SOCIAL SCIENCE METHODOLOGY: A WIDE-RANGING INTRODUCTION

John Gerring
Dino Christenson
Department of Political Science
Boston University

Cambridge University Press

Draft: January 20, 2016

Word count (excluding Part IV): 139,000
Target: 200,000

Comments welcome!
Contents

Preface...........................................................................................................................................................18
I. BUILDING BLOCKS ..................................................................................................................................22
  1. A Unified Framework...............................................................................................................................24
  2. Arguments ............................................................................................................................................37
  3. Concepts and Measures.........................................................................................................................58
  4. Analyses................................................................................................................................................78
II. CAUSALITY ..........................................................................................................................................94
  5. Causal Frameworks ...............................................................................................................................95
  6. Causal Hypotheses and Analyses .......................................................................................................122
  7. Experimental Designs..........................................................................................................................141
  8. Large-N Observational Designs.........................................................................................................162
  9. Case Study Designs ............................................................................................................................185
  10. Diverse Tools of Causal Inference ...............................................................................................199
III. PROCESS AND PRESENTATION ....................................................................................................211
  11. Reading and Reviewing ....................................................................................................................212
  12. Brainstorming..................................................................................................................................229
  13. Data Gathering..................................................................................................................................242
  14. Writing...............................................................................................................................................280
  15. Speaking.........................................................................................................................................300
  16. Ethics...............................................................................................................................................311
References..................................................................................................................................................320
Index..........................................................................................................................................................331
Detailed Contents

Notation ................................................................................................................................................... 16
Acknowledgments .................................................................................................................................. 17
Preface........................................................................................................................................................... 18
A Wide-Ranging Approach ................................................................................................................... 19
Outline and Features .............................................................................................................................. 20
I. BUILDING BLOCKS ............................................................................................................................ 22
1. A Unified Framework......................................................................................................................... 24
   The Purpose of Unity............................................................................................................................. 26
   Examples.................................................................................................................................................. 28
   Worker-training programs .............................................................................................................. 28
   Social capital ................................................................................................................................... 29
   Democracy ...................................................................................................................................... 30
   Conclusions ..................................................................................................................................... 32
Key Terms................................................................................................................................................ 34
Inquiries.................................................................................................................................................... 34
Resources ................................................................................................................................................. 35
   Web sites......................................................................................................................................... 35
   History and philosophy of social science ........................................................................................ 35
   Worker-training programs .............................................................................................................. 35
   Democracy ...................................................................................................................................... 36
   Social capital ................................................................................................................................... 36
2. Arguments ............................................................................................................................................ 37
   Descriptive Arguments ........................................................................................................................ 37
      Figure 2.1: Taxonomy of Descriptive Arguments....................................................................... 38
   Associations .................................................................................................................................... 39
      Figure 2.2: Possible Relationships among Two Factors ............................................................. 40
   Syntheses.......................................................................................................................................... 41
   Simple Typologies ............................................................................................................................. 42
   Periodization ................................................................................................................................... 43
   Matrix Typologies ............................................................................................................................. 43
      Table 2.1: A Matrix Typology: Regime-types ............................................................................. 44
II. CAUSALITY

5. Causal Frameworks

Table 5.1: Causal Frameworks

Motivational Frameworks
- Interests
- Norms
- Psychology

Structural Frameworks
- Material factors
- Human capital/Demography
- Institutions

Interactive Frameworks
- Adaptation
- Coordination
- Diffusion
- Networks
- Path dependence

Building a Theory
- An Example: Birth
- Formal models

Conclusions

Key Terms

Inquiries

Resources

Causal frameworks, generally considered

Interests (that which benefits an actor)

Norms (values, beliefs, ideology, culture, identity)

Information (ideas, knowledge, learning)
Psychology (cognition, emotion, personality, genetics) ............................................................... 116
Material factors (resources, modernization, geography, technology) ............................................. 117
Institutions (formal rules and organizations) ................................................................................ 117
Human capital/Demography (health, education, skills, migration, life-cycle events) ...................... 118
Diffusion (demonstration effects, exposure) ................................................................................ 118
Adaptation (feedback, competition, selection, evolution, efficiency, functionalism) ...................... 119
Coordination (collective action, externalities, transaction costs) .................................................... 119
Networks (associational structures, social capital) .......................................................................... 120
Path-dependence (contingent moments leading to fixed trajectories) .......................................... 120
Simulations .................................................................................................................................... 120
Formal models ............................................................................................................................... 121

6. Causal Hypotheses and Analyses .................................................................................................. 122
Causality ........................................................................................................................................ 122

Figure 6.1: A Simple Causal Graph .......................................................................................... 125

Causal Graphs ............................................................................................................................ 126

Criteria of a Causal Hypothesis................................................................................................... 126

