term. A word that designates a **concept**. Other more specialized usages are also found in this volume, as in **error term**.

test of significance. See **significance test**.

theory. The conceptual and explanatory understandings that are an essential point of departure in conducting research, and that in turn are revised in light of research. Different analytic traditions have divergent norms about the appropriate structure and content of these understandings. A **causal model** draws on, and is part of, a theory.

19. The assumptions relevant to different tools of causal inference merit brief comment here. The conditional independence assumption, which employs the experimental tradition as a metaphor, is directly relevant to causal inference based on stratification. Other inferential tools, such as regression analysis, employ related assumptions, including the specification assumption. For many purposes, such as helping analysts focus on the potential problem of missing variable bias in causal inference, it is productive to emphasize the similarities between these two assumptions. However, the distinctive strengths of different research tools (e.g., stratification versus regression) often depend on the contrasts among the many different sets of assumptions that serve to justify these tools.

thick analysis. See **thick versus thin analysis**.

thick description. A description or characterization of the meaning of human behavior from the standpoint of the individuals whose behavior is being observed (Geertz 1973). This is not to be confused with detailed description, which may or may not be thick in this sense. This term is often used interchangeably with **interpretation** and *Verstehen*.

thick versus thin analysis. A distinction that captures different styles of research and sources of analytic leverage. Some investigators utilize thick analysis, in the sense that they have a rich knowledge of cases.²⁰ If this knowledge is utilized effectively, it can greatly strengthen descriptive and causal inference. By contrast, researchers who deal with large numbers of cases more frequently rely on thin analysis, in the sense that they depend not on detailed knowledge of cases, but rather on the inferential leverage that derives from statistical tools applied to a large N. Whereas the capacity to use statistical tests is a distinctive strength of quantitative research, the leverage gained from thick analysis is a characteristic strength of qualitative research. **Thick description**, which is concerned with interpreting meaning, should be seen as one tool of thick analysis, as defined here.

thin analysis. See **thick versus thin analysis**.

thought experiment. Reasoning about phenomena that have not been observed. See **counterfactual analysis**.

time-series analysis. Analysis focused on change over time. It is also called **longitudinal analysis**. Contrast with **cross-sectional analysis**.

tipping point. A discontinuity or inflection in a process of change over time. Thus, it is a point at which a previous trend ends and a new one begins.

tool. A specific research procedure or practice. Some tools are highly systematized and have elaborate mathematical underpinnings: **probability theory**, **regression analysis**, **significance tests**, and **covariance structure models**. Increasing the number of observations is likewise a tool that has routinely been justified on the grounds that it increases **inferential leverage**. Other tools involve practices and procedures that are not explicitly

20. This usage is adapted from Coppedge (1999). A related distinction is made by KKV (154) in contrasting the “descriptive richness” of nominal categories with the “facilitation of comparison” at higher levels of measurement.

rooted in statistics or mathematics. This second group of tools includes **within-case analysis**, **process tracing**, **triangulation**, procedures for avoiding **conceptual stretching**, qualitative **validity** assessment, and strategies for the comparison of **matching** and **contrasting cases**. Methods of data collection are also tools, such as public opinion research, focus groups, participant observation, event scoring, content analysis, archival research, the construction of unobtrusive measures, and systematic collection of secondary data. See **goals**, **trade-off**.

trade-off. Incompatibility among desired objectives.

triangulation. Research procedure that employs empirical evidence derived from more than one method or from more than one type of data. Triangulation can strengthen the **validity** of both descriptive and causal inference.²¹ See **nested inference**.

truncation. A selection process that omits cases located in some specific part of the distribution of values for a given variable. Omitting cases above or below a given value is the form of truncation of concern in the present volume.²²

The difference between truncation and “censoring” is that with truncated samples, no data are available on any of the omitted cases. By contrast, some data are available for the cases subject to censoring. See **selection bias**.

typology. A coordinated set of categories or types that establishes theoretically relevant analytic distinctions. It is often formed by cross-tabulating two or more nominal or ordinal variables, with the cells in the resulting table becoming the categories in the typology. Each category commonly has a name. A typology is usually, but not always, a nominal (or occasionally an ordinal) scale. See **level of measurement**.

uncertainty. Lack of complete knowledge.

underspecified model. A model with the problem of **missing variable bias**. More specifically, theoretically relevant variables are missing which,

21. The idea of triangulation and of multimethod triangles can be dated to Campbell and Fiske (1959: 38–39), who in turn cite the philosopher Feigl (1958) as the source of this concept.

22. This is sometimes called “outer” truncation. By contrast, “inner” truncation omits cases within a given range of values but includes cases above and below that range.

if added to the model, would change the estimates of causal effects for the already-included variables. See **specification**.

unidentifiability. See **identifiability**.

unit homogeneity. The strong assumption for causal inference that the units in an analysis are completely identical in all relevant respects except for the dependent and independent variables of interest. A somewhat weaker assumption is defined above as **causal homogeneity**. Although KKV uses the label “unit homogeneity,” its framework instead relies centrally on the idea of causal homogeneity. Hence, in discussing their arguments, we use the label “causal homogeneity.”

units of analysis. See **units of observation**.

units of observation. The individuals, institutions, entities, or objects about which data are collected. In studies based on **data-set observations**, each unit typically receives a score on each variable. This should not be confused with **level of analysis**, in that, at any given level of analysis, researchers may make different choices about units of observation. Also called cases or **units of analysis**.

universe of cases. The set of cases about which the analyst seeks to make inferences. Research may focus on a **sample** of cases from within this universe, with the goal of making inferences to the universe. Alternatively, in some studies the set of cases under analysis is the universe. Identifying a conceptually and theoretically appropriate definition of the universe is a basic task of research. Universe of cases is often used interchangeably with **population**. See **scope conditions**.

unstandardized slope. A regression coefficient that is not adjusted to account for the differing means and variances of the variables entered into the analysis. The unstandardized slope has the advantage that it is not affected by the variance of the independent variables; it has the disadvantage that the unstandardized slopes associated with different explanatory variables are typically not expressed in the same measurement units, and hence may be hard to compare. Contrast with **standardized slope**.

validity. The adequacy of descriptive and causal inference. See **external validity**, **internal validity**, **measurement validity**, **reliability**.

value. The **score** assumed by a variable for a particular case.

variable. A systematized understanding of similarities and differences among observed phenomena. Different levels of measurement reflect some of the alternative logical forms that this systematized understanding can take.

The term variable is sometimes used interchangeably with **concept** and with **indicator**. See: **antecedent, background, dependent, endogenous, exogenous, explanatory, independent, intervening, latent, missing, omitted, and outcome variable**. See also **missing variable bias, level of measurement, and thick versus thin analysis**.

variable-oriented research. Analysis that typically focuses on a large number of cases and on systematically analyzing a well-defined set of variables for these cases. This term is identified with Ragin (1987). Variable-oriented researchers may engage in fine-grained examination of cases, but their attention is centered more strongly on understanding the cases in terms of this set of variables. Contrast with **case-oriented research**.

Verstehen. A description or characterization of the meaning of human behavior from the standpoint of the individuals who are being observed. Often used interchangeably with **interpretation** and **thick description**.

within-case analysis. The internal analysis of one or a few cases. Within-case analysis takes two principal forms, the first of which is of central concern in the present volume.

The first type, especially identified with the qualitative tradition, focuses on internal evidence about patterns of causation connected with an overall outcome distinctively associated with the particular case or cases. Familiar examples include in-depth studies of macrolevel events such as wars, revolutions, and regime change, although the focus may be at other levels of analysis as well. In such within-case analysis, scholars work with only one observation on the dependent variable (e.g., war broke out, revolution was averted, or democracy collapsed). Correspondingly, new evidence is introduced, but the number of observations (i.e., the N) is not increased. The additional evidence added by such within-case analysis contributes to evaluating explanations of this single outcome on the basis of **causal-process observations**.

In the second type of within-case analysis, researchers collect observations on the dependent variable and all the independent variables for multiple (spatial or temporal) subunits of the original case. In this

instance, the number of observations (i.e., the N) increases, and this can be seen, within the framework of KKV, as an important example of increasing the number of observations as a means of gaining inferential leverage. When scholars study subunits in this way, within-case analysis in effect becomes **cross-case analysis** and focuses on **data-set observations**.

within-case control. A procedure that uses predictions about causal mechanisms to distinguish between systematic and random aspects of a given outcome within a single case. Researchers achieve within-case control by exploring causal processes to determine which aspects of a decision or an outcome were influenced by a set of hypothesized systematic variables, and which were influenced by other, idiosyncratic factors.

Bibliography

- Acemoglu, Daron, Simon Johnson, and James A. Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review* 91, no. 5 (December): 1369–1401.
- Achen, Christopher H. 1982. *Interpreting and Using Regression*. Beverly Hills, Calif.: Sage Publications.
- . 1983. "Toward Theories of Data: The State of Political Methodology." In Ada W. Finifter, ed., *Political Science: The State of the Discipline*. Washington, D.C.: American Political Science Association.
- . 1986. *The Statistical Analysis of Quasi-Experiments*. Berkeley: University of California Press.
- . 2000. "Warren Miller and the Future of Political Data Analysis." *Political Analysis* 8, no. 2: 142–46.
- . 2002. "Toward a New Political Methodology: Microfoundations and ART." *Annual Review of Political Science*, vol. 5. Palo Alto, Calif.: Annual Reviews.
- Achinstein, Peter. 1983. *The Nature of Explanation*. New York: Oxford University Press.
- Adcock, Robert, and David Collier. 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95, no. 3 (September): 529–46.
- Alesina, Alberto, Sule Özler, Nouriel Roubini, and Phillip Swagel. 1996. "Political instability and economic growth." *Journal of Economic Growth* 1: 189–211.
- Allen, William Sheridan. 1965. *The Nazi Seizure of Power: The Experience of a Single German Town, 1930–1935*. Chicago: Quadrangle Books.
- Alvarez, R. Michael, Geoffrey Garrett, and Peter Lange. 1991. "Government Partisanship, Labor Organization, and Macroeconomic Performance." *American Political Science Review* 85, no. 2 (June): 539–56.
- Andrich, David. 1988. *Rasch Models for Measurement*. Newbury Park, Calif.: Sage Publications.
- Angrist, Joshua D., and Victor Lavy. 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Student Achievement." *Quarterly Journal of Economics* 114: 533–75.

- Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.
- Ansolabehere, Stephen, James M. Snyder, Jr., and Charles Stewart III. 2000. "Old Voters, New Voters, and the Personal Vote: Using Redistricting to Measure the Incumbency Advantage." *American Journal of Political Science* 44, no. 1: 17–34.
- APSA-CP. 1996. "Replication Debate." *APSA-CP, Newsletter of the APSA Organized Section in Comparative Politics* 7, no. 1 (Winter): 5–13.
- . 2003. "Symposium on Bridging the Quantitative-Qualitative Divide." *APSA-CP, Newsletter of the APSA Organized Section in Comparative Politics* 14, no. 1 (Winter): 8–24.
- Arceneaux, Kevin, Donald Green, and Alan Gerber. 2006. "Comparing Experimental and Matching Methods Using a Large-Scale Voter Mobilization Experiment." *Political Analysis* 14: 37–62.
- Arendt, Hannah. 1958. *The Origins of Totalitarianism*. Cleveland, Ohio: World.
- Arrow, Kenneth J. 1951. *Social Choice and Individual Values*. New Haven: Yale University Press.
- Babbie, Earl R. 2004. *The Practice of Social Research*. 10th edition. Belmont, Calif.: Wadsworth.
- Bailey, Kenneth D. 1994. *Typologies and Taxonomies: An Introduction to Classification Techniques*. Beverly Hills, Calif.: Sage Publications.
- Banerjee, Abhijit, and Lakshmi Iyer. 2005. "History, Institutions, and Economic Performance: The Legacy of Colonial Land Tenure Systems in India." *The American Economic Review* 95, no. 4: 1190–1213.
- Baron, John. 1838. *The Life of Edward Jenner*. Vol. I. London: Henry Colburn.
- Barro, Robert J. 1996. "Democracy and growth." *Journal of Economic Growth* 1 (March): 1–27.
- . 1997. *Determinants of Economic Growth: A Cross-Country Empirical Study*. Cambridge: MIT Press.
- Barro, Robert J. and Jong Hua Lee. 1996. "International measures of schooling years and schooling quality." *American Economic Review* 86 (May): 218–23.
- Barro, Robert J. and Xavier Sala-i-Martin. 1995. *Economic Growth*. Cambridge: MIT Press.
- Bartels, Larry M. 1991. "Instrumental and 'Quasi-Instrumental' Variables." *American Journal of Political Science* 35, no. 3 (August): 777–800.
- . 1996. "Pooling Disparate Observations." *American Journal of Political Science* 40, no. 3 (August): 905–42.
- . 1997. "Specification Uncertainty and Model Averaging." *American Journal of Political Science* 41, no. 2 (April): 641–74.
- Bartels, Larry M., and Henry E. Brady. 1993. "The State of Quantitative Political Methodology." In Ada W. Finifter, ed., *Political Science: The State of the Discipline II*. Washington, D.C.: American Political Science Association.
- Bartels, Larry M., and John Zaller. 2001. "Presidential Vote Models: A Recount." *PS: Political Science & Politics* 34, no. 1 (March): 9–20.
- Bates, Robert H. 1981. *Markets and States in Tropical Africa: The Political Basis of Agricultural Policies*. Berkeley: University of California Press.
- . 1983. *Essays on the Political Economy of Rural Africa*. New York: Cambridge University Press.

- Baum, Matthew A., and David A. Lake. 2003. "The political economy of growth: Democracy and human capital." *American Journal of Political Science* 47, no. 2: 333–47.
- Bayles, Kenneth W. 2000. "The bactericidal action of penicillin: New clues to an unsolved mystery." *Trends in Microbiology* 8: 274–78.
- Bazin, Herve. 2000. *The Eradication of Smallpox*, translated into English by Andrew and Glenise Morgan. London: Academic Press.
- Beck, Nathaniel. 2006. "Is Causal-Process Observation an Oxymoron?" *Political Analysis* 14, no. 3 (Summer): 347–52.
- Beck, Nathaniel, and Jonathan N. Katz. 1995. "What to Do (and Not to Do) with Time-Series Cross-Section Data in Comparative Politics." *American Political Science Review* 89, no. 3 (September): 634–47.
- Beck, Nathaniel, and Jonathan N. Katz. 2006. "Top Twenty Commentaries." *American Political Science Review* 100, no. 4 (November): 676–77.
- Beck, Nathaniel, Jonathan N. Katz, R. Michael Alvarez, Geoffrey Garrett, and Peter Lange. 1993. "Government Partisanship, Labor Organization, and Macroeconomic Performance: A Corrigendum." *American Political Science Review* 87, no. 4 (December): 945–48.
- Bendix, Reinhard. 1963. "Concepts and Generalizations in Comparative Sociological Studies." *American Sociological Review* 28, no. 4 (August): 532–39.
- Bennett, Andrew. 1999. *Condemned to Repetition? The Rise, Fall, and Reprise of Soviet-Russian Military Interventionism, 1973–1996*. Cambridge, MA: MIT Press.
- . 2003. "Trust Bursting Out All Over: The Soviet Side of German Unification." In William Wohlforth, ed., *Cold War Endgame*. University Park: Pennsylvania State University Press.
- . 2005. "The Guns that Didn't Smoke: Ideas and the Soviet Non-use of Force in 1989." *Journal of Cold War Studies* 7, no. 2 (Spring): 81–109.
- Bennett, Andrew, and Colin Elman. 2006. "Qualitative Research: Recent Developments in Case Study Methods." *Annual Review of Political Science*. Palo Alto: Annual Review Publications.
- Bennett, Andrew, and Alexander L. George. 1997a. "Case Study Methods and Research on the Democratic Peace." Paper presented at the American Political Science Association Conference, Washington D.C., August.
- . 1997b. "Process Tracing in Case Study Research." Paper presented at the MacArthur Foundation Workshop on Case Study Methods, Belfer Center for Science and International Affairs (BCSIA), Harvard University, October 17–19.
- Berelson, Bernard R., Paul F. Lazarsfeld, and William N. McPhee. 1954. *Voting: A Study of Opinion Formation in a Presidential Campaign*. Chicago: University of Chicago Press.
- Berger, Daniel. 2009. "Taxes, Institutions, and Local Governance: Evidence from a Natural Experiment in Colonial Nigeria." Manuscript, Department of Politics, New York University.
- Berk, Richard A. 2004. *Regression Analysis: A Constructive Critique*. Newbury Park, Calif.: Sage Publications.
- Berk, Richard A., and David A. Freedman. 2008. "On Weighting Regressions by Propensity Scores." *Evaluation Review* 32, no. 4: 392–409.