Table 6.2: Causal Hypotheses: Criteria ..................................................................................... 127

Clarity ........................................................................................................................................... 128
Manipulability ............................................................................................................................. 128
Precedence .................................................................................................................................. 129
Impact .......................................................................................................................................... 130
Mechanism ................................................................................................................................. 131

Causal Analysis .......................................................................................................................... 131

Table 6.3: Causal Research Designs: Criteria ............................................................................ 132

Covariation .................................................................................................................................. 132

Figure 6.2: Covariational Patterns ............................................................................................ 134

Comparability ............................................................................................................................. 135

Conclusions ................................................................................................................................. 136
Key Terms .................................................................................................................................... 137
Inquiries ...................................................................................................................................... 138
Resources .................................................................................................................................... 140

7. Experimental Designs ............................................................................................................. 141
Begin where you are ............................................................................................................................ 230
Get off your home turf ....................................................................................................................... 231
Play with ideas ...................................................................................................................................... 232
Practice dis-belief ................................................................................................................................. 234
Observe empathically .......................................................................................................................... 236
Theorize wildly ..................................................................................................................................... 236
Think ahead .......................................................................................................................................... 238
Conclusions .......................................................................................................................................... 239
Key Terms............................................................................................................................................. 240
Inquiries................................................................................................................................................. 240
Resources .............................................................................................................................................. 240
13. Data Gathering .................................................................................................................................. 242
Standardized Surveys ........................................................................................................................... 243
   Medium, location .......................................................................................................................... 243
   Respondents ................................................................................................................................. 244
   Compensation ............................................................................................................................... 245
   Introductory statement ................................................................................................................ 245
   Questionnaire ............................................................................................................................... 246
   Response-categories ..................................................................................................................... 248
   Knowledge, certainty, and salience .............................................................................................. 248
   Sensitive subjects .......................................................................................................................... 249
   Longitudinal data .......................................................................................................................... 250
   Summary ....................................................................................................................................... 251
Less Structured Interpersonal Settings ............................................................................................. 252
   Interviews ...................................................................................................................................... 253
   Focus groups ................................................................................................................................. 258
   Ethnography .................................................................................................................................. 259
Unobtrusive Measures .......................................................................................................................... 262
   Surreptitious measures ................................................................................................................ 262
   Ex post measures .......................................................................................................................... 263
Data Assessment .................................................................................................................................. 265
   Divergent sources ......................................................................................................................... 266
   Replicability ................................................................................................................................... 269
Validity tests .................................................................................................................................. 270
Conclusions ................................................................................................................................... 271
Key Terms ........................................................................................................................................ 272
Inquiries .......................................................................................................................................... 272
Resources ...................................................................................................................................... 274
Data-gathering methods, generally considered ................................................................. 274
Survey research ........................................................................................................................ 274
Interviewing .............................................................................................................................. 275
Focus groups ............................................................................................................................ 275
Ethnography ............................................................................................................................. 276
Unobtrusive measures ............................................................................................................. 277
Integrating Sources ................................................................................................................... 277
Replicability .............................................................................................................................. 278

14.  Writing ....................................................................................................................................... 280
   Genres .......................................................................................................................................... 281
   Organization ............................................................................................................................. 282
   Introduction .............................................................................................................................. 283
   Literature Review ................................................................................................................... 284
      Table 14.1: Crossnational Studies of Development and Democracy ................................ 285
   Thesis ......................................................................................................................................... 285
   Method ...................................................................................................................................... 286
   Evidence and Supporting Arguments .................................................................................... 286
   Conclusion ............................................................................................................................... 286
   End Matter ............................................................................................................................... 287
   Variations ............................................................................................................................... 287
   Style ......................................................................................................................................... 287
   Rules ......................................................................................................................................... 288
   Sources .................................................................................................................................... 290
   Quotation formats .................................................................................................................. 292
   Citation formats ...................................................................................................................... 292
   Editing ....................................................................................................................................... 294
   Conclusions ............................................................................................................................. 294
   Key Terms ............................................................................................................................... 296
### Notation