- Bevir, Mark, and Asaf Kedar. 2008. "Concept Formation in Political Science: An Anti-Naturalist Critique of Qualitative Methodology." *Perspectives on Politics* 6, No. 3 (September): 503–517.
- Blalock, Hubert M., Jr. 1982. *Conceptualization and Measurement in the Social Sciences*. Beverly Hills, Calif.: Sage Publications.
- Blattman, Christopher. 2008. "From Violence to Voting: War and Political Participation in Uganda." *American Political Science Review* 103, no. 2: 231–47.
- Boix, Charles. 1998. *Political Parties, Growth and Equity: Conservative and Social Democratic Economic Strategies in the World Economy*. New York: Cambridge University Press.
- Bollen, Kenneth A. 1989. *Structural Equations with Latent Variables*. New York: Wiley.
- . 1993. "Liberal Democracy: Validity and Method Factors in Cross-National Measures." *American Journal of Political Science* 37, no. 4 (November): 1207–30.
- Bollen, Kenneth A., and Richard Lennox. 1991. "Conventional Wisdom on Measurement: A Structural Equation Perspective." *Psychological Bulletin* 110, no. 2: 305–15.
- Bourguignon, Francois, and Thierry Verdier. 2000. "Oligarchy, democracy, inequality, and growth." *Journal of Development Economics* 62 (August): 285–313.
- Brady, Henry E., and Stephen Ansolehebere. 1989. "The Nature of Utility Functions in Mass Publics." *American Political Science Review* 83, no. 1 (March): 165–92.
- Brady, Henry E., and John McNulty. 2004. "The Costs of Voting: Evidence from a Natural Experiment." Paper presented at the annual meeting of the Society for Political Methodology, Stanford University, July 29–31, 2004.
- Braumoeller, Bear F., and Gary Goertz. 2000. "The Methodology of Necessary Conditions." *American Journal of Political Science* 44, no. 4 (October): 844–58.
- . 2002. "Watching Your Posterior: Bayes, Sampling Assumptions, Falsification, and Necessary Conditions." *Political Analysis* 10, no. 2 (Spring): 198–203.
- Bristowe, John S. and James S. Hutchinson. 1876. *A Treatise on the Theory and Practice of Medicine*. Philadelphia: Henry C. Lea.
- Bristowe, John S., et al. 1879. *Diseases of the Intestines and Peritoneum*. New York: William Wood and Company.
- Brody, Baruch A., and Richard E. Grandy, eds. 1989. *Readings in the Philosophy of Science*. 2nd edition. Englewood Cliffs, N.J.: Prentice-Hall.
- Brooks, Stephen, and William Wohlforth. 2000/01. "Power, Globalization, and the End of the Cold War: Reevaluating a Landmark Case for Ideas." *International Security* 25, no. 3 (Winter): 5–53.
- . 2002. "From Old Thinking to New Thinking in Qualitative Research." *International Security* 26, no. 4 (Spring): 93–111.
- Brown, Archie. 1996. *The Gorbachev Factor*. New York: Oxford University Press.
- Brunetti, Aymo. 1997. "Political variables in cross-country growth analysis." *Journal of Economic Surveys* 11 (June): 163–90.
- Buck, Carol, et al., eds. 1989. *The Challenge of Epidemiology: Issues and Selected Readings*. Geneva: World Health Organization.
- Budd, William. 1873. *Typhoid Fever: Its Nature, Mode of Spreading, and Prevention*. London: Longmans, Green, and Co.
- Bulloch, William. 1938. *The History of Bacteriology*. London: Oxford University Press.

- Bunce, Valerie. 1981. *Do New Leaders Make a Difference? Executive Succession and Public Policy under Capitalism and Socialism*. Princeton: Princeton University Press.
- Cain, Bruce, John Ferejohn, and Morris Fiorina. 1987. *The Personal Vote: Constituency Service and Electoral Independence*. Cambridge, Mass.: Harvard University Press.
- Campbell, Donald T. 1975. "'Degrees of Freedom' and the Case Study." *Comparative Political Studies* 8, no. 2 (July): 178–93.
- . 1988. *Methodology and Epistemology for Social Science*. Chicago: University of Chicago Press.
- Campbell, Donald T., and Robert F. Boruch. 1975. "Making the Case for Randomized Assignment to Treatments by Considering the Alternatives: Six Ways in Which Quasi-Experimental Evaluations in Compensatory Education Tend to Underestimate Effects." In Carl A. Bennett and Arthur A. Lumsdaine, eds., *Evaluation and Experiment: Some Critical Issues in Assessing Social Programs*. New York: Academic Press.
- Campbell, Donald T., and Donald W. Fiske. 1959. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." *Psychological Bulletin* 56, no. 2 (March): 81–105.
- Campbell, Donald T., and H. Laurence Ross. 1968. "The Connecticut Crackdown on Speeding: Time-Series Data in Quasi-Experimental Analysis." *Law and Society Review* 3, no. 1 (August): 33–53.
- Campbell, Donald T., and Julian C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally.
- Campbell, James E. 2000. *The American Campaign: U.S. Presidential Campaigns and the National Vote*. College Station, Tex.: Texas A&M University Press.
- Caporaso, James A. 1995. "Research Design, Falsification, and the Qualitative-Quantitative Divide." *American Political Science Review* 89, no. 2: 457–60.
- Card, David, and Alan B. Krueger. 1994. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania." *American Economic Review* 84, no. 4: 772–93.
- Carpenter, Kenneth J., ed. 1981. *Pellagra*. Stroudsburg, PA.: Hutchinson Ross Pub. Co.
- . 2000. *Beriberi, White Rice, and Vitamin B*. Berkeley: University of California Press.
- Chattopadhyay, Raghavendra, and Esther Duflo. 2004. "Women as Policy Makers: Evidence from a Randomized Experiment in India." *Econometrica* 72, no. 5: 1409–43.
- Checkel, Jeffrey. 1997. *Ideas and International Political Change: Soviet/Russian Behavior and the End of the Cold War*. New Haven: Yale University Press.
- Chehabi, H. E., and Juan L. Linz. 1998. *Sultanistic Regimes*. Baltimore: Johns Hopkins University Press.
- Cholera Inquiry Committee. 1855. *Report on the Cholera Outbreak in the Parish of St. James, Westminster during the Autumn of 1854*. London: Churchill.
- Clarke, Kevin A. 2002. "The Reverend and the Ravens: Comment on Seawright." *Political Analysis* 10, no. 2 (Spring): 194–97.
- . 2005. "The Phantom Menace: Omitted Variable Bias in Econometric Research." *Conflict Management and Peace Science* 22: 341–52.

- Cliff, Norman. 1983. "Some Cautions Concerning the Application of Causal Modeling Methods." *Multivariate Behavioral Research* 18: 115–26.
- Cohen, Bernard P. 1989. *Developing Sociological Knowledge: Theory and Method*. 2nd edition. Chicago: Nelson-Hall.
- Cole, Stephen R., et al. 2010. "Illustrating Bias Due to Conditioning on a Collider." *International Journal of Epidemiology* 39: 417–20.
- Collier, David. 1995. "Translating Quantitative Methods for Qualitative Researchers: The Case of Selection Bias." *American Political Science Review* 89, no. 2 (June): 461–66.
- . 1999. "Data, Field Work and Extracting New Ideas at Close Range." *APSA-CP: Newsletter of the APSA Organized Section in Comparative Politics* 10, no. 1 (Winter): 1–2, 4–6.
- Collier, David, and Robert Adcock. 1999. "Democracy and Dichotomies: A Pragmatic Approach to Choices about Concepts." *Annual Review of Political Science*, vol. 2. Palo Alto: Annual Reviews.
- Collier, David, and Colin Elman. 2008. "Qualitative and Multi-Method Research: Organizations, Publication, and Reflections on Integration." In Janet Box-Steffensmeier, Henry E. Brady, and David Collier, eds., *The Oxford Handbook of Political Methodology*. Oxford: Oxford University Press.
- Collier, David, and Steven Levitsky. 1997. "Democracy with Adjectives: Conceptual Innovation in Comparative Research." *World Politics* 49, no. 3 (April): 430–51.
- Collier, David, James Mahoney, and Jason Seawright. 2004. "Claiming Too Much: Warnings about Selection Bias." In Henry E. Brady and David Collier, eds., *Rethinking Social Inquiry*. Lanham, Md.: Rowman & Littlefield. Available online—see Preface.
- Collier, David, Jasjeet S. Sekhon, and Philip B. Stark. 2010. "Editors' Introduction: Inference and Shoe Leather." In David A. Freedman, *Statistical Models and Causal Inference: A Dialogue with the Social Sciences*. David Collier, Jasjeet Sekhon, and Philip B. Stark, eds. Cambridge: Cambridge University Press.
- Collier, Ruth Berins. 1999. *Paths Toward Democracy: The Working Class and Elites in Western Europe and South America*. New York: Cambridge University Press.
- Collier, Ruth Berins, and David Collier. 1991. *Shaping the Political Arena: Critical Junctures, the Labor Movement, and Regime Dynamics in Latin America*. Princeton: Princeton University Press. Reprinted, 2002, Notre Dame, Ind.: University of Notre Dame Press.
- Cook, Thomas D., and Donald T. Campbell. 1979. *Quasi-Experimentation: Design and Analysis for Field Settings*. Boston: Houghton Mifflin.
- Coombs, Clyde H., Robyn M. Dawes, and Amos Tversky. 1970. *Mathematical Psychology: An Elementary Introduction*. Englewood Cliffs, N.J.: Prentice-Hall.
- Copas, John B., and H. G. Li. 1997. "Inference for Non-Random Samples." *Journal of the Royal Statistical Society* 59 (Series B): 55–77.
- Coppedge, Michael. 1999. "Thickening Thin Concepts and Theories: Combining Large-N and Small in Comparative Politics." *Comparative Politics* 31, no. 4 (July): 465–76.
- . 2001. "Explaining Democratic Deterioration in Venezuela Through Nested Induction." Paper presented at the annual meeting of the American Political Science Association, San Francisco, September 2–5.

- Cowden, Jonathan A., and Thomas Hartley. 1993. "Complex Measures and Sociotropic Voting." In John R. Freeman, ed., *Political Analysis*, vol. 4. Ann Arbor: University of Michigan Press.
- Cox, David R. 1958. *Planning of Experiments*. New York: John Wiley & Sons.
- . 1977. "The Role of Significance Tests." *Scandinavian Journal of Statistics* 4: 49–70.
- . 1992. "Causality: Some Statistical Aspects." *Journal of the Royal Statistical Society* 155 (Series A): 291–301.
- Cox, David R., and N. Wermuth. 1996. *Multivariate Dependencies*. London: Chapman Hall.
- Cox, Gary, Frances Rosenbluth, and Michael F. Thies. 2000. "Electoral Rules, Career Ambitions, and Party Structure: Conservative Factions in Japan's Upper and Lower Houses." *American Journal of Political Science* 44: 115–22.
- Darnell, Adrian C. 1994. *A Dictionary of Econometrics*. Cheltenham, U.K.: Elgar.
- Deaton, Angus. 2009. "Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development." The Keynes Lecture, British Princeton University.
- Deere, Donald, Kevin M. Murphy, and Finis Welch. 1995. "Sense and Nonsense on the Minimum Wage." *Regulation: The Cato Review of Business and Government* 18, no. 1: 47–56.
- De Haan, Jakob, and C. Siermann. 1995. "A Sensitivity Analysis of the Impact of Democracy on Economic Growth." *Empirical Economics* 20, No. 2 (June): 197–215.
- Dehejia, Rajeev. 2005. "Practical Propensity Score Matching: a reply to Smith and Todd." *Journal of Econometrics* 125, no. 1: 355–64.
- Dehejia, Rajeev H., and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94: 1053–62.
- De Soto, Hernando. 1989. *The Other Path: The Economic Answer to Terrorism*. New York: Basic Books.
- . 2000. *The Mystery of Capital: Why Capitalism Triumphs in the West and Fails Everywhere Else*. New York: Basic Books.
- Diaconis, Persi. 1998. "A Place for Philosophy? The Rise of Modeling in Statistical Science." *Quarterly of Applied Mathematics* 56, no. 4 (December): 797–805.
- Diamond, Jared, and James A. Robinson, eds. 2010. *Natural Experiments in History*. Cambridge: Harvard University Press.
- Dietz, Thomas, R. Scott Frey, and Linda Kalof. 1987. "Estimation with Cross-National Data: Robust and Nonparametric Methods." *American Sociological Review* 52, no. 3 (June): 380–90.
- Dijkstra, Theo K., ed. 1988. "On Model Uncertainty and Its Statistical Implications." Proceedings of a workshop held in Groningen, the Netherlands, September 25–26, 1986. *Lecture Notes in Economics and Mathematical Systems*, no. 307. Berlin, New York: Springer.
- Dion, Douglas. 1998. "Evidence and Inference in the Comparative Case Study." *Comparative Politics* 30, no. 2 (January): 127–45.
- Di Tella, Rafael, Sebastian Galiani, and Ernesto Schargrodsky. 2007. "The Forma-

- tion of Beliefs: Evidence from the Allocation of Land Titles to Squatters." *Quarterly Journal of Economics* 122: 209–41.
- Doherty, Daniel, Donald Green, and Alan Gerber. 2006. "Personal Income and Attitudes toward Redistribution: A Study of Lottery Winners." *Political Psychology* 27 (3): 441–58. Earlier version circulated as a working paper, Institution for Social and Policy Studies, Yale University, June 30, 2005.
- Drezner, Daniel W. 1999. *The Sanctions Paradox: Economic Statecraft and International Relations*. Cambridge: Cambridge University Press.
- Druckman, James N., Donald P. Green, James H. Kuklinski, and Arthur Lupia. 2006. "The Growth and Development of Experimental Research in Political Science." *American Political Science Review* 100, no. 4: 627–635.
- Duflo, Esther, and Michael Kremer. 2006. "Using Randomization in Development Economics Research: A Toolkit." Working paper, Departments of Economics, MIT and Harvard.
- Duncan, Otis Dudley. 1984a. *Notes on Social Measurement: Historical and Critical*. New York: Russell Sage Foundation.
- . 1984b. "Measurement and Structure: Strategies for the Design and Analysis of Subjective Survey Data." In Charles F. Turner and Elizabeth Martin, eds., *Surveying Subjective Phenomena*. New York: Russell Sage Foundation.
- Dunnill, Michael S. 2001. *The Plato of Praed Street: The Life and Times of Almroth Wright*. London: Royal Society of Medicine Press.
- Dunning, Thad. 2008a. "Improving Causal Inference: Strengths and Limitations of Natural Experiments." *Political Research Quarterly* 61 no. 2: 282–93.
- . 2008b. "Natural and Field Experiments: The Role of Qualitative Methods." *Qualitative and Multi-Method Research* 6, no. 2: 17–22.
- . 2008c. "Model Specification in Instrumental-Variables Regression." *Political Analysis* 16, no. 3: 290–302.
- . 2009. "The Salience of Ethnic Categories: Field and Natural Experimental Evidence from Indian Village Councils." Working paper, Department of Political Science, Yale University.
- Durham, J. Benson. 1999. "Economic Growth and Political Regimes." *Journal of Economic Growth* 4, no. 1: 81–111.
- Eckstein, Harry. 1975. "Case Study and Theory in Political Science." In Fred I. Greenstein and Nelson W. Polsby, eds., *Handbook of Political Science*, vol. 7. Reading, Mass.: Addison-Wesley.
- Eden, Lynn. 2004. *Whole World on Fire: Organizations, Knowledge, and Nuclear Weapons Devastation*. Ithaca: Cornell University Press.
- Eliot, Charles W., ed. 1910 [1897]. *Scientific Papers: Physiology, Medicine, Surgery, Geology*. Vol. 38 in *The Harvard Classics*. New York: P. F. Collier & Son.
- Elster, Jon. 1999. *Alchemies of the Mind: Rationality and the Emotions*. New York: Cambridge University Press.
- Engerman, Stanley L., Elisa Mariscal and Kenneth L. Sokoloff. 1998. "Schooling, suffrage, and the persistence of inequality in the Americas, 1800–1945." Unpublished paper, Department of Economics, UCLA.
- English, Robert. 2000. *Russia and the Idea of the West: Gorbachev, Intellectuals, and the End of the Cold War*. New York: Columbia University Press.

- . 2002. "Power, Ideas, and New Evidence on the Cold War's End: A Reply to Brooks and Wohlforth." *International Security* 26, no. 4 (Spring): 70–92.
- Epstein, Lee, and Gary King. 2002. "The Rules of Inference." *The University of Chicago Law Review* 69, no. 1 (Winter): 1–133.
- Evans, Richard J. 1987. *Death in Hamburg: Society and Politics in the Cholera Years, 1830–1910*. Oxford: Oxford University Press.
- Eyler, John M. 1979. *Victorian Social Medicine: The Ideas and Methods of William Farr*. Baltimore: Johns Hopkins University Press.
- Fearon, James D., and David D. Laitin. 2008. "Integrating Qualitative and Quantitative Methods." In Janet Box-Steffensmeier, Henry E. Brady, and David Collier, eds., *The Oxford Handbook of Political Methodology*. Oxford: Oxford University Press.
- Feigl, Herbert. 1958. "The Mental and the Physical." In Herbert Feigl, Michael Scriven, and Grover Maxwell, eds., *Minnesota Studies in the Philosophy of Science, vol. 2: Concepts, Theories and the Mind-Body Problem*. Minneapolis: University of Minnesota Press.
- Feng, Yi. 1997. "Democracy, political stability, and economic growth." *British Journal of Political Science* 27 (July): 391–418.
- Feng, Yi, and Paul J. Zak. 1999. "The determinants of democratic transitions." *Journal of Conflict Resolution* 43, no. 2: 162–77.
- Fenner, Frank, et al. 1988. *Smallpox and its Eradication*. Geneva: World Health Organization.
- Fenno, Richard F. 1977. "U.S. House Members in Their Constituencies: An Exploration." *American Political Science Review* 71, no. 3 (September): 883–917.
- . 1978. *Home Style: House Members in Their Districts*. Boston: Little, Brown.
- Ferraz, Claudio, and Frederico Finan. 2008. "Exposing Corrupt Politicians: The Effect of Brazil's Publicly Released Audits on Electoral Outcomes." *Quarterly Journal of Economics* 123, no. 2: 703–45.
- Feynman, Richard Phillips. 1965. *The Character of Physical Law*. Cambridge, Mass.: MIT Press.
- Fish, M. Steven. 1995. *Democracy from Scratch: Opposition and Regime in the New Russian Revolution*. Princeton: Princeton University Press.
- Fisher, Sir Ronald Aylmer. 1935. *The Design of Experiments*. Edinburgh: Oliver & Boyd.
- Fleming, Alexander. 1929. "On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of *B. influenzae*." *British Journal of Experimental Pathology* 10: 226–36.
- Freedman, David A. 1983. "A Note on Screening Regression Equations." *American Statistician* 37, no. 2 (May): 152–55.
- . 1991. "Statistical Models and Shoe Leather." In Peter Marsden, ed., *Sociological Methodology*. San Francisco: Jossey-Bass.
- . 1992a. "As Others See Us: A Case Study in Path Analysis." In J. P. Shaffer, ed., *The Role of Models in Nonexperimental Social Science: Two Debates*. Washington, D.C.: American Educational Research Association and American Statistical Association.
- . 1992b. "A Rejoinder on Models, Metaphors and Fables." In J. P. Shaffer, ed., *The Role of Models in Nonexperimental Social Science: Two Debates*. Washington,

- D.C.: American Educational Research Association and American Statistical Association.
- . 1999. "From Association to Causation: Some Remarks on the History of Statistics." *Statistical Science* 14: 243–58.
- . 2005. *Statistical Models: Theory and Practice*. New York: Cambridge University Press.
- . 2006. "Statistical Models for Causation: What Inferential Leverage Do They Provide?" *Evaluation Review* 30: 691–713.
- . 2008a. "On regression adjustments to experimental data." *Advances in Applied Mathematics* 40: 180–93.
- . 2008b. "On regression adjustments in experiments with several treatments." *Annals of Applied Statistics* 2: 176–96.
- . 2009. *Statistical Models: Theory and Practice*. Cambridge: Cambridge University Press, 2nd edition.
- . 2010. *Statistical Models and Causal Inference: A Dialogue with the Social Sciences*. David Collier, Jasjeet Sekhon, and Philip B. Stark, eds. Cambridge: Cambridge University Press.
- Freedman, David A., and Richard A. Berk. 2010. "Statistical Assumptions as Empirical Commitments." In David A. Freedman, *Statistical Models and Causal Inference: A Dialogue with the Social Sciences*, David Collier, Jasjeet S. Sekhon, and Philip B. Stark, eds. New York: Cambridge University Press.
- Freedman, David A., and David Lane. 1983. "A Nonstochastic Interpretation of Reported Significance Levels." *Journal of Business and Economic Statistics* 1: 292–98.
- Freedman, David A., Robert Pisani, and Roger Purves. 2007. *Statistics*. 4th edition. New York: W. W. Norton & Company.
- Friedman, Milton. 1962. *Capitalism and Freedom*. Chicago: University of Chicago Press.
- Galiani, Sebastian, and Ernesto Scharrotsky. 2004. "The Health Effects of Land Titling." *Economics and Human Biology* 2: 353–72.
- Garrett, Geoffrey. 1998. *Partisan Politics in the Global Economy*. New York: Cambridge University Press.
- Gasiorowski, Mark J. 2000. "Democracy and macroeconomic performance in underdeveloped countries: An empirical analysis." *Comparative Political Studies* 33 (April): 319–49.
- Geertz, Clifford. 1973. "Thick Description: Toward an Interpretive Theory of Culture." In C. Geertz, ed., *The Interpretation of Cultures*. New York: Basic Books.
- Gelman, Andrew, and Gary King. 1994. "A Unified Method of Evaluating Electoral Systems and Redistricting Plans." *American Journal of Political Science* 38, no. 2 (May): 514–54.
- George, Alexander L. 1979. "Case Studies and Theory Development: The Method of Structured, Focused Comparison." In Paul Gordon Lauren, ed., *Diplomacy: New Approaches in History, Theory and Policy*. New York: Free Press.
- George, Alexander L., and Andrew Bennett. 2005. *Case Studies and Theory Development*. Cambridge, Mass.: MIT Press.
- George, Alexander L., and Timothy J. McKeown. 1985. "Case Studies and Theories of Organizational Decision Making." In Robert F. Coulam and Richard A. Smith,

- eds., *Advances in Information Processing in Organizations*, vol. 2. Greenwich, Conn.: JAI Press.
- George, Alexander L., and Richard Smoke. 1974. *Deterrence in American Foreign Policy: Theory and Practice*. New York: Columbia University Press.
- Gerber, Alan S., and Donald P. Green. 2008. "Field Experiments and Natural Experiments." In Janet Box-Steffensmeier, Henry E. Brady, and David Collier, eds., *The Oxford Handbook of Political Methodology*. New York: Oxford University Press, 357–81.
- Gerring, John. 2001/2011 forthcoming. *Social Science Methodology: A Criterial Framework*. New York: Cambridge University Press.
- Giesbrecht, Peter, et al. 1998. "Staphylococcal cell wall: Morphogenesis and fatal variations in the presence of penicillin." *Microbiology and Molecular Biology Reviews* 62: 1371–1414.
- Gill, Christopher J., Lora Sabin, and Christopher H. Schmid. 2005. "Why Clinicians Are Natural Bayesians." *British Medical Journal* 330 (May): 1080–83.
- Gillespie, John V., and Betty A. Nesvold. 1971. *Macro-Quantitative Analysis: Conflict, Development and Democratization*. Beverly Hills, Calif.: Sage.
- Gilligan, Michael J., and Ernest J. Sergenti. 2008. "Do UN Interventions Cause Peace? Using Matching to Improve Causal Inference." *Quarterly Journal of Political Science* 3, no 2: 89–122.
- Glaeser, Edward L., Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer. 2004. "Do Institutions Cause Growth?" *Journal of Economic Growth* 9, no. 3: 271–303.
- Glaser, Barney G., and Anselm L. Strauss. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago: Aldine.
- Glazer, Amihai, and Marc Robbins. 1985. "Congressional Responsiveness to Constituency Change." *American Journal of Political Science* 29, no. 2: 259–73.
- Goemans, Hein. 2000. *War and Punishment: The Causes of War Termination and the First World War*. Princeton: Princeton University Press.
- Goertz, Gary. 2003. "The Substantive Importance of Necessary Condition Hypotheses." In Gary Goertz and Harvey Starr, eds., *Necessary Conditions: Theory, Methodology, and Applications*. Lanham, Md.: Rowman & Littlefield.
- Goertz, Gary, and Harvey Starr, eds. 2003. *Necessary Conditions: Theory, Methodology, and Applications*. Lanham, Md.: Rowman & Littlefield.
- Goldberger, Joseph. 1914. "The etiology of pellagra." *Public Health Reports* 29: 1683–86.
- Goldberger, Joseph, et al. 1926. "A further study of butter, fresh beef, and yeast as pellagra preventives, with consideration of the relation of factor P-P of pellegra (and black tongue of dogs) to vitamin B1." *Public Health Reports* 41: 297–318.
- Goldsmith, Margaret. 1946. *The Road to Penicillin*. London: Lindsay Drummond.
- Goldstone, Jack A. 1991. *Revolution and Rebellion in the Early Modern World*. Berkeley: University of California Press.
- Goldthorpe, John H. 2001. "Causation, Statistics, and Sociology." *European Sociological Review* 17, no. 1: 1–20.
- Gourevitch, Peter Alexis. 1978. "The International System and Regime Formation: A Critical Review of Anderson and Wallerstein." *Comparative Politics* 10, no. 3 (April): 419–38.

- Granger, Clive W. J., ed. 1990. *Modelling Economic Series: Readings in Econometric Methodology*. Oxford: Clarendon.
- Green, Donald. 2009. "Regression Adjustments to Experimental Data: Do David Freedman's Concerns Apply to Political Science?" Manuscript, Department of Political Science, Yale University.
- Greene, Kenneth F. 2002. "Defeating Dominance: Opposition Party Building and Democratization in Mexico." Doctoral dissertation, Department of Political Science, University of California, Berkeley.
- Greene, William H. 2000. *Econometric Analysis*. 4th edition. Upper Saddle River, N.J.: Prentice-Hall.
- Griffin, Larry J. 1992. "Temporality, Events, and Explanation in Historical Sociology: An Introduction." *Sociological Methods and Research* 20, no. 4: 403–27.
- Griliches, Zvi. 1986. "Economic Data Issues." In Z. Griliches and Michael Intriligator, eds., *Handbook of Econometrics*, vol. 3. Amsterdam: North Holland.
- Grofman, Bernard, Thomas L. Brunell, and William Koetzle. 1998. "Why Gain in the Senate but Midterm Loss in the House? Evidence from a Natural Experiment." *Legislative Studies Quarterly* 23, no. 1: 79–89.
- Grofman, Bernard, Robert Griffin, and Gregory Berry. 1995. "House Members Who Become Senators: Learning from a 'Natural Experiment.'" *Legislative Studies Quarterly* 20, no. 4: 513–29.
- Gujarati, Damodar N. 1988. *Basic Econometrics*. 2nd edition. New York: McGraw-Hill.
- Haggard, Stephan, and Robert R. Kaufman. 1995. *The Political Economy of Democratic Transitions*. Princeton: Princeton University Press.
- Hanushek, Eric A., and John E. Jackson. 1977. *Statistical Methods for Social Scientists*. Orlando, Fla.: Academic Press.
- Hare, Ronald. 1970. *The Birth of Penicillin and the Disarming of Microbes*. London: Allen & Unwin.
- Heberle, Rudolf. 1963. *Landbevölkerung und Nationalsozialismus: Eine soziologische Untersuchung der politischen Willensbildung in Schleswig-Holstein 1918 bis 1932*. Revised edition. Stuttgart: Deutsche Verlags-Anstalt.
- . 1970. *From Democracy to Nazism: A Regional Case Study on Political Parties in Germany*. New York: Grosset & Dunlap.
- Heckman, James J. 1988. "The microeconomic evaluation of social programs and economic institutions." In *Chung-Hua Series of Lectures by Invited Eminent Economists*, 14. Taipei: Institute of Economics, Academia Sinica.
- . 1990. "Varieties of Selection Bias." *American Economic Review* 80, no. 2: 313–18.
- . 1992. "Randomization and Social Policy Evaluation." In Charles F. Manski and Irwin Garfinkel, eds., *Evaluating Welfare and Training Programs*. Cambridge, Mass.: Harvard University Press.
- . 2000. "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective." *Quarterly Journal of Economics* 115: 45–97.
- Heckman, James, and Sergio Urzua. 2009. "Comparing IV with Structural Models: What Simple IV Can and Cannot Identify." NBER Working Paper #14706.
- Hedström, Peter. 2008. "Studying Mechanisms to Study Causal Inferences in Quan-

- titative Research." In Janet Box-Steffensmeier, Henry E. Brady, and David Collier, eds., *The Oxford Handbook of Political Methodology*. Oxford: Oxford University Press.
- Helliwell, J.F. 1994. "Empirical linkages between democracy and economic growth." *British Journal of Political Science* 24 (April), Part 2: 225–48.
- Hempel, Carl G. 1965. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: Free Press.
- . 1966. *Philosophy of Natural Science*. Englewood Cliffs, N.J.: Prentice-Hall.
- . 1970 [1952]. *Fundamentals of Concept Formation in Empirical Science*. Chicago: University of Chicago Press.
- Hendry, David F. 1980. "Econometrics—Alchemy or Science." *Economica* 47, no. 188 (November): 387–406.
- Hendry, David F., and Jean-François Richard. 1982. "On the Formulation of Empirical Models in Dynamic Econometrics." *Journal of Econometrics* 20, no. 3: 3–33. Also in Granger, 1990.
- Hibbs, Douglas A. 1987. "On the Political Economy of Long-Run Trends in Strike Activity." In Hibbs, *The Political Economy of Industrial Democracies*. Cambridge, Mass.: Harvard University Press.
- Hicks, Alexander. 1988. "Social Democratic Corporatism and Economic Growth." *Journal of Politics* 50, no. 3 (August): 677–704.
- Hicks, Alexander, and William David Patterson. 1989. "On the Robustness of the Left Corporatist Model of Economic Growth." *Journal of Politics* 51, no. 3 (August): 662–75.
- Hidalgo, F. Daniel, Suresh Naidu, Simeon Nichter, and Neal Richardson. Forthcoming. "Occupational Choices: Economic Determinants of Land Invasions." *Review of Economics and Statistics*.
- Hill, Austin Bradford. 1991 [1937]. *Principles of Medical Statistics*. 12th edition. London: Arnold.
- Ho, Daniel E., and Kosuke Imai. 2008. "Estimating Causal Effects of Ballot Order from a Randomized Natural Experiment: California Alphabet Lottery 1978–2002." *Public Opinion Quarterly* 72, no. 2: 216–40.
- Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2007. "Matching as Nonparametric Preprocessing for Improving Parametric Causal Inference." *Political Analysis* 15: 199–236.
- Hodges, Joseph L., and Erich L. Lehmann. 1964. *Basic Concepts of Probability and Statistics*. San Francisco: Holden-Day.
- Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81, no. 396 (December): 945–60.
- Homer-Dixon, Thomas F. 1999. *Environment, Scarcity, and Violence*. Princeton: Princeton University Press.
- Hoole, Francis, and Dina Zinnes, eds. 1976. *Quantitative International Politics: An Appraisal*. New York: Praeger.
- Humphreys, Noel A., ed. 1885. *Vital Statistics: A Memorial Volume of Selections from the Reports and Writings of William Farr*. London: Edward Stanford.
- Huntington, Samuel P. 1991. *The Third Wave: Democratization in the Late Twentieth Century*. Norman: University of Oklahoma Press.

- . 1996. *The Clash of Civilizations and the Remaking of World Order*. New York: Simon and Schuster.
- Hyde, Susan. 2007. "The Observer Effect in International Politics: Evidence from a Natural Experiment." *World Politics* 60: 37–63.
- Imai, Kosuke. 2005. "Do Get-Out-The-Vote Calls Reduce Turnout? The Importance of Statistical Methods for Field Experiments." *American Political Science Review* 99, no. 2 (May): 283–300.
- Imbens, Guido. 2009. "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." Manuscript, Department of Economics, Harvard University.
- Jackman, Robert W. 1985. "Cross-national statistical research and the study of comparative politics." *American Journal of Political Science* 29 (February): 161–82.
- . 1987. "The Politics of Growth in the Industrial Democracies, 1974–80: Leftist Strength or North Sea Oil?" *Journal of Politics* 49, no. 1 (February): 242–56.
- . 1989. "The Politics of Economic Growth, Once Again." *Journal of Politics* 51, no. 3 (August): 646–61.
- Jackson, John E. 1983. "Election Night Reporting and Voter Turnout." *American Journal of Political Science* 27, no. 4 (November): 615–35.
- Jenner, Edward. 1798. *An Inquiry into the Causes and Effects of the Variolae Vaccinae, a Disease Discovered in Some of the Western Counties of England, Particularly Gloucestershire, and Known by the Name of the Cow Pox*. Reprinted in Eliot 1910 [1897].
- . 1801. *The Origin of the Vaccine Inoculation*. London: D. N. Shury.
- Katzenstein, Peter J. 1985. *Small States in World Markets: Industrial Policy in Europe*. Ithaca, N.Y.: Cornell University Press.
- Kennedy, Peter. 1998. *A Guide to Econometrics*. Cambridge, Mass.: MIT Press.
- Keohane, Robert O. 2003. "Disciplinary Schizophrenia: Implications for Graduate Education in Political Science." *Qualitative Methods* 1, no. 1 (Spring): 9–12.
- Kerlinger, Fred N. 1979. *Behavioral Research: A Conceptual Approach*. New York: Holt, Rinehart, and Winston.
- Key, V. O., Jr. 1984 [1949]. *Southern Politics in State and Nation*. Knoxville: University of Tennessee Press.
- Khong, Yuen Foong. 1992. *Analogies at War: Korea, Munich, Dien Bien Phu and the Vietnam Decisions of 1965*. Princeton: Princeton University Press.
- Kim, Jae-On, and Charles W. Mueller. 1978. *Introduction to Factor Analysis: What It Is and How to Do It*. London: Sage Publications.
- King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton: Princeton University Press.
- . 1995. "The Importance of Research Design in Political Science." *American Political Science Review* 89, no. 2 (June): 475–80.
- Kitschelt, Herbert. 1994. *The Transformation of European Social Democracy*. New York: Cambridge University Press.
- Kittel, Bernhard. 1999. "Sense and Sensitivity in Pooled Analysis of Political Data." *European Journal of Political Research* 35: 225–53.
- Knopf, Jeffrey W. 1998. *Domestic Society and International Cooperation: The Impact of Protest on U.S. Arms Control Policy*. Cambridge: Cambridge University Press.
- Kohli, Atul. 1987. *The State and Poverty in India*. New York: Cambridge University Press.

- Kornhauser, William. 1959. *The Politics of Mass Society*. Glencoe, Ill.: Free Press.
- Krantz, David L., R. Duncan Luce, Patrick Suppes, and Amos Tversky. 1971, 1989, 1990. *Foundations of Measurement*, vols. 1, 2, and 3. New York: Academic Press.
- Krasno, Jonathan S., and Donald P. Green. 2008. "Do Televised Presidential Ads Increase Voter Turnout? Evidence from a Natural Experiment." *Journal of Politics* 70, no. 1: 245–61.
- Kriekhaus, Jonathan. 2004. "The regime debate revisited: A sensitivity analysis of democracy's economic effects." *British Journal of Political Science* 34: 635–55.
- Kriesi, Hanspeter, Ruud Koopmans, Jan Willem Duyvendak, and Marco G. Giugni. 1995. *New Social Movements in Western Europe: A Comparative Analysis*. Minneapolis: University of Minnesota Press.
- Kubik, Jan. 1994. *The Power of Symbols against the Symbols of Power: The Rise of Solidarity and the Fall of State Socialism in Poland*. University Park, Penn.: Pennsylvania State University Press.
- Kuhn, Thomas. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
- Kurzman, Charles, Regina Werum, and Ross E. Burkhart. 2002. "Democracy's effect on economic growth: A pooled time-series analysis, 1951–1980." *Studies in Comparative International Development* 37 (Spring): 3–33.
- Laba, Roman. 1991. *The Roots of Solidarity: A Political Sociology of Poland's Working-Class Democratization*. Princeton: Princeton University Press.
- Laitin, David D. 1986. *Hegemony and Culture: Politics and Religious Change among the Yoruba*. Chicago: The University of Chicago Press.
- . 1995. "Disciplining Political Science." *American Political Science Review* 89, no. 2 (June): 454–56.
- . 2002. "Comparative Politics: The State of the Subdiscipline." In Ira Katznelson and Helen Milner, eds., *Political Science: State of the Discipline*. New York: Norton.
- Lakatos, Imre, and Alan Musgrave, eds. 1970. *Criticism and the Growth of Knowledge*. Cambridge: Cambridge University Press.
- Lane, Robert E. 1962. *Political Ideology: Why the American Common Man Believes What He Does*. New York: Free Press.
- Lang, Janet, Kenneth J. Rothman, and C. L. Cann. 1998. "That Confounded P-Value." *Epidemiology* 9, no. 1 (January): 7–8.
- Lange, Peter, and Geoffrey Garrett. 1985. "The Politics of Growth: Strategic Interaction and Economic Performance in the Advanced Industrial Democracies, 1974–1980." *Journal of Politics* 47 (August): 792–827.
- . 1987. "The Politics of Growth Reconsidered." *Journal of Politics* 49 (February): 257–74.
- Larson, Deborah Welch. 1997. *Anatomy of Mistrust: U.S.-Soviet Relations During the Cold War*. Ithaca: Cornell University Press.
- Lave, Charles, and James G. March. 1975. *An Introduction to Models in the Social Sciences*. New York: Harper & Row.
- Layne, Christopher. 1994. "Kant or Cant: The Myth of the Democratic Peace." *International Security* 19, no. 2: 5–49.
- Lazarsfeld, Paul F. 1955. "Interpretation of Statistical Relations as a Research Opera-

- tion." In Paul F. Lazarsfeld and Morris Rosenberg, *The Language of Social Research*. New York: Free Press.
- Leamer, Edward E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: Wiley.
- . 1983. "Let's Take the Con Out of Econometrics." *American Economic Review* 73, no. 1 (March): 31–43.
- . 1990. "Specification Problems in Econometrics." In J. Eatwell, M. Milgate and P. Newman (eds.), *Econometrics: The New Palgrave*. New York: W.W. Norton: 238–45.
- Leblang, David A. 1997. "Political democracy and economic growth: Pooled cross-sectional and time-series evidence." *British Journal of Political Science* 27 (July): 453–66.
- Lederer, Emil. 1940. *State of the Masses: The Threat of the Classless Society*. New York: W. W. Norton.
- Lee, David S. 2008. "Randomized Experiments from Non-random Selection in U.S. House Elections." *Journal of Econometrics* 142, no. 2: 675–97.
- Lerman, Amy. 2008. "Bowling Alone (With My Own Ball and Chain): The Effects of Incarceration and the Dark Side of Social Capital." Manuscript, Department of Politics, Princeton University.
- Levine, Ross, and David Renelt. 1992. "A Sensitivity Analysis of Cross-Country Growth Regressions." *The American Economic Review* 82, no. 4 (September): 942–63.
- Lewis-Beck, Michael, and Tom W. Rice. 1992. *Forecasting Elections*. Washington, D.C.: Congressional Quarterly.
- Lieberman, Evan S. 2003a. "Nested Analysis in Cross-National Research." *APSA-CP: Newsletter of the APSA Comparative Politics Section* 14, no. 1 (Winter): 17–20.
- . 2003b. *Race and Regionalism in the Politics of Taxation in Brazil and South Africa*. New York: Cambridge University Press.
- Liebersohn, Stanley. 1985. *Making It Count: The Improvement of Social Research and Theory*. Berkeley: University of California Press.
- Lijphart, Arend. 1971. "Comparative Politics and the Comparative Method." *American Political Science Review* 65, no. 3 (September): 682–93.
- . 1975 [1968]. *The Politics of Accommodation: Pluralism and Democracy in the Netherlands*. Berkeley: University of California Press.
- Lindauer, David, and Lant Pritchett. 2002. "What's the Big Idea? Three Generations of Development Advice." *Economia* 3, no. 1: 1–28.
- Lipton, Michael. 1976. *Why Poor People Stay Poor: Urban Bias in World Development*. Cambridge, Mass.: Harvard University Press.
- Liu, Ta Chung. 1960. "Under-Identification, Structural Estimation, and Forecasting." *Econometrica* 28: 855–65.
- Loehlin, John C. 2004. *Latent Variable Models: An Introduction to Factor, Path, and Structural Equation Analysis*. Mahwah: Lawrence Erlbaum Associates, Publishers.
- Lott, John R., Jr. 2000. "Gore Might Lose a Second Round: Media Suppressed the Bush Vote." *Philadelphia Inquirer*, Tuesday, 14 November 2000, p. 23A.
- Loudon, Irvine. 2000. *The Tragedy of Childbed Fever*. Oxford: Oxford University Press.

- Lucas, Robert E., Jr. 1976. "Econometric Policy Evaluation: A Critique." In K. Brunner and A. Meltzer, eds., *The Phillips Curve and Labor Markets*, vol. 1. Carnegie-Rochester Conferences on Public Policy, Supplementary Series to the *Journal of Monetary Economics*. North-Holland, Amsterdam: North-Holland.
- Lyall, Jason. 2009. "Does Indiscriminate Violence Incite Insurgent Attacks? Evidence from Chechnya." *Journal of Conflict Resolution* 53, no. 3: 331–62.
- Mahoney, James. 1999. "Nominal, Ordinal, and Narrative Appraisal in Macrocausal Analysis." *American Journal of Sociology* 104, no. 4 (January): 1154–96.
- . 2010. "After KKV: The New Methodology of Qualitative Research." *World Politics* 62, no. 1: 120–47.
- Manski, Charles F. 1995. *Identification Problems in Social Sciences*. Cambridge, Mass.: Harvard University Press.
- Marshall, Monty G., and Keith Jagers. 1998. *Polity IV Project: Political Regime Characteristics and Transitions, 1800–2002*. Accessed May 2000. <http://www.systemicpeace.org/polity/polity4.htm>.
- Martin, Lisa L. 1992. *Coercive Cooperation: Explaining Multilateral Economic Sanctions*. Princeton: Princeton University Press.
- Mason, Linda, Kathleen Frankovic, and Kathleen Hall Jamieson. 2001. "CBS News Coverage of Election Night 2000." Accessed 5 November 2003 at www.cbsnews.com/htdocs/c2k/pdf/REPFINAL.pdf.
- Mauldon, Jane, Janet Malvin, Jon Stiles, Nancy Nicosia, and Eva Y. Seto. 2000. "Impact of California's Cal-Learn Demonstration Project: Final Report." UC Data, University of California at Berkeley.
- McAdam, Doug. 1988. *Freedom Summer*. New York: Oxford University Press.
- McAdam, Doug, Sidney Tarrow, and Charles Tilly. 2001. *Dynamics of Contention*. New York: Cambridge University Press.
- McKeown, Timothy J. 2004. "Case Studies and the Limits of the Quantitative Worldview." In Henry E. Brady and David Collier, eds., *Rethinking Social Inquiry*. Lanham, Md.: Rowman & Littlefield. Available online—see Preface.
- McKim, Vaughn R., and Stephen P. Turner, eds. 1997. *Causality in Crisis? Statistical Methods and the Search for Causal Knowledge in the Social Sciences*. Notre Dame, Ind.: Notre Dame Press.
- Meehl, Paul E. 1978. "Theoretical Risks and Tabular Asterisks." *Journal of Consulting and Clinical Psychology* 46, no. 4: 806–34.
- Messick, Samuel. 1975. "The Standard Problem: Meaning and Values in Measurement and Evaluation." *American Psychologist* 30: 955–66.
- . 1989. "Validity." In Robert L. Linn, ed., *Educational Measurement*. 3rd edition. New York: Macmillan.
- Michell, Joel. 1990. *An Introduction to the Logic of Psychological Measurement*. Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Michels, Robert. 1915. *Political Parties: A Sociological Study of the Oligarchical Tendencies of Modern Democracy*. London: Jarrold & Sons.
- Migdal, Joel S. 1988. *Strong Societies and Weak States: State-Society Relations and State Capabilities in the Third World*. Princeton: Princeton University Press.
- Miguel, Edward. 2004. "Tribe or Nation: Nation Building and Public Goods in Kenya versus Tanzania." *World Politics* 56, no. 3: 327–62.

- Miguel, Edward, Shanker Satyanath, and Ernest Sergenti. 2004. "Economic Shocks and Civil Conflict: An Instrumental Variables Approach." *Journal of Political Economy* 122: 725–53.
- Mill, John Stuart. 1974 [1843]. "Of the Four Methods of Experimental Inquiry" and "Of the Chemical, or Experimental, Method in the Social Science." *A System of Logic, Raciocinative and Inductive*. Toronto: University of Toronto Press.
- Minier, Jenny A. 1998. "Democracy and growth: Alternative approaches." *Journal of Economic Growth* 3 (September): 241–66.
- Mirer, Thad W. 1995. *Economic Statistics and Econometrics*. 3rd edition. Englewood Cliffs, N.J.: Prentice-Hall.
- Moore, Barrington, Jr. 1967. *Social Origins of Dictatorship and Democracy: Lord and Peasant in the Making of the Modern World*. Boston: Beacon.
- Moravcsik, Andrew. 1998. *The Choice for Europe: Social Purpose and State Power from Messina to Maastricht*. Ithaca: Cornell University Press.
- Morgan, Stephen L., and Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York: Cambridge University Press.
- Morton, Rebecca B., and Kenneth C. Williams. 2008. "Experimentation in Political Science." In Janet Box-Steffensmeier, Henry E. Brady, and David Collier, eds., *The Oxford Handbook of Political Methodology*. New York: Oxford University Press.
- Moss, Pamela A. 1992. "Shifting Conceptions of Validity in Educational Measurement: Implications for Performance Assessment." *Review of Educational Research* 62 (Fall): 229–58.
- . 1995. "Themes and Variations in Validity Theory." *Educational Measurement: Issues and Practice* 14 (Summer): 5–13.
- Munck, Gerardo L. 1998. "Canons of Research Design in Qualitative Analysis." *Studies in Comparative International Development* 33, no. 3 (Fall): 18–45.
- . 2001. "Game Theory and Comparative Politics: New Perspectives and Old Concerns." *World Politics* 53, no. 2 (January): 173–204.
- . 2004. "Tools for Qualitative Research." In Henry E. Brady and David Collier, eds., *Rethinking Social Inquiry*. Lanham, Md.: Rowman & Littlefield. Available online—see Preface.
- Nelson, Michael A., and Ram D. Singh. 1998. "Democracy, economic freedom, fiscal policy, and growth in LDCs: A fresh look." *Economic Development and Cultural Change* 46 (July): 677–96.
- Neyman, Jerzy Splawa, with D. M. Dabrowska, and T. P. Speed. 1990 [1923]. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science* 5, no. 4 (November): 465–72. Originally published by Neyman in Polish in the *Annals of Agricultural Sciences*.
- Ní Bhrolcháin, Máire. 2001. "'Divorce Effects' and Causality in the Social Sciences." *European Sociological Review* 17, no. 1: 33–57.
- Nuland, Sherwin. 1979. "The enigma of Semmelweis—An interpretation." *Journal of the History of Medicine and Allied Sciences* 34: 255–72.
- Nutton, Vivian, ed. 2008. *Pestilential Complexities*. London: Wellcome Trust.
- Olson, Mancur. 1983. "The political economy of comparative growth rates." In D.C. Mueller (Ed.), *The political economy of growth*. New Haven: Yale University Press.

- . 1990. *How bright are the northern lights? Some questions about Sweden*. Lund: Lund University Press.
- Owen, John M., IV. 1997. *Liberal Peace, Liberal War: American Politics and International Security*. Ithaca: Cornell University Press.
- Oye, Kenneth. 1996. "Explaining the End of the Cold War: Morphological and Behavior Adaptations to the Nuclear Peace?" In Richard Ned Lebow and Thomas Risse-Kappen, eds., *International Relations Theory and the End of the Cold War*. New York: Columbia University Press.
- Paluck, Elizabeth Levy. 2008. "The Promising Integration of Qualitative Methods and Field Experiments." *Qualitative and Multi-Method Research* 6, no. 2: 23–30.
- Perrot, Michelle. 1986. "On the Formation of the French Working Class." In Ira Katznelson and Aristide Zolberg, eds., *Working Class Formation: Nineteenth Century Patterns in Western Europe and the United States*. Princeton: Princeton University Press.
- Pinto, Pablo M., and Jeffrey M. Timmons. 2005. "The political determinants of economic performance: Political competition and the sources of growth." *Comparative Political Studies* 38 (February): 26–50.
- Piore, Michael J. 1979. "Discovering Qualitative Research." *Administrative Science Quarterly* 24 (December): 560–69.
- . 2006. "Qualitative Research: Does it Fit in Economics?" *European Management Review* 3, no. 1: 17–23.
- Pitkin, Hanna. 1967. *The Concept of Representation*. Berkeley: University of California Press.
- Polanyi, Karl. 1944. *The Great Transformation: The Political and Economic Origins of Our Time*. New York, Toronto: Farrar & Rinehart.
- Porckess, Roger. 1991. *The HarperCollins Dictionary of Statistics*. New York: HarperCollins.
- Porter, Michael E. 1990. *The Competitive Advantage of Nations*. New York: Free Press.
- Porter, Roy. 1997. *The Greatest Benefit to Mankind*. New York: W. W. Norton & Company.
- Posner, Daniel N. 2004. "The Political Salience of Cultural Difference: Why Chewas and Tumbukas Are Allies in Zambia and Adversaries in Malawi." *American Political Science Review* 98, no. 4: 529–45.
- . 2005. *Institutions and Ethnic Politics in Africa*. Cambridge: Cambridge University Press.
- Powell, Robert. 1999. *In the Shadow of Power: States and Strategies in International Politics*. Princeton: Princeton University Press.
- Pratt, J. W., and Robert Schlaifer. 1984. "On the Nature and Discovery of Structure." *Journal of the American Statistical Association* 79, no. 385 (March): 9–21.
- PS: *Political Science & Politics*. 1995. "Symposium on Verification/Replication." In PS: *Political Science & Politics* 28, no. 3 (September): 443–99.
- Przeworski, Adam. 1995. Contribution to a symposium on "The Role of Theory in Comparative Politics." *World Politics* 48, no. 1 (October): 16–21.
- Przeworski, Adam, and Fernando Limongi. 1993. "Political regimes and economic growth." *Journal of Economic Perspectives* 7, no. 3: 51–69.
- Przeworski, Adam, and Henry Teune. 1970. *The Logic of Comparative Social Inquiry*. New York: Wiley.

- Przeworski, Adam, Michael Alvarez, Jose Antonio Cheibub, and Fernando Limongi. 2000. *Democracy and Development: Political Institutions and Well-Being in the World, 1950–1990*. Cambridge: Cambridge University Press.
- Putnam, Robert D. 1993. *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton: Princeton University Press.
- Ragin, Charles C. 1987. *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*. Berkeley: University of California Press.
- . 2000. *Fuzzy-Set Social Science*. Chicago: University of Chicago.
- . 2004. "Turning the Tables: How Case-Oriented Research Challenges Variable-Oriented Research." In Henry E. Brady and David Collier, eds., *Rethinking Social Inquiry*. Lanham, Md.: Rowman & Littlefield. Available online—see Preface.
- Ragin, Charles C., and Howard S. Becker. 1992. *What Is a Case? Exploring the Foundations of Social Inquiry*. New York: Cambridge University Press.
- Rasch, Georg. 1980 [1960]. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: University of Chicago Press.
- Rihoux, Benoît. 2006. "Two Methodological Worlds Apart? Praises and Critiques from a European Comparativist." *Political Analysis* 14, no. 3: 332–35.
- Roberts, Fred S. 1976. *Discrete Mathematical Models, with Applications to Social, Biological, and Environmental Problems*. Englewood Cliffs, N.J.: Prentice-Hall.
- Rock, Stephen R. 1989. *Why Peace Breaks Out: Great Power Rapprochement in Historical Perspective*. Chapel Hill: University of North Carolina Press.
- . 2000. *Appeasement in International Politics*. Lexington: University Press of Kentucky.
- Rodrik, Dani. 1999. "Democracies pay higher wages." *Quarterly Journal of Economics* CXIV: 707–38.
- Roe, Daphne A. 1973. *A Plague of Corn*. Ithaca, N.Y.: Cornell University Press.
- Rogowski, Ronald. 1995. "The Role of Theory and Anomaly in Social-Scientific Inference." *American Political Science Review* 89, no. 2 (June): 467–70.
- Rosenbaum, Paul R. 1984. "From Association to Causation in Observational Studies: The Role of Tests of Strongly Ignorable Treatment Assignment." *Journal of the American Statistical Association* 79, no. 385 (March): 41–48.
- . 2002. *Observational Studies*. 2nd edition. New York: Springer.
- Rosenberg, Morris. 1968. *The Logic of Survey Analysis*. New York: Basic Books.
- Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66, no. 5 (October): 688–701.
- . 1977. "Assignment to Treatment on the Basis of a Covariate." *Journal of Educational Statistics* 2: 1–26.
- . 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *Annals of Statistics* 6, no.1 (January): 34–58.
- . 1980. "Discussion of 'Randomization Analysis of Experimental Data: The Fisher Randomization Test' by D. Basu." *Journal of the American Statistical Association* 75, no. 351 (September): 575–82.
- . 1990. "Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies." *Statistical Science* 5, no. 4 (November): 472–80.
- Rueschemeyer, Dietrich, Evelyne Huber Stephens, and John D. Stephens. 1992. *Capitalist Development and Democracy*. Chicago: University of Chicago Press.

- Russell, Bertrand. 1969. *The Autobiography of Bertrand Russell, 1914–1944*. New York: Bantam Books.
- Sagan, Scott. 1993. *The Limits of Safety: Organizations, Accidents, and Nuclear Weapons*. Princeton: Princeton University Press.
- Salmon, Wesley C. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.
- Sartori, Giovanni. 1970. "Concept Misformation in Comparative Politics." *American Political Science Review* 64, no. 4 (December): 1033–53.
- Scharfstein, Daniel O., Andrea Rotnitzky, and James M. Robins. 1999. "Adjusting for Non-Ignorable Dropout Using Semi-Parametric Non-Response Models." *Journal of the American Statistical Association* 94: 1096–146.
- Schrodt, Philip A. 2006. "Beyond the Linear Frequentist Orthodoxy." *Political Analysis* 14, no. 3: 335–39.
- Schultz, Kenneth A. 2001. *Democracy and Coercive Diplomacy*. New York: Cambridge University Press.
- Schwandt, Thomas A. 1997. *Qualitative Inquiry: A Dictionary of Terms*. Thousand Oaks, Calif.: Sage.
- Sciubba, Roberto, and Rossana Sciubba Pace. 1976. *Le comunità di base in Italia*. Rome: Coines.
- Scriven, Michael. 1975. "Causation as Explanation." *Nous* 9: 3–10. *Methods and Research* 20, no. 4: 428–55.
- Seawright, Jason. 2002a. "Testing for Necessary and/or Sufficient Causation: Which Cases Are Relevant?" *Political Analysis* 10, no. 2: 178–93.
- . 2002b. "What Counts as Evidence? Reply." *Political Analysis* 10, no. 2: 204–7.
- Sekhon, Jasjeet S. 2009. "Opiates for the Matches: Matching Methods for Causal Inference." *Annual Review of Political Science* 12: 487–508.
- Sekhon, Jasjeet S., and Rocio Titiunik. 2009. "Redistricting and the Personal Vote: When Natural Experiments Are Neither Natural Nor Experiments." Working paper, Department of Political Science, UC Berkeley.
- Simmelweis, Ignaz. 1981 [1860]. *The Etiology, Concept, and Prophylaxis of Childbed Fever*. English translation by F. P. Murphy. Birmingham: Classics of Medicine Library.
- Shafer, Michael. 1988. *Deadly Paradigms: The Failure of U.S. Counterinsurgency Policy*. Princeton: Princeton University Press.
- Shen, Jian-Guang. 2002. "Democracy and growth: An alternative empirical approach." BOFIT Discussion Papers No. 13. Helsinki: Bank of Finland Institute for Economics in Transition.
- Shepard, Lorrie. 1993. "Evaluating Test Validity." *Review of Research in Education* 19: 405–50.
- Sherman, Lawrence, and Heather Strang. 2004. "Experimental Ethnography: The Marriage of Qualitative and Quantitative Research." *The Annals of the American Academy of Political and Social Sciences* 595, 204–22.
- Shugart, Matthew, Erika Moreno, and Luis E. Fajardo. 2001. "Deepening Democracy through Renovating Political Practices: The Struggle for Electoral Reform in Colombia." Paper presented at the conference on Democracy, Human Rights, and Peace in Colombia, Kellogg Institute, University of Notre Dame, March 26–27.

- Shultz, Kenneth S., Matt L. Riggs, and Janet L. Kottke. 1998. "The Need for an Evolving Concept of Validity in Industrial and Personnel Psychology: Psychometric, Legal, and Emerging Issues." *Current Psychology* 17 (Winter): 265–86.
- Sil, Rudra. 2000. "The Division of Labor in Social Science Research: Unified Methodology or 'Organic Solidarity'?" *Polity* 32, no. 4 (Summer): 499–531.
- Simon, Herbert A., and Y. Iwasaki. 1988. "Causal Ordering, Comparative Statistics, and Near Decomposability." *Journal of Econometrics* 39: 149–73.
- Sirowy, Larry, and Alex Inkeles. 1991. "The effects of democracy on economic growth and inequality: A review." In Alex Inkeles, ed., *On Measuring Democracy: Its Consequences and Concomitants*. New Brunswick: Transaction Books.
- Skocpol, Theda. 1979. *States and Social Revolutions: A Comparative Analysis of France, Russia, and China*. New York: Cambridge University Press.
- Smith, Jeffrey A., and Petra E. Todd. 2005. "Does matching overcome LaLonde's critique of nonexperimental estimators?" *Journal of Econometrics* 125, no. 1: 305–53.
- Snow, John. 1965 [1855]. *On the Mode of Communication of Cholera*. London: John Churchill, New Burlington Street, England, 2nd edition. Reprinted in *Snow on Cholera*, London: Humphrey Milford: Oxford University Press.
- Snyder, Jack. 1984. *The Ideology of the Offensive: Military Decision Making and the Disasters of 1914*. Ithaca: Cornell University Press.
- . 1984/85. "Richness, Rigor, and Relevance in the Study of Soviet Foreign Policy." *International Security* 9, no. 3 (Winter): 89–108.
- . 1987/88. "The Gorbachev Revolution: A Waning of Soviet Expansionism?" *International Security* 12, no. 3 (Winter): 93–131.
- . 1990. "Averting Anarchy in the New Europe." *International Security* 14, no. 4 (Spring): 5–41.
- . 1991. *Myths of Empire: Domestic Politics and International Ambition*. Ithaca: Cornell University Press.
- Solow, Robert M. 1956. "A contribution to the theory of economic growth." *Quarterly Journal of Economics* 70 (February): 65–94.
- Sovey, Allison J., and Donald P. Green. 2009. "Instrumental Variables Estimation in Political Science: A Readers' Guide." Manuscript, Department of Political Science, Yale University.
- Stasavage, David. 2003. "Transparency, Democratic Accountability, and the Economic Consequences of Monetary Institutions." *American Journal of Political Science* 47, no. 3: 389–402.
- Steiger, James H. 2001. "Driving Fast in Reverse." *Journal of the American Statistical Association* 96, no. 453 (March): 331–38.
- Stein, Janice Gross. 1994. "Political Learning by Doing: Gorbachev as Uncommitted Thinker and Motivated Learner." *International Organization* 48, no. 2 (Spring): 155–83.
- Stevens, S. S. 1946. "On the Theory of Scales of Measurement." *Science* 103: 677–80.
- . 1951. "Mathematics, Measurement, and Psychophysics." In S. S. Stevens, ed., *Handbook of Experimental Psychology*. New York: Wiley.
- Stigler, Stephen M. 1986. *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge, Mass.: Belknap Press of Harvard University Press.
- Stock, James H., and Mark W. Watson. 2003. *Introduction to Econometrics*. Boston: Addison-Wesley.

- Stokes, Susan Carol. 2001. *Mandates and Democracy: Neoliberalism by Surprise in Latin America*. New York: Cambridge University Press.
- . 2009. "A Defense of Observational Research." Manuscript, Department of Political Science, Yale University.
- Stolzenberg, Ross M., and Daniel A. Relles. 1990. "Theory Testing in a World of Constrained Research Design: The Significance of Heckman's Censored Sampling Bias Correction for Nonexperimental Research." *Sociological Methods and Research* 18, no. 4: 395–415.
- Stone, Charles J. 1985. "Additive Regression and Other Nonparametric Models." *Annals of Statistics* 13, no. 2 (June): 689–705.
- Stone, Richard. 1993. "The Assumptions on Which Causal Inferences Rest." *Journal of the Royal Statistical Society* 55, no. 2 (Series B): 455–66.
- Stouffer, Samuel A., et al. 1949. *The American Soldier*. Princeton: Princeton University Press.
- Strauss, Anselm, and Juliet Corbin. 1994. "Grounded Theory Methodology: An Overview." In Norman K. Denzin and Yvonna S. Lincoln, eds., *Handbook of Qualitative Research*. Thousand Oaks, Calif.: Sage Publications.
- Sundquist, James L. 1973. *Dynamics of the Party System: Alignment and Realignment of Political Parties in the United States*. Washington, D.C.: The Brookings Institution.
- Tannenwald, Nina. 1999. "The Nuclear Taboo: The United States and the Normative Basis of Nuclear Non-Use." *International Organization* 53, no. 3 (Summer): 433–68.
- . 2005. "Ideas and Explanation: Advancing the Theoretical Agenda." *Journal of Cold War Studies*, 7, no. 2 (Spring): 13–42.
- Tarrow, Sidney G. 1988. "Old Movements in New Cycles of Protest: The Career of an Italian Religious Community." In B. Klandermans, Hanspeter Kriesi, and Sidney Tarrow, eds., *From Structure to Action: Comparing Social Movements Across Cultures*. International Social Movement Research Series, vol. 1. Greenwich, Conn.: JAI Press.
- . 1989. *Democracy and Disorder: Protest and Politics in Italy, 1965–1975*. New York: Oxford University Press.
- . 1994. *Power in Movement: Social Movements, Collective Action, and Politics*. New York: Cambridge University Press.
- . 1995. "Bridging the Quantitative-Qualitative Divide in Political Science." *American Political Science Review* 89, no. 2 (June): 475–81.
- Terris, Milton, ed. 1964. *Goldberger on Pellagra*. Baton Rouge: Louisiana State University Press.
- Thistlewaite, Donald L., and Donald T. Campbell. 1960. "Regression-discontinuity Analysis: An Alternative to the Ex-post Facto Experiment." *Journal of Educational Psychology* 51, no. 6 (December): 309–17.
- Thompson, E. P. 1963. *The Making of the English Working Class*. New York: Vintage Books.
- Tilly, Charles. 1990. *Coercion, Capital, and European States, 990–1990 A.D.* Cambridge, Mass.: Blackwell.
- . 1993. *European Revolutions, 1492–1992*. Oxford: Blackwell.
- . 1994. "States and Nationalism in Europe, 1492–1992." *Theory and Society* 23, no. 1: 131–46.

- . 2001. "Mechanisms in Political Processes." *Annual Review of Political Science*, vol. 4. Palo Alto: Annual Reviews.
- Timonius, Emanuel, and John Woodward. 1714. "An account, or history, of the procuring the small pox by incision, or inoculation; as it has for some time been practised at Constantinople." *Philosophical Transactions* 29: 72–82.
- Tintner, Gerhard. 1952. *Econometrics*. New York: Wiley.
- Titiunik, Rocío. 2008. "Drawing Your Senator From a Jar: Term Length and Legislative Behavior." Working paper, Department of Political Science, University of Michigan.
- Truman, David Bicknell. 1951. *The Governmental Process: Political Interests and Public Opinion*. New York: Knopf.
- van Deth, Jan W. 1995. "Comparative Politics and the Decline of the Nation-State in Western Europe." *European Journal of Political Research* 27: 443–62.
- Van Evera, Stephen. 1997. *Guide to Methods for Students of Political Science*. Ithaca: Cornell University Press.
- Verba, Sidney, Kay Lehman Schlozman, and Henry E. Brady. 1995. *Voice and Equality: Civic Voluntarism in American Politics*. Cambridge, Mass.: Cambridge University Press.
- Vinten-Johansen, Peter, Howard Brody, Nigel Paneth, Stephen Rachman, and Michael Rip. 2003. *Cholera, Chloroform, and the Science of Medicine*. New York: Oxford University Press.
- Vogt, W. Paul. 1999. *Dictionary of Statistics & Methodology: A Nontechnical Guide for the Social Sciences*. 2nd edition. Thousand Oaks, Calif.: Sage Publications.
- Walker, Henry A., and Bernard P. Cohen. 1985. "Scope Statements: Imperatives for Evaluating Theory." *American Sociological Review* 50, no. 3 (June): 288–301.
- Wallerstein, Immanuel Maurice. 1974. *The Modern World-System*, vol. 1. New York: Academic.
- Wallerstein, Michael. 2000. "Trying to Navigate Between Scylla and Charybdis: Mismatched and Unidentified Models in Comparative Politics." *APSA-CP: Newsletter of the APSA Comparative Politics Section* 11, no. 2 (Summer): 1–2, 4, 21.
- . 2001. "Bridging the Quantitative/Non-Quantitative Divide." *APSA-CP: Newsletter of the APSA Comparative Politics Section* 12, no. 2 (Summer): 1–2, 23.
- Walsh, Christopher. 2003. *Antibiotics: Actions, Origins, Resistance*. Washington, D. C.: ASM Press.
- Walt, Stephen M. 1996. *Revolution and War*. Ithaca: Cornell University Press.
- Weber, Max. 1949. *The Methodology of the Social Sciences*. Glencoe, Ill.: Free Press.
- Weber, Steven. 1991. *Cooperation and Discord in U.S.-Soviet Arms Control*. Princeton: Princeton University Press.
- Webster, William S. 1998. "Teratogen update: Congenital rubella." *Teratology* 58: 13–23.
- Weyland, Kurt. 2005. "Book Review." *Perspectives on Politics* 3, no. 2 (June): 392–93.
- White, Halbert. 1994. *Estimation, Inference, and Specification Analysis*. Cambridge: Cambridge University Press.
- Wickham-Crowley, Timothy P. 1992. *Guerrillas and Revolution in Latin America: A Comparative Study of Insurgents and Regimes Since 1956*. Princeton: Princeton University Press.

- Wohlforth, William. 1994/95. "Realism and the End of the Cold War." *International Security* 19, no. 3 (Winter): 91-129.
- Wonnacott, Ronald J., and Thomas H. Wonnacott. 1979. *Econometrics*. New York: Wiley.
- Wooldridge, Jeffrey M. 2009. *Introductory Econometrics*. Cincinnati: South-Western College.
- World Bank. 1998. *World Development Indicators*. Washington, D.C.: World Bank.
- Zelikow, Philip, with Condoleezza Rice. 1995. *Germany Unified and Europe Transformed: A Study in Statecraft*. Cambridge: Harvard University Press.

Acknowledgment of Permission to Reprint Copyrighted Material

In the present volume, chapters 1, 2, 8, 9, and 12 are entirely original material, and these chapters appeared in the first edition of *Rethinking Social Inquiry*. Chapter 14 is new material and is original to this edition. Chapters 3, 4, 5, 6, 7, 10, 11, and 13 are to varying degrees reprints of earlier articles; some have new summary tables, revised headings and/or new titles, and some have been heavily rewritten.

We wish to thank several publishers for permission to reprint:

Earlier versions of Chapters 3 and 4 were originally published as Larry M. Bartels, "Symposium on *Designing Social Inquiry*, Part 1," and Henry E. Brady, "Symposium on *Designing Social Inquiry*, Part 2: Doing Good and Doing Better," both in *The Political Methodologist* 6, no. 2 (Spring 1995): 8–11 and 11–19, copyright American Political Science Association, reprinted with permission of the APSA's Society for Political Methodology.

Earlier versions of Chapters 5, 6, and 7¹ were originally published as Ronald Rogowski, "The Role of Theory and Anomaly in Social-Scientific Inference," Sidney Tarrow, "Bridging the Quantitative-Qualitative Divide in Political Science," and Gary King, Robert O. Keohane, and Sidney Verba, "The Importance of Research Design in Political Science," *American Political Science Review* 89, no. 2 (June 1995): 467–70, 471–74, and 475–81, copyright American Political Science Association, reprinted with permission of Cambridge University Press.

1. These three chapters were part of the symposium on King, Keohane, and Verba (1994), published in the *American Political Science Review* and edited by Mark Lichbach. Two additional, valuable contributions to this symposium are Laitin (1995) and Caporaso (1995) are cited at various points in the present volume.

Chapter 10 is a slightly abridged version of David A. Freedman, "On Types of Scientific Enquiry," and Chapter 11 is an extensively rewritten version of Andrew Bennett, "Process Tracing: A Bayesian Perspective," *The Oxford Handbook of Political Methodology*, edited by Janet Box-Steffensmeier, Henry E. Brady, and David Collier, Oxford University Press (2008): 300–18 and 702–721, reprinted with permission of Oxford University Press.

Chapter 13 is an extensively rewritten version of Jason Seawright, "Democracy and Growth," *Regimes and Democracy in Latin America*, edited by Gerardo Munck, Oxford University Press (2007), reprinted with permission of Oxford University Press.

Chapters 6, 7, 8, and 9 from the first edition of *Rethinking Social Inquiry* are available online, posted on the Rowman & Littlefield website and accessible with the instructions on the copyright page of this second edition. Permissions for this online posting are indicated on the website.

Subject Index

- anomaly, theoretical. *See* theory
as-if randomization, 277–80, 280–81, 282–83 (table 14.1), 284, 285, 286, 287, 288, 289, 300, 301, 314; plausibility of, 292–300, 293n27, 297–98n39, 298n40, 298n42, 302, 303–4, 306–7, 308, 309
- assumptions, 16, 18, 22, 23, 27, 28, 40–49, 52, 54, 59, 179, 314; causal homogeneity, 25, 41–43, 52, 88, 142, 316; conditional independence, 44–49, 74–76, 164, 172–77, 320; constant causal effects, 41n10, 321; independence of observations, 41, 41n8, 43–44, 43n13, 332; specification assumption, 172–77, 262, 263, 264, 351; unit homogeneity, 41–42, 41n9, 42n11, 357
- autocorrelation, 41n8, 82, 314
- Bayesian inference, 150, 166–67, 314
- bias. *See* error
- Boolean algebra, 315
- case evidence: Cold War, end of, 214–19; Eijkman and Beriberi, 227–29; and experiments, 308–10; Fashoda Crisis, 211–12; Fleming and Penicillin, 231; Goldberger and Pelagra, 229–31; Jenner and vaccination, 222–23; Semmelweis and Puerperal Fever, 223–25; World War I, end of, 212–14. *See also* cholera, Snow on; process tracing
- cases, 52, 92n6, 182, 315; cases versus observations, 52, 182–84
- case selection, 34–35, 51–53, 118–20, 315; selecting on the dependent variable, 93, 94, 95n8, 116–18, 120–21, 349
- case studies, 21, 28, 91–97, 116–18, 315–16; case-oriented versus variable-oriented research, 315; contrasting cases, 10, 20, 321; cross-case analysis, 24, 128n2, 141, 143, 323; crucial case, 323; deviant case, 146, 326; least-likely case, 90n2, 242, 335; matching cases, defined, 336; most-likely case, 21, 91, 131, 339; negative case, 340; positive case, 61, 343; within-case analysis, 10, 21, 24, 129, 141, 143, 155, 197, 358. *See also* case evidence; constructivism; interpretation; qualitative and quantitative methods: critique of subordinating qualitative methods to norms of mainstream quantitative methods
- causality, 70–76; causal effect, 6, 7, 38, 44, 45, 52, 72, 73n10, 74, 76, 95n8,

- 117, 142–43, 173, 176, 258n16, 263, 265, 266, 290, 296, 300, 316; causal heterogeneity, 43, 63, 316; causal homogeneity (*see* assumptions); causal inference (*see* inference); causal mechanism, 72, 104 (table 6.1), 185, 317; causal model, 6, 36, 41–42, 145, 146, 149, 257 (figure 13.2), 317, 350–51; causal process, 43, 150, 151n20, 158, 318; causal-process observations (*see* observations); causal sequence, defined, 318; causation, multiple and conjunctural, 318; constant causal effects (*see* assumptions). *See also* causality, tests for
- causality, tests for, 210–11; doubly decisive test, 210 (table 10.1), 211, 326–27; hoop test, 210, 210 (table 10.1), 211, 212, 216, 217, 218–19, 331; smoking gun test, 210–11, 210 (table 10.1), 212, 213–14, 350; straw-in-the-wind test, 210 (table 10.1), 211, 217, 354
- cause, 145–52, 318; counterfactual definition of causation (*see* counterfactual analysis); deterministic cause, 37, 145–52, 326; necessary cause, 145–52, 147 (table 8.1), 340; probabilistic cause, 145–52, 147 (table 8.1), 151n20, 343; sufficient cause, 145–52, 147 (table 8.1), 354
- cholera, Snow on, 225–27, 282
- Cold War, end of. *See* case evidence
- comparative method, 10, 15n1, 319; comparative-historical analysis, 106, 391; contextualized comparison, 321; critical juncture, 323; method of agreement, 21, 146, 337, 338n9; method of difference, 21, 146, 338; path dependence, 343; Qualitative Comparative Analysis, 344; tipping point, 104 (table 10.1), 105, 355; within-case control, 359. *See also* cases
- concepts, 20, 76, 319–20; conceptual stretching, 135, 320; contrast space, 321; extension of concepts, 153n21, 330; formation of concepts, 76, 132–40, 320; intension of concepts, 153n21, 333; terms, defined, 354
- conditional independence. *See* assumptions
- confounders. *See* variables
- constant causal effects. *See* assumptions
- constructivism, 321. *See also* interpretation
- context, 20, 21, 24, 25, 53, 54, 60, 134–35, 139–40, 178, 278, 300, 321. *See also* observations: causal process observations
- counterfactual analysis, 71, 322; counterfactual definition of causation, 25, 38, 44–45, 163, 276n4, 322–23
- creativity in research, 130–31
- critical juncture. *See* comparative method
- cross-case analysis. *See* case selection
- cross-sectional analysis, defined, 323
- data, defined, 323; data mining, 55, 62, 170–72, 298, 324; data point, defined, 324; data set, defined, 324; data-set observations (*see* observations); missing data, 8, 24, 195–96, 263; piece of data, defined, 324
- deductive analysis, defined, 324
- dependent variable. *See* variables
- description, defined, 325
- descriptive inference. *See* inference
- determinate research design. *See* research design
- deterministic cause. *See* cause
- diffusion, 44, 109, 326
- econometrics, 6n5, 16n2, 62, 76, 138, 175, 288n23, 327
- Eijkman and Beriberi. *See* case evidence
- empirical, defined, 327
- error, defined, 327–28; bias, 39, 45, 76, 87, 148, 211, 254, 254n12, 314; complexification based on extreme cases, defined, 319; measurement

- error, 54, 60, 76, 80n14, 86–87, 93, 114, 132, 134, 138, 148–49, 151n20, 254n12, 337; random error, 7, 169, 281, 281n11, 346; sampling error, 38, 82n17, 149, 167, 169, 348; selection bias, 40, 52, 69, 82n17, 114, 118–21, 140–45, 140n11, 149, 155, 349; standard error, 7, 43, 275, 296–97n36, 298–99n39; standard error in experiments, 275, 299–300n44, 300n45; systematic error, defined, 354
- ethnographic research, defined, 328
- experiments, 16n2, 23, 73, 74, 75–76, 114n2, 162–65, 162n2, 173, 274, 329; control group, 322; experimental data, defined, 329; lab (true) experiments, 306–7. *See also as-if* randomization; natural experiments; quasi-experiments; randomization
- explanation, 70–76, 70n5, 77, 118, 227, 329; rival explanations, 161n1, 162, 168, 170, 177, 191, 193, 197, 207–8, 209–10, 211–19
- falsifiability, 50–51, 157; falsifiable, defined, 330
- Fashoda Crisis. *See* case evidence
- field research, defined, 330
- Fleming and Penicillin. *See* case evidence
- goals, 21, 129, 330; intermediate, 39–40, 154–55, 330; overarching, 39, 153–54, 330–31
- Goldberger and Pellagra. *See* case evidence
- hermeneutics, defined, 331. *See also* interpretation
- hypotheses and theory, refinement of, 8, 25, 57, 127–30
- hypothesis, defined, 331; and *ex post facto* formation, defined, 329; iterated refinement of, 335; null hypothesis, defined, 341
- independence of observations. *See* assumptions
- indeterminate research design. *See* research design
- inductive analysis, defined, 333; elaboration model, 131, 327
- inference, 18, 35, 90, 125–59 *passim*, 333; causal inference, 2, 6, 9, 10, 20–21, 22–26, 28, 35n2, 37–8, 41n8, 50 (figure 2.1), 54–56, 59, 60, 63, 70–71n5, 72n7, 73, 77, 104 (table 6.1), 119–20, 128–29, 133, 138–40, 153–56, 161–99 *passim*, 317; descriptive inference, 35n2, 36–37, 50 (figure 2.1), 53–44, 70, 70n5, 119–20, 128–29, 137n9, 140, 153–56, 197–98, 325–26; design-based, 275, 277–81, 289, 292; evaluating based on interpretability, 157, 170, 197, 334; fundamental problem of causal inference, 37, 44, 114, 317; logical foundations of causal inference, 28–36, 56–62, 229–66; inferential leverage, 11, 26, 27 (table 2.1), 174 (table 10.1), 218–20, 229–66, 291; model-based, 275, 277–81, 289, 291, 306; nested inference, 181–82, 192, 340
- instrumental variables. *See* natural experiments
- interpretation, 7, 53, 62–63, 109, 186, 334; and *Verstehen*, defined, 358. *See also* constructivism; thick description; thick versus thin analysis
- Jenner and vaccination. *See* case evidence
- large N. *See* N (number of observations)
- level of analysis, 24, 55, 121, 157, 178, 208, 268, 335
- level of measurement, 177–78, 335–6, 344–45; nominal, 133, 155; ordinal, 137
- leverage. *See* inference
- LISREL, defined, 336. *See also* statistical modeling
- longitudinal analysis, defined, 336

- mainstream quantitative methods, 15–17, 33–63, 148–52, 161–77, 186n30, 187, 191–92, 295, 336; limitations of, 3–9, 16, 22–26, 67–82, 83–88, 162–65, 240–41, 247–71; strengths of, 17–18, 33–59, 69–70, 101–10 *passim*, 111–22 *passim*, 129–30, 148–49, 177–82. *See also* qualitative and quantitative methods
- matching. *See* research design
- measurement, 76–81, 132–40, 337; indicator, defined, 333; measurement error (*see* error); measurement model, defined, 337; measurement theory, defined, 337; measurement validity, defined, 337; operationalization, defined, 342; reliability, defined, 347; score, defined, 348; subtype, defined, 354; value, defined, 358. *See also* level of measurement; psychometrics; typology
- model/modeling, defined, 339. *See also* Rubin-Holland model; statistical modeling; theory
- multi-method research, defined, 339
- N (number of observations), 169, 178–79, 339; increasing the N, 87–88, 121–12, 134, 135, 157–58, 183–84, 192–95, 325; large N, 2, 17, 22, 51n21, 59–60, 68n1, 72, 106, 122, 155, 162, 164–65, 178–79, 179 (table 9.1), 189, 191, 196, 233, 248, 335; small N, 10, 22, 40, 52, 55, 60, 68n1, 121, 157, 162, 164–65, 167, 178–79, 179 (table 9.1), 183–84, 189, 190, 209, 350
- natural experiments, 9, 162n3, 204, 231n3, 276, 277–78, 280–86, 282–83 (table 14.1), 291, 305–6, 305 (figure 14.4), 307, 340; instrumental-variables (IV) designs, 282–83 (table 14.1), 288–89, 289nn21–22, 308, 333; regression-discontinuity (RD) designs, 282–83 (table 14.1), 286–88, 286n17, 288n20, 298n40, 307–8, 346–47; treatment in, 290n24, 301, 302n48, 303–4; versus matching, 289–91. *See also as-if* randomization; randomization
- necessary cause. *See* cause
- Neyman-Rubin-Holland model, 163–64, 204, 276, 276n4, 277n7, 293n34, 340–41. *See also* Rubin-Holland model
- no-variance designs. *See* research design
- observable implications, 50–51, 52, 55, 87, 113, 115, 117, 121, 121n6, 122, 156, 208, 341
- observational data, 2, 16n2, 20, 22–23, 28, 44, 46, 62, 151, 162–65, 169, 171, 172–73, 174, 221, 232, 240, 259, 281, 289, 292, 306, 341
- observational studies, 162–65, 341
- observations, 182–84, 341; causal-process observations, 2, 2n3, 24–25, 141, 184–96, 202n2, 221–36, 237–42, 318 (*see also* process tracing); data-set observations, 2n3, 24, 183n25, 184–96, 202n2, 237–42, 324; rectangular data set, 24, 182–3, 203, 346; increasing the number of observations (*see* N (number of observations))
- observations versus cases, 51–53, 182–84. *See also* qualitative and quantitative methods: critique of subordinating qualitative methods to norms of mainstream quantitative methods
- outlier, defined, 342. *See also* case studies: deviant case
- probabilistic, defined, 343
- probabilistic cause. *See* cause
- probability theory, 39 (table 2.1), 343
- process tracing, 29, 88, 103–4, 104 (table 6.1), 201–2, 207–19, 221–36, 237–42, 243, 343–44. *See also* observations: causal-process observations; qualitative and quantitative methods: critique of subordinating quali-

- tative methods to norms of mainstream quantitative methods; typology
- psychometrics, 16n2, 20, 135–40 *passim*, 166, 198, 344
- qualitative and quantitative methods, 177–82; critique of subordinating qualitative methods to norms of mainstream quantitative methods, 186–87, 191–96; linking qualitative and quantitative research, 29–30, 85–6, 101–110 *passim*, 111–122 *passim*, 184–96, 306–8; qualitative methods, 10, 15–17, 20–26, 29, 63, 77, 85–86; qualitative versus quantitative methods, 68n1, 177–82, 191–96, 344–45; quantitative template (*see* mainstream quantitative methods). *See also* statistical rationale for qualitative research
- quantitative methods. *See* mainstream quantitative methods; qualitative and quantitative methods
- quasi-experiments, 75, 162–63, 162n3, 345–46; interrupted time-series design, 105, 163, 281–82n14, 334
- randomization, defined, 346; random assignment, 46, 75, 114n2, 164–65, 281n14, 290, 295n33, 304–5, 342n11 (*see also as-if* randomization); random sample, defined, 346
- regression analysis, 5–8, 15–16, 16n2, 18, 23, 38, 43, 46n17, 55, 62–63, 76, 139, 141–43, 142n13, 148, 149, 155, 170, 171, 171n15, 175, 175–76n19, 186, 188, 203–4, 238, 241–42, 247–71 *passim*, 273–75, 277n6, 278–80, 289, 291, 296, 296–97n36, 297, 298n41, 299, 299nn43–44, 306, 346; interaction term, defined, 334. *See also* mainstream quantitative methods: limitations of; statistical modeling
- regression discontinuity. *See* natural experiments
- replication, 56, 347
- research cycle, 39–50, 347
- research design, 111–22 *passim*, 194–95, 273–76, 299, 347; determinate design, 55n23, 326; determinate versus indeterminate designs, 51–52, 168–70; indeterminate design, 55, 332–33; matching design, 290–91n25, 336–37 (*see also* natural experiments: versus matching); no-variance design, 53, 128–29, 131, 141, 142–43, 147 (table 8.1), 149–51, 341; strong versus weak, 279–80, 291, 296, 306–8; trade-offs involved in, 304, 307. *See also* cholera, Snow on; research design: determinate versus indeterminate designs
- research problem/question, 50, 347
- research program, defined, 347
- rethinking social inquiry, 1–10; alternative methodological tools, 20–21; toward an alternative view of methodology, 22–26. *See also* inference: evaluating based on interpretability; mainstream quantitative methods; qualitative and quantitative methods; research design: determinate versus indeterminate designs; statistical rationale for qualitative research; standards, shared
- Rubin-Holland model, 42nn11–12, 72–74. *See also* Neyman-Rubin-Holland model
- sample, defined, 347–48; sampling error (*see* error)
- scientific inquiry, types of, 221–36. *See also* observations
- scientific research, 34–35, 116–18, 161n1, 221–36, 348
- scope conditions/restrictions, 24–25, 348
- selection bias. *See* error
- Semmelweis and Puerperal Fever. *See* case evidence
- shared standards. *See* standards, shared
- small N: and method of controlled

- comparison, 322. *See also* N (number of observations)
- specification assumption. *See* assumptions
- standard error. *See* error
- standards, shared, 19–22, 26, 30–31, 349–50; technification and the quest for, 198–99. *See also* goals; trade-offs
- statistical modeling: correlation, 149, 176, 255, 257 (figure 13.2c), 274, 322; covariance structure models, 136, 138, 138n10, 323; credibility of, 247–271 passim, 274, 278–79, 304; degrees of freedom, 184, 192–93, 193nn35–36, 194, 209–10, 299n44, 325; difference-of-means test, 280n10, 282–3 (table 14.1), 296, 296–97n36, 297, 297n38, 300, 300n45; efficiency, 39n5, 114, 150, 150n19, 327; endogeneity, 55–56, 59, 60, 62, 75, 168n12, 176, 254, 254n12, 267n22, 327; estimation, defined, 328; estimator, defined, 328; expected value, defined, 329; heteroskedasticity, defined, 331; identifiability, 168, 168n12, 170, 331; identification, defined, 331; identification problem, defined, 331–32; multicollinearity, 51n21, 55, 130, 157, 168, 168nn10–11, 263–64, 339; parameter, defined, 342; parameter estimation, defined, 342–43; power of a statistical test, 61, 169, 170, 343; significance tests, 7–8, 18, 23, 167, 171n15, 350; slope, standardized and unstandardized, defined, 352, 357; specification search, 25, 62, 170–72, 351–52; spurious correlation, 46–47, 97, 175–76n19, 209, 352; statistical tests as a criterion for distinguishing quantitative from qualitative research, 180; underspecified model, defined, 356–57. *See also* assumptions; econometrics; error; mainstream quantitative methods; qualitative and quantitative methods; regression analysis; uncertainty
- statistical rationale for qualitative research, 16, 26, 197–98
- statistical theory, 16, 26, 86, 142, 165–68, 186n30, 352–53. *See also* statistical rationale for qualitative research
- stochastic, defined, 353
- stratification, defined, 353–54
- strengths of mainstream quantitative methods. *See* qualitative and quantitative methods, mainstream quantitative methods
- sufficient cause. *See* cause
- theory, 67–82 passim, 83–88, 354; concreteness, 133, 320; parsimony, defined, 343; refinement of theory (*see* hypotheses and theory, refinement of); specifying the theory, 50–51; theoretical anomaly, 89–98. *See also* concepts; statistical theory
- thick description, 180, 180nn22–23, 355
- thick versus thin analysis, 153n21, 180–81, 184, 237–42, 355. *See also* observations
- thin analysis. *See* thick versus thin analysis
- thought experiments, 355. *See also* counterfactual analysis
- time-series analysis, defined, 355
- tools, 19–22, 153–56, 355–56; for bridging the quantitative-qualitative divide, 104 (table 6.1); for qualitative research, 29, 58 (table 2.2), 62–63; quantitative tools employed in KKV, summary of, 39 (table 2.1). *See also* case studies; comparative method; statistical modeling
- trade-offs, 21, 22, 26, 61, 131–33, 153–59, 198, 276, 291, 304, 307, 356; trade-offs in KKV, 156. *See also* research design: trade-offs involved in
- triangulation, 104 (table 6.1), 108–10, 121–22, 356

- truncation, 141–42, 143–44, 149, 356
- typology, 134, 202, 204, 356; and classification, 319–20; of natural experiments, 275–76, 276n4, 303–6; of tests in process tracing, 210–11, 210 (table 10.1)
- uncertainty, 6, 9, 11, 15, 23, 26, 41, 43, 53–55, 71, 183, 185, 234–35, 237, 311
- unit homogeneity. *See* assumptions
- units of analysis, 37, 158, 178n20, 182, 357
- universe of cases, 22n1, 23–24, 107–12, 146, 155, 210, 211, 215, 218, 234, 311–12
- validity, 54, 133, 138, 139–40, 357; Campbell's checklists of threats to, 75, 163, 315; external validity, 119n5, 162n2, 301, 330; internal validity, 75, 77, 119n5, 162n2, 334
- variables, *passim*, 312; antecedent, defined, 313; confounder, 252–53, 254–59 *passim*, 257 (figure 13.2), 261, 265, 277, 278, 280, 284, 285, 286, 289, 290, 291, 292, 295n33, 296, 297, 299, 320; control, defined, 322; dependent, defined, 325; and dichotomy, 326; endogenous, defined, 327; exogenous, defined, 328; independent, defined, 332; intervening, defined, 334; latent, defined, 335; missing (omitted), 28, 59, 93, 119n5, 175–76n19, 193, 194, 252, 253n10, 266, 338; missing (omitted) variable bias, 46n17, 55, 56, 87, 114, 176, 258, 277, 288, 338–9; outcome variable, defined, 342
- variable-oriented research, defined, 358
- within-case analysis. *See* case selection
- World War I, end of. *See* case evidence

Name Index

- Acemoglu, Daron, 289n21
Achen, Christopher H., 6, 75, 86, 87,
166n6, 168n12, 175n18, 199n39,
262, 273, 274, 281n14
Achinstein, Peter, 71n6
Adcock, Robert, 135, 135n4, 137n8,
139
Allen, William Sheridan, 30, 91, 92,
92n6, 93, 95, 116, 118, 128
Alvarez, R. Michael, 43, 144n15
Andrich, David, 79
Angrist, Joshua D., 287, 292 (figure
14.1), 293, 296 (figure 14.2), 298,
298nn40–41, 301 (figure 14.3)
Ansolabehere, Stephen, 137, 282 (table
14.1)
Arceneaux, Kevin, 290
Arendt, Hannah, 92
Arrow, Kenneth J., 84

Babbie, Earl, 131
Bagehot, Walter, 138
Bailey, Kenneth, 139
Banerjee, Abhijit, 282 (table 14.1),
302n52
Baron, John, 234
Barro, Robert J., 250, 253, 255, 260
(table 13.3), 267
Bartels, Larry, 7, 22n3, 28, 58, 59, 60,
61, 86, 88, 125, 127, 133, 134, 138,
152, 165, 166n6, 179, 264

Bates, Robert H., 30, 94–96, 96n10,
119, 120–21, 121n6, 127, 128,
128n2
Baum, Matthew A., 250n8, 260 (table
13.3), 269
Bayles, Kenneth W., 236
Bazin, Herve, 233–34
Beck, Nathaniel (Neal), 4, 44, 248n2
Becker, Howard S., 182
Bendix, Richard, 195
Bennett, Andrew, 10, 201, 202, 207,
207n1, 211, 214n4, 215n5, 216,
217, 218, 237, 344
Berelson, Bernard R., 86
Berger, Daniel, 282 (table 14.1),
302n52
Berk, Richard, 6, 7n7, 8, 166n6, 232,
290
Berry, Gregory, 283 (table 14.1), 294,
294nn30–31, 298, 301, 302
Blalock, Hubert M., 153n21, 154, 157
Blattman, Christopher, 282 (table 14.1)
Boix, Charles, 134
Bollen, Kenneth, 77, 80–81, 134, 136,
138, 138n10
Boruch, Robert F., 163
Bourguignon, Francois, 254n13
Brady, Henry E., 10, 17, 27–28, 30–31,
44n14, 47, 58, 59, 60, 61, 86, 125,
127, 132, 133–34, 134n3, 137, 152,

- 163, 164, 188–89, 203, 203n4, 207, 221, 232, 262, 274, 278, 281, 282 (table 14.1), 292 (figure 14.1), 293, 293n27, 296 (figure 14.2), 297, 301, 301 (figure 14.3), 302, 309
- Braumoeller, Bear F., 148–49, 150, 151
- Bristowe, John S., 233
- Brody, Baruch A., 71n6, 72n8
- Brooks, Stephen, 214, 215, 216, 217
- Brunell, Thomas L., 283 (table 14.1)
- Brunetti, Aymo, 250n7
- Buck, Carol, 233
- Budd, William, 233
- Bulloch, William, 234
- Bunce, Valerie, 108
- Burkhart, Ross E., 250n6
- Cain, Bruce, 72
- Campbell, Donald T., 75, 75n13, 76, 77, 78, 80, 114, 114n2, 131, 162, 162n3, 163, 166, 166n7, 209–10, 265, 281n14, 286–87, 286n17, 287n18, 301, 315, 315n2, 345, 356n21
- Campbell, James F., 179
- Cann, C. L., 166n6
- Caporaso, James A., 111n1, 113, 114, 114n2, 119n5, 163, 165
- Card, David, 282 (table 14.1), 292 (figure 14.1), 293, 296 (figure 14.2), 297, 301, 301 (figure 14.3), 302, 304
- Carpenter, Kenneth J., 228, 235
- Chattopadhyay, Raghavendra, 282 (table 14.1), 292, 292 (figure 14.1), 296, 296 (figure 14.2), 297, 301 (figure 14.3), 302, 304, 307
- Checkel, Jeffrey, 215n5
- Chehabi, H. E., 48n18
- Clarke, Kevin A., 151, 258, 259, 264
- Cohen, Bernard P., 153n21, 154, 157, 195
- Cole, Stephen R., 258n16
- Collier, David, 5, 27, 29, 30, 58 (table 2.2), 60, 61, 63, 84, 111n1, 113, 116, 118, 119n5, 120, 125, 129, 131, 134, 135, 135n4, 137n8, 139, 141, 186, 221, 232, 243n1, 262, 274, 299n43, 309
- Collier, Ruth Berins, 134, 178
- Cook, Thomas D., 75, 77, 163, 166n7, 315n2
- Coombs, Clyde H., 136n7, 139
- Copas, J. B., 166n6
- Coppedge, Michael, 153n21, 179, 180n22, 181n24, 340n10, 345n13, 355n20
- Corbin, Juliet, 131
- Cowden, Jonathan A., 87
- Cox, David R., 166n6, 186n29, 301n43, 301n44
- Cox, Gary, 282 (table 14.1)
- Darnell, Adrian C., 175n18, 262, 313n1
- Dawes, Robyn M., 136n7, 139
- Deaton, Angus, 302n47, 308n54
- De Haan, Jakob, 250n7
- Dehejia, Rajeev, 300
- De Soto, Hernando, 285n15
- Diaconis, Persi, 266n39
- Diamond, Jared, 276n5
- Dietz, Thomas, 248n2
- Dijkstra, Theo K., 166n6
- Dion, Douglas, 147, 148, 149, 150
- Di Tella, Rafael, 282 (table 14.1), 285
- Doherty, Daniel, 282 (table 14.1), 283, 283n13, 290, 290n26, 292 (figure 14.1)
- Drezner, Daniel W., 207n1
- Druckman, James N., 274
- Duflo, Esther, 280, 282 (table 14.1), 292, 292 (figure 14.1), 296, 296 (figure 14.2), 297, 301 (figure 14.3), 302, 304, 307
- Duncan, Otis Dudley, 77, 78, 232
- Dunnill, Michael S., 236
- Dunning, Thad, 9, 204, 274, 276, 276n5, 277, 281, 281n12, 282 (table 14.1), 288n20, 289, 292, 297n38, 300, 304, 308, 309, 314, 333, 337, 340, 350
- Durham, J. Benson, 250n7, 260 (table 13.3)

- Eckstein, Harry, 90n2, 91, 130, 146, 323
 Eden, Lynn, 207n1
 Eijkman, Christiaan, 227–29
 Eliot, Charles W., 233
 Elman, Colin, 211, 243n1
 Elster, Jon, 186n28, 195
 Engerman, Stanley L., 254n13
 English, Robert, 215n5, 216
 Epstein, Lee, 3, 34n1
 Evans, Richard J., 227, 234
 Eyler, John M., 235
- Fajardo, Luis E., 194
 Farr, William, 234–35
 Fearon, James, 202, 248n2, 249
 Feigl, H., 356n21
 Feng, Yi, 250, 254n13, 260 (table 13.3), 267–68
 Fenner, Frank, 234
 Fenno, Richard F., 84, 102
 Ferejohn, John, 72
 Ferraz, Claudio, 282 (table 14.1)
 Feynman, Richard Phillips, 90, 115, 115n3
 Finan, Frederico, 282 (table 14.1)
 Fiorina, Morris, 72
 Fish, M. Steven, 79–80
 Fisher, Sir Ronald Aylmer, 112, 280, 300n45
 Fiske, D. W., 78, 80, 356n21
 Fleming, Alexander, 231, 232, 236
 Frankovic, Kathleen, 238n2
 Freedman, David A., 2, 6, 7n7, 8, 8n8, 10, 16n2, 75, 85, 143, 167, 171, 186n29, 202, 202n2, 207, 208n3, 209, 232, 234, 235, 264, 269, 274, 274n2, 278, 279, 281, 281n12, 283 (table 14.1), 284, 290, 295, 296, 299nn43–44, 300, 300n45, 309, 310, 348, 350, 353n17
 Frey, R. Scott, 248n2
 Friedman, Milton, 250
- Galiani, Sebastian, 282 (table 14.1), 284, 285, 292, 292 (figure 14.1), 296, 296 (figure 14.2), 297, 301 (figure 14.3), 304, 307, 309
 Garrett, Geoffrey, 43, 179
 Gasiorowski, Mark J., 250, 260 (table 13.3), 262
 Geertz, Clifford, 180, 180n22, 334, 345n13
 Gelman, Andrew, 79
 George, Alexander L., 103–4, 187n32, 191, 207, 207n1, 211, 237
 Gerber, Alan, 274, 276, 276n5, 281, 282 (table 14.1), 283, 283n13, 290, 292, 292n26
 Gerring, John, 153, 153n21, 156, 157, 313n1
 Giesbrecht, Peter, 236
 Gill, Christopher J., 208
 Gillespie, John V., 248n2
 Gilligan, Michael, 287
 Glaeser, Edward L., 250n7
 Glaser, Barney G., 131
 Glazer, Amihai, 282 (table 14.1)
 Goemans, Hein, 212–14
 Goertz, Gary, 216, 217, 218, 219
 Goldberger, Joseph, 229–30, 232, 235
 Goldsmith, Margaret, 236
 Goldstone, Jack, 47
 Goldthorpe, John H., 166n6, 186–87, 186nn29–30, 187n31, 192, 269
 Gourevitch, Peter Alexis, 91, 92, 93, 95
 Grandy, Richard E., 71n6, 72n8
 Granger, Clive W. J., 172
 Green, Donald P., 274, 276, 276n5, 279, 281, 282–83 (table 14.1), 283, 283n13, 289, 290, 292, 292n26, 299n44, 304n52
 Greene, Kenneth F., 127
 Greene, William H., 138, 167, 168n12, 175n18, 195, 262
 Griffin, Larry, J., 105
 Griffin, Robert, 283 (table 14.1), 294, 294nn30–31, 298, 301, 302
 Griliches, Zvi, 195
 Grofman, Bernard, 283 (table 14.1), 292 (figure 14.1), 294, 294n30, 296 (figure 14.2), 298, 301, 301 (figure 14.3), 302
 Gujarati, Damodar N., 175n18

- Haggard, Stephan, 250
 Hanushek, Eric A., 175n18, 262
 Hare, Ronald, 235, 236
 Hartley, Thomas, 87
 Heberle, Rudolf, 93–94, 95
 Heckman, James J., 6, 75, 144, 267n22, 279, 302n47
 Helliwell, J.F., 250n8
 Hempel, Carl G., 72, 90, 90n2, 93n7, 135, 139
 Hendry, David F., 172
 Hibbs, Douglas A., 178
 Hicks, Alexander, 179
 Hidalgo, F. Daniel, 283 (table 14.1), 287n21
 Hill, Austin Bradford, 186n29
 Ho, Daniel E., 265n21, 283 (table 14.1)
 Hodges, Joseph L., 37
 Holland, Paul W., 37, 38, 41, 42nn11–12, 44n14, 71, 72, 72n7, 73, 73n10, 74, 76, 114, 138, 163, 164, 204, 276, 276n4, 277n7, 295n34, 316, 340–41, 347
 Homer-Dixon, Thomas F., 207n1
 Hoole, Francis, 248n2
 Huntington, Samuel P., 267
 Hutchinson, James S., 233
- Imai, Kosuke, 265n21, 283 (table 14.1)
 Imbens, Guido, 302n47, 308n54
 Inkeles, Alex, 250
 Iwasaki, Y., 186n29
 Iyer, Lakshmi, 282 (table 14.1), 304n52
- Jackman, Robert W., 179, 248n2
 Jagers, Keith, 257 (figure 13.2)
 Jamieson, Kathleen Hall, 238n2
 Jenner, Edward, 222–23, 231, 233, 234
 Johnson, Simon, 287n21
- Kalof, Linda, 248n2
 Katz, Jonathan N., 44, 248n2
 Katzenstein, Peter J., 30, 94, 95, 95n9, 96, 119–20
 Kaufman, Robert R., 250
 Kennedy, Peter, 175n18, 261, 262, 313n1
- Keohane, Robert O., 1, 3n4, 17, 26, 29–30, 33, 67, 69n3, 83, 89, 101, 111n1, 125, 126, 128, 128nn1–2, 129, 140, 141n12, 156n24, 187, 187n33, 199n39, 209, 274, 313n1
 Kerlinger, Fred N., 76
 Key, V. O., 86
 Khong, Yuen Foong, 207n1
 Kim, Jae-On, 138
 King, Gary, 1, 3, 3n4, 17, 26, 29–30, 33, 59n25, 67, 69n3, 71n6, 79, 80n14, 83, 89, 101, 111n1, 125, 126, 128, 128nn1–2, 129, 140, 141n12, 156n24, 187, 187n33, 199n39, 209, 274, 313n1
 Kitschelt, Herbert, 127
 Kittel, Bernhard, 44
 Knopf, Jeffrey W., 207n1
 Koetzle, William, 283 (table 14.1)
 Kohli, Atul, 93
 Kornhauser, William, 92, 95
 Kottke, Janet L., 137
 Krantz, David L., 77, 78, 79, 80n15, 136
 Krasno, Jonathan S., 283 (table 14.1), 302n52
 Kremer, Michael, 280
 Krieckhaus, Jonathan, 250n8, 251
 Kriesi, Hanspeter, 108
 Krueger, Alan B., 282 (table 14.1), 291, 292 (figure 14.1), 295, 296 (figure 14.2), 299, 300, 301 (figure 14.3), 302
 Kubik, Jan, 107
 Kuhn, Thomas, 97
 Kurzman, Charles, 260 (table 13.3), 250n6
- Laba, Roman, 107–8
 Laitin, David D., 111n1, 113, 115, 115n3, 116, 126, 133, 134, 154, 202, 248n2, 249, 304n52
 Lakatos, Imre, 347n14
 Lake, David A., 250n8, 260 (table 13.3), 269
 Lane, Robert E., 84, 167
 Lang, Janet, 166n6
 Lange, Peter, 43, 179

- Larson, Deborah Welch, 207n1
 Lave, Charles, 89n1, 96
 Lavy, Victor, 282 (table 14.1), 285, 291, 292 (figure 14.1), 296 (figure 14.2), 298, 298nn40–41, 301, 301 (figure 14.3), 302, 308, 309
 Layne, Christopher, 211
 Lazarsfeld, Paul F., 131, 327
 Leamer, Edward E., 6n5, 85, 162, 163, 165, 166n6, 169, 171, 172, 176, 262, 264
 Leblang, David A., 250, 250n6
 Lederer, Emil, 92
 Lee, David S., 283 (table 14.1), 288n20, 292 (figure 14.1), 293
 Lee, Jong Hua, 255
 Lehmann, Erich L., 37
 Lennox, Richard, 77
 Lerman, Amy, 283 (table 14.1), 288n20, 309
 Levine, Ross, 172
 Levitsky, Steven, 84, 137n8
 Lewis-Beck, Michael, 179
 Li, H. G., 166n6
 Lichbach, Mark Irving, 90n3, 387n1
 Lieberman, Evan, 181n24, 340n11
 Lieberman, Stanley, 60, 157, 166n6
 Lijphart, Arend, 30, 91, 91n5, 93, 94, 95, 97, 116–17, 117n4, 128, 183
 Limongi, Fernando, 144n15, 250, 250n7, 262, 270
 Lindauer, David, 248n3, 271
 Linz, Juan J., 48n18
 Lipton, Michael, 96
 Littlewood, J. E., 90n4
 Liu, T. C., 166n36
 Lott, John R., Jr., 188, 203, 238, 238n2, 239, 240–42
 Loudon, Irvine, 235
 Lucas, R. E., Jr., 166n6
 Mahoney, James, 3, 4, 29, 58 (table 2.2), 60, 61, 63, 125, 129, 141, 177, 186, 202n2
 Manski, Charles F., 144, 166n6
 March, James G., 89n1, 96
 Mariscal, Elisa, 254n13
 Marshall, Monty G., 257 (figure 13.2)
 Martin, Lisa L., 85, 104
 Mason, Linda, 238n2
 McAdam, Doug, 106–7, 186n28
 McKeown, Timothy J., 58 (table 2.2), 59, 60, 63, 103, 104, 125, 128, 131, 146, 166, 186
 McKim, Vaughn R., 166n6
 McNulty, John, 281, 282 (table 14.1), 292 (figure 14.1), 293, 293n27, 296 (figure 14.2), 297, 301, 301 (figure 14.3), 302
 Messick, Samuel, 139–40
 Michell, Joel, 136n7, 137, 139
 Michels, Robert, 84
 Migdal, Joel S., 148
 Miguel, Edward, 283 (table 14.1), 288, 304n52, 308
 Mill, John Stuart, 10, 146, 337, 337n8, 338, 338n9
 Minier, Jenny A., 250n6
 Mirer, Thad W., 175n8
 Montesquieu, Charles Louis de Sec-
 ondat, Baron de, 92
 Moore, Barrington, Jr., 93, 95
 Moravcsik, Andrew, 207n1
 Moreno, Erika, 194
 Morgan, Pamela S., 1n1
 Morgan, Stephen L., 6n6, 265
 Morton, Rebecca B., 274
 Moss, Pamela A., 137, 140
 Mueller, Charles W., 138
 Munck, Gerardo L., 27, 34n1, 58 (table 2.2), 60, 61, 63, 125, 129, 131, 134, 134n3, 135, 141, 142, 145, 148, 157, 172, 186
 Naidu, Suresh, 283 (table 14.1)
 Nelson, Michael A., 250n6
 Nesvold, Betty A., 248n2
 Neyman, Jerzy Splawa, 37, 138, 163–64, 204, 276, 276n4, 277n7, 295n34, 299n43, 340–41
 Ni Bhrolcháin, Máire, 166n6
 Nichter, Simeon, 283 (table 14.1)
 Nuland, Sherwin, 235
 Nutton, Vivian, 233

- Olson, Mancur, 269
 Owen, John M., IV, 207n1
 Oye, Kenneth, 214
- Pace, Rossana Sciubba, 109
 Paluck, Elizabeth Levy, 307, 307n55
 Patterson, William David, 179
 Pearson, Karl, 16n2, 353n17
 Perrot, Michelle, 105
 Piore, Michael, 131, 202
 Pisani, Robert, 8, 143, 167, 171, 235, 281, 300n45
 Pischke, Jörn-Steffen, 296, 296n41
 Pinto, Pablo M., 250n8
 Pitkin, Hanna, 81, 81n16
 Polanyi, Karl, 84
 Popper, Karl, 115n3, 129
 Porkess, Roger E., 177
 Porter, Michael, 94
 Porter, Roy, 222, 233
 Posner, Daniel, 281, 283 (table 14.1), 292 (figure 14.1), 293, 296 (figure 14.2), 297n39, 301 (figure 14.3), 302, 302n46, 303–4, 303n49
 Powell, Robert, 131
 Pratt, J. W., 138, 253n10, 263
 Pritchett, Lant, 248n3, 271
 Przeworski, Adam, 10, 21, 43, 101, 131, 144, 144n15, 153n21, 154, 155, 157, 250, 250n7, 251, 259, 260 (table 13.3), 262, 265, 267, 267n22, 268n23, 270, 338, 338n9
 Purves, Roger, 143, 167, 171, 235, 281, 300n45
 Putnam, Robert D., 85, 102–3
- Ragin, Charles C., 18, 58 (table 2.2), 59, 60, 61, 63, 125, 128–29, 131, 135, 141, 142, 146, 148, 149, 150, 151, 153n21, 172, 181, 182, 315, 316, 317, 318, 331, 344, 358
 Rasch, Georg, 77, 79
 Relles, Daniel, 143
 Renelt, David, 172
 Rice, Condoleezza, 218
 Rice, Tam W., 179
 Richard, Jean-François, 172
- Richardson, Neal, 283 (table 14.1)
 Riggs, Matt L., 137
 Robbins, Marc, 282 (table 14.1)
 Roberts, Fred, 136n7, 139
 Robins, James M., 166n6
 Robinson, James A., 276n5, 289n21
 Rock, Stephen R., 207n1
 Rodrik, Dani, 254n13
 Roe, Daphne A., 230
 Rogowski, Ronald, 28, 29, 30, 59 (table 2.2), 60, 61, 63, 111n1, 115, 116, 117, 117n4, 118, 119, 120, 121, 125, 128, 128nn1–2, 131, 141, 154, 186, 233
 Rosenbaum, Paul R., 138, 186n29, 342n11
 Rosenbluth, Frances, 282 (table 14.1)
 Ross, Laurence H., 163, 166n7, 281n14, 315n2
 Rothman, Kenneth J., 166n6
 Rotnitzky, Andrea, 166n6
 Rubin, Donald B., 37, 41, 42, 42nn11–12, 44n14, 71, 72–73, 73n10, 74, 76, 138, 163–64, 164n4, 204, 276, 276n4, 277n7, 286n17, 295n34, 299n43, 316, 340–41, 347
 Rueschmeyer, Dietrich, 178
 Russell, Bertrand, 90n4
- Sabin, Lora, 208
 Sagan, Scott, 207n1
 Sala-i-Martin, Xavier, 253
 Salmon, Wesley C., 72
 Sartori, Giovanni, 153n21
 Satyanath, Shanker, 283 (table 14.1), 288, 308
 Scharfstein, David O., 166n6
 Schargrodsky, Ernesto, 282 (table 14.1), 284, 285, 292, 292 (figure 14.1), 296, 296 (figure 14.2), 297, 301 (figure 14.3), 304, 307, 309
 Schlaifer, Robert, 138, 253n10
 Schlozman, Kay Lehman, 47
 Schmid, Christopher H., 208
 Schwandt, Thomas A., 313n1
 Sciubba, Roberto, 109
 Scriven, Michael, 72, 72n8

- Seawright, Jason, 9, 27, 29, 30, 31, 58
 (table 2.2), 59, 60, 61, 63, 125, 129,
 141, 147, 148, 149n18, 150, 151,
 186, 203, 204, 221, 232, 262, 274,
 306, 209, 333, 336, 337, 346, 348
- Sekhon, Jasjeet, 5, 274, 277, 281,
 286n20, 289n25, 291, 297n43
- Semmelweis, Ignac, 223–24, 232, 235
- Sergenti, Ernest J., 283 (table 14.1), 288,
 289, 308
- Shafer, Michael, 207n1
- Shen, Jian-Guang, 250n6
- Shepard, Lorrie, 136, 137
- Sherman, Lawrence, 307n55
- Shugart, Matthew, 194
- Shultz, Kenneth S., 137
- Siermann, C., 250n7
- Sil, Rudra, 153n21
- Simon, Herbert A., 186n29
- Singh, Ram D., 250n6
- Sirowy, Larry, 250
- Skocpol, Theda, 150
- Smith, Jeffrey A., 288
- Smoke, Richard, 207n1
- Snow, John, 202, 225–27, 231, 233,
 234, 235, 281n12, 283 (table 14.1),
 284, 292, 292 (figure 14.1), 295, 296
 (figure 14.2), 296n35, 301, 301
 (figure 14.3), 304, 307, 309
- Snyder, Jack, 161n1, 207n1, 214, 216,
 217, 218
- Snyder, James M., 282 (table 14.1)
- Sokoloff, Kenneth L., 254n13
- Solow, Robert M., 259
- Sovey, Allison J., 287
- Stanley, Julian C., 75, 114, 162, 163,
 166n7, 284n14, 286n17, 301, 315n2
- Starr, Harvey, 151
- Stasavage, David, 283 (table 14.1)
- Stein, Janice Gross, 215n5
- Stephens, Evelyn Huber, 178
- Stephens, John D., 178
- Stevens, S. S., 76, 78, 80n14
- Stewart, Charles, 282 (table 14.1)
- Stigler, Stephen M., 16n2, 353n17,
- Stock, James H., 238n3
- Stokes, Susan Carol, 189, 196, 295n33
- Stolzenberg, Ross M., 143
- Stone, Charles, 170n14
- Stone, Richard, 45, 138, 175n18
- Stouffer, Samuel A., 86
- Strang, Heather, 307n55
- Strauss, Anselm, 131
- Sundquist, James L., 134
- Tannenwald, Nina, 189–90, 215, 217
- Tarrow, Sidney, 29, 30, 58 (table 2.2),
 61, 63, 108, 109, 111n1, 113, 122,
 125, 186, 186n28, 187, 188, 188n34
- Terris, Milton, 235
- Teune, Henry, 10, 21, 43, 101, 131,
 153n21, 154, 155, 157, 338, 338n9
- Thies, Michael F., 282 (table 14.1)
- Thistlethwaite, Donald L., 265, 284,
 285, 285n18
- Thompson, E. P., 84
- Tilly, Charles, 92, 105, 106, 178,
 186n28
- Timmons, Jeffrey M., 250n8
- Timonius, Emanuel, 234
- Tintner, Gerhard, 76
- Titiunik, Rocio, 283 (table 14.1),
 288n20, 293
- Tocqueville, Alexis de, 92, 108
- Todd, Petra E., 288
- Truman, David Bicknell, 91, 95, 116,
 117
- Turner, Stephen P., 166n6
- Tversky, Amos, 139
- Urzua, Sergio, 302n47
- Van Evera, Stephen, 202, 210–11,
 313n1, 344n12
- Verba, Sidney, 1, 3n4, 17, 26, 29–30, 33,
 47, 67, 69n3, 83, 89, 101, 111n1,
 125, 126, 128, 128nn1–2, 129, 140,
 141n12, 156n24, 187, 187n33, 209,
 274, 313n1
- Verdier, Thierry, 254n13
- Vinten-Johansen, Peter, 234, 235
- Vogt, W. Paul, 168n10, 175n18, 177,
 183n26, 313n1

- Wahba, Sadek, 290
Walker, Henry A., 195
Wallerstein, Immanuel, 91, 92, 95
Wallerstein, Michael, 166n6, 187n32
Walsh, Christopher, 236
Walt, Stephen M., 207n1
Watson, Mark W., 238n3
Weber, Steven, 207n1
Weber, Max, 195, 334
Wermuth, N., 186n29
Werum, Regina, 250n6
White, Halbert, 172
Wickham-Crowley, Timothy P., 147,
178
Williams, Kenneth C., 274
Winship, Christopher, 6n6, 265
Wohlforth, William, 214, 215, 216, 217
Wonnacott, Ronald J., 175n18
Wonnacott, Thomas H., 175n18
Woodward, John, 234
Wooldridge, Jeffrey M., 238n3, 274

Yule, G. Udny, 16n2, 353n17

Zak, Paul J., 254n13
Zaller, John, 179
Zelikow, Philip, 218
Zinnes, Dina, 248n2

About the Contributors

Larry M. Bartels is Professor of Politics and Public Affairs, Donald E. Stokes Professor of Public and International Affairs, and Director of the Center for the Study of Democratic Politics at Princeton University. His methodological work has appeared in the *American Journal of Political Science*, *Political Methodology*, and *Political Analysis*. He has also written extensively on electoral politics, public opinion, and political accountability. His most recent book is *Unequal Democracy: The Political Economy of the New Gilded Age* (2008), which won the Gladys M. Kammerer Award for the year's best book on U.S. national policy, among other honors. Bartels is an elected Fellow of the American Academy of Arts and Sciences, an elected Fellow of the American Academy of Arts and Sciences and the American Academy of Political and Social Science, a past president of the Political Methodology Section of the American Political Science Association, and a former chair of the Board of Overseers of the American National Election Studies.

Andrew Bennett is Professor of Government at Georgetown University. Through his publications and institutional leadership, he has been a major contributor to new work on qualitative and multi-method research. He is coauthor of an influential study—"Do We Preach What We Practice?" (2003)—which documents the substantial imbalance between the extensive use of qualitative methods in political science and the small number of graduate departments offering courses that teach these methods. He is the coauthor, with Alexander L. George, of *Case Studies and Theory Development in the Social Sciences* (2005). Bennett's extensive work in the field of international relations includes *Condemned to Repetition? The Rise, Fall, and Reprise of Soviet-Russian Military Interventionism, 1973-1996* (1999), and the collaborative volume *Friends in Need: Burden Sharing in the Persian Gulf War* (1997), both of which use typological theorizing, process tracing, and other

case study methods. Recent articles include "Trust Bursting Out All Over: The Soviet Side of German Unification" (2002), and "The Guns that Didn't Smoke: Ideas and the Soviet Non-Use of Force in 1989" (2005). Bennett is President of the Consortium for Qualitative Research Methods.

Henry E. Brady is Dean of the Goldman School of Public Policy and Class of 1941 Monroe Deutsch Professor of Political Science and Public Policy at the University of California, Berkeley. He received his Ph.D. in Economics and Political Science from MIT, and he is currently President of the American Political Science Association. Brady's latest research focuses on political participation in America, Estonia, and Russia, the dynamics of public opinion and political campaigns, and public policy for voting systems, social welfare programs, and higher education. He has coauthored *Letting the People Decide: Dynamics of a Canadian Election*, *Expensive Children in Poor Families*, *Counting all the Votes*, and *Voice and Equality: Civic Voluntarism in American Politics*. He has co-edited and contributed to *The Oxford Handbook of Political Methodology* and *Capturing Campaign Effects*. Brady has also published numerous articles on political participation, public opinion, and methodology. He is an elected Fellow of the American Academy of Arts and Sciences and the American Association for the Advancement of Science.

David Collier is Robson Professor of Political Science at the University of California, Berkeley, where he has served as Department Chair and Chair of the Center for Latin American Studies. His research focuses on democracy and authoritarianism, Latin American politics, comparative-historical analysis, and methodology. Collier's books include *Shaping the Political Arena* (with Ruth Berins Collier, 1991, reissued in 2002), which won the Best Book Prize of the APSA Comparative Politics Section. He recently co-edited three books on methods: *The Oxford Handbook of Political Methodology* (2008), *Concepts and Method in Social Science* (2009), and *Statistical Models and Causal Inference* (2010). Within the American Political Science Association, he has served as President of the Organized Section for Comparative Politics, Vice President of the Association, and founding President of the Organized Section for Qualitative and Multi-Method Research. Collier is an elected Fellow of the American Academy of Arts and Sciences.

Thad Dunning is Associate Professor of Political Science at Yale University and a research fellow at Yale's Whitney and Betty MacMillan Center for International and Area Studies as well as the Institution for Social and Policy Studies. He studies comparative politics, political economy, and qualitative and quantitative methodology. His book *Crude Democracy: Natural Resource Wealth and Political Regimes* (2008) contrasts the democratic and authoritarian effects of natural resource wealth. Dunning has written on a

range of methodological topics, including multi-method research, instrumental-variables analysis, and the use of natural experiments in the social sciences. His current substantive work on ethnic and other cleavages draws on qualitative fieldwork and field and natural experiments in Latin America, India, and Africa. His research has appeared in the *American Political Science Review*, *Comparative Political Studies*, *International Organization*, *The Journal of Conflict Resolution*, *Political Analysis*, *Studies in Comparative International Development*, and other journals. He holds a Ph.D. in Political Science and an M.A. in Economics from the University of California, Berkeley.

David A. Freedman (1938–2008) was Professor of Statistics at the University of California, Berkeley. He was a distinguished mathematical statistician whose theoretical research included the analysis of martingale inequalities, Markov processes, de Finetti's theorem, consistency of Bayes estimators, sampling, the bootstrap, and procedures for testing and evaluating models and methods for causal inference. Freedman published widely on the application—and misapplication—of statistics in works within a variety of social sciences, including epidemiology, demography, public policy, and law. He emphasized exposing and checking the assumptions that underlie standard methods, as well as understanding how those methods behave when the assumptions are false—for example, how regression models behave when fitted to data from randomized experiments. He had a remarkable talent for integrating carefully honed statistical arguments with compelling empirical applications and illustrations. Freedman was an elected Fellow of the American Academy of Arts and Sciences, and he received the National Academy of Science's John J. Carty Award for his "profound contributions to the theory and practice of statistics."

Robert O. Keohane is Professor of International Affairs at Princeton University. He is the author of *After Hegemony: Cooperation and Discord in the World Political Economy* (1984) and *Power and Governance in a Partially Globalized World* (2002). He is coauthor (with Joseph S. Nye, Jr.) of *Power and Interdependence* (third edition 2001), and (with Gary King and Sidney Verba) of *Designing Social Inquiry* (1994). He has served as the editor of the journal *International Organization* and as President of the International Studies Association and the American Political Science Association. He won the Grawemeyer Award for Ideas Improving World Order and the Johan Skytte Prize in Political Science. He is an elected Fellow of the American Academy of Arts and Sciences and a member of the American Philosophical Society and National Academy of Sciences. He has received honorary degrees from the University of Aarhus, Denmark, and Science Po in Paris,

and was the Harold Lasswell Fellow of the American Academy of Political and Social Science.

Gary King, Albert J. Weatherhead III University Professor at Harvard, teaches in the Department of Government and serves as Director of the Institute for Quantitative Social Science. His work develops and applies empirical methods in many areas of social science research, focusing on innovations that range from statistical theory to practical application. King's authored, coauthored, and edited books include *Unifying Political Methodology: The Likelihood Theory of Statistical Inference* (1989), *Designing Social Inquiry: Scientific Inference in Qualitative Research* (1994), *A Solution to the Ecological Inference Problem* (1997), *Demographic Forecasting* (2008), and *The Future of Political Science* (2009). He has made major contributions to creating computational software for the social sciences, having developed fifteen open source software packages. King is an elected Fellow of the American Statistical Association, American Association for the Advancement of Science, and American Academy of Arts and Sciences, and an elected Member of the National Academy of Sciences.

Gerardo L. Munck is Professor in the School of International Relations at the University of Southern California. His research focuses on political regimes and democratization, Latin American politics, and research methods. He is author of *Authoritarianism and Democratization: Soldiers and Workers in Argentina, 1976–83* (1998) and numerous articles and book chapters, including "Game Theory and Comparative Politics: New Perspectives and Old Concerns," published in *World Politics* (2001). In 2002 he coauthored the lead article in a symposium on "Conceptualizing and Measuring Democracy: Evaluating Alternative Indices," published in *Comparative Political Studies*. This article subsequently won the Award for Conceptual Innovation in Democratization Studies of the IPSA Committee on Concepts and Methods. In 2009 he published *Measuring Democracy: A Bridge between Scholarship and Politics*. Munck has been Chief Technical Advisor for the UNDP *Report on Democratic Development in Latin America*.

Ronald Rogowski is Professor and former Chair of Political Science, University of California, Los Angeles. At UCLA, he has also been Interim Dean and Vice Provost of the International Institute Studies, as well as Founding Director of its Global Studies program. Rogowski received his B.A. in Political Science and Mathematics from the University of Nebraska, and Ph.D. in Politics from Princeton University. He also studied at the Free University of Berlin. Rogowski has held research appointments at Harvard University, the Center for Advanced Study in the Behavioral Sciences, and the *Wissenschaftskolleg* in Berlin. He has also taught at Princeton and Duke, and was

Visiting Stassen Professor at the University of Minnesota. Rogowski is author of *Rational Legitimacy* (1974) and *Commerce and Coalitions* (1989). He likewise is co-author of *Electoral Systems and Consumer Power* (Cambridge, 2010) and co-editor of *Essential Readings in Comparative Politics* (3rd ed., 2009). His current research focuses on the economic and political causes and effects of economic inequality. Rogowski has served as Vice-President and Program Co-chair of the American Political Science Association and President of the APSA Comparative Politics Section, and is currently Lead Editor of the *American Political Science Review*. He is an elected Fellow of the American Academy of Arts and Sciences.

Jason Seawright, Assistant Professor of Political Science at Northwestern University, specializes in comparative party systems, Latin American politics, and methodology. He is author of "Testing for Necessary and/or Sufficient Causation: Which Cases are Relevant?" and "What Counts as Evidence?," published in the methodology journal *Political Analysis* (2002); as well as "Qualitative Comparative Analysis vis-à-vis Regression" and "Assumptions, Causal Inference, and the Goals of QCA" (2005), both published in *Studies in Comparative International Development*. Seawright is also the coauthor of "Case Selection Techniques in Case Study Research: A Menu of Qualitative and Quantitative Options" (2008), published in *Political Research Quarterly*. His study "Political Participation and Representational Distortion: The Nexus Between Associationalism and Partisan Politics" appeared in 2010. He holds a Ph.D. in Political Science and an M.A. in Statistics from the University of California, Berkeley.

Sidney Tarrow is Maxwell M. Upson Professor of Government and Professor of Sociology at Cornell University, where he teaches comparative European politics, political sociology, and collective action. Tarrow's numerous books on social movements and European politics include *Democracy and Disorder: Protest and Politics in Italy, 1965–1975* (1989), *Power in Movement: Social Movements, Collective Action, and Politics* (1998); *Dynamics of Contention* (with Doug McAdam and Charles Tilly, 2001); *Contentious Europeans: Protest and Politics in an Emerging Polity* (with Doug Imig, 2001); and *The New Transnational Activism* (2005). He is currently working on the historical relations among war, state-building, and movements for rights. Tarrow is an elected Fellow of the American Academy of Arts and Sciences.

Sidney Verba is Carl H. Pforzheimer University Professor Emeritus and Research Professor of Government at Harvard. His research focuses on political participation, electoral politics, participatory democracy, civic culture, and methodology. Verba's most recent book is *The Private Roots of Public Action* (with Nancy Burns and Kay Lehman Schlozman, 2001), which

won the American Political Science Association's Victoria Schuck Prize. His many coauthored books include *Elites and the Idea of Equality: A Comparison of Japan, Sweden, and the United States* (1987), *Voice and Equality: Civic Voluntarism in American Politics* (1995), *Equality in America: The View from the Top* (1995), and *Designing Social Inquiry: Scientific Inference in Qualitative Research* (1994). Verba is former President of the American Political Science Association, a member of the National Academy of Sciences, and an elected Fellow of the American Academy of Arts and Sciences. He received Uppsala University's Johan Skytte Prize for Distinguished Contribution to Political Science.