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B$</td>
<td>Background causes of $Y$ that are uncorrelated with (orthogonal to) $X$</td>
</tr>
<tr>
<td>$C$</td>
<td>Confounder, or potential confounder – correlated with $X$</td>
</tr>
<tr>
<td>$E$</td>
<td>Expected value</td>
</tr>
<tr>
<td>$H_0$</td>
<td>Hypothesis about $X \rightarrow Y$</td>
</tr>
<tr>
<td>$M$</td>
<td>Mechanism or pathway connecting $X$ to $Y$</td>
</tr>
<tr>
<td>$N$</td>
<td>Observations or (occasionally) units/cases</td>
</tr>
<tr>
<td>$O$</td>
<td>Observation</td>
</tr>
<tr>
<td>$P$</td>
<td>Probability</td>
</tr>
<tr>
<td>$U$</td>
<td>Error term, or uncertainty</td>
</tr>
<tr>
<td>$T$</td>
<td>Time-period</td>
</tr>
<tr>
<td>$X$</td>
<td>Causal factor of theoretical interest, aka condition, covariate, exogenous variable, explanatory variable, explanans, independent variable, input, intervention, parent, predictor, right-side variable, treatment</td>
</tr>
<tr>
<td>$X=1$</td>
<td>Treatment condition</td>
</tr>
<tr>
<td>$X=0$</td>
<td>Control condition</td>
</tr>
<tr>
<td>$\Delta X$</td>
<td>Change in $X$, e.g., from $X=0$ to $X=1$</td>
</tr>
<tr>
<td>$Y$</td>
<td>Causal outcome, aka dependent variable, descendant, effect, endogenous variable, explanandum, left-side variable, output, response</td>
</tr>
<tr>
<td>$Y_{X=1}$</td>
<td>Outcome for member of the treatment group (i.e., where $X=1$)</td>
</tr>
<tr>
<td>$Y_{X=0}$</td>
<td>Outcome for member of the control group (i.e., where $X=0$)</td>
</tr>
<tr>
<td>$X \rightarrow Y$</td>
<td>$X$ causes $Y$</td>
</tr>
<tr>
<td>$X \leftrightarrow Y$</td>
<td>$X$ causes $Y$ and $Y$ causes $X$</td>
</tr>
<tr>
<td>$X - Y$</td>
<td>$X$ co-varies with $Y$</td>
</tr>
<tr>
<td>$Z$</td>
<td>Antecedent cause (antecedent to $X$), which may be used as an instrumental variable</td>
</tr>
<tr>
<td>^</td>
<td>‘hat’ – denotes fitted (estimated) values for a variable</td>
</tr>
</tbody>
</table>
ACKNOWLEDGMENTS

We are grateful for comments and suggestions received from Joe Bizup (Chapter 14), Colin Elman (Chapter 13), Diana Kapiszewski (Chapter 13), Laurel Smith-Doerr (Chapter 16), Peter Spiegler (Chapter 5), Arun Swamy (Chapter 5), Susan Wishinsky (Chapter 11) and reviewers for Cambridge University Press. We would also like to thank Taylor Boas for terrific comments throughout the Statistics section, as well as Max Palmer for a number of helpful suggestions there. We are grateful to our students Christina Jarymowycz, Joshua Yesnowitz and Matthew Maguire, who worked on many aspects of the manuscript. Finally, we want to thank our editor at the Press, John Haslam, who shepherded this book along the path to publication, rendering wise counsel and keeping our noses to the proverbial grindstone.

Chapters 1-4, 6-8, 10, 12 and 17 draw on material published originally in Gerring’s Social Science Methodology: A Unified Framework (Cambridge University Press, 2012). Chapter 9 is based loosely on Gerring’s “Selecting Cases for Intensive Analysis: A Diversity of Methods and Criteria” (Sociological Methods & Research, 2016, joint with Lee Cojocaru) and Case Study Research: Principles and Practices, 2d ed. (Cambridge University Press, 2016a). Readers may refer to these works for a more detailed treatment of these subjects.
Once upon a time, the practice of social science could be understood as the application of commonsense and intuition – something you might develop in the course of growing up. This is no longer true, or only partly true. Although commonsense and intuition are still useful, the social science disciplines have moved well beyond what can be understood without specialized training.

If you want to become an artist, musician, engineer – or pretty much anything, these days – developing your technique in these highly specialized areas is essential. It takes great dedication, countless hours of concentrated work, and professional guidance. The same may be said for social science in the contemporary era. One may mourn the death of the amateur social scientist. But one might as well reconcile oneself to the fact.

In response, methods courses have proliferated at both the undergraduate and graduate level. Likewise, methodological skills are in high demand in the social sciences and their cognate professions. Successful careers in government, communications, education, social work, business, law, and all of the policy fields require a solid grounding in methodology. Whether one is applying for graduate programs or for a job, the material covered in this book should stand one in good stead.

Indeed, a working knowledge of social science tools of analysis may prove more crucial for one’s career than whatever substantive knowledge one acquires in the course of a college education. What one knows is less important than what one can do, and what one can do depends on a working knowledge of methodology.

These developments may be viewed as part of a broader sea-change, driven by the rise of computers and the internet. With sophisticated IT tools at our disposal, factual knowledge about a subject is no longer at a premium and can usually be obtained from a Google search or from a specially designed database in milliseconds. Likewise, any repetitive procedure can be programmed as a set of algorithms on a computer. This means that the value of an education is no longer in the facts or established protocols you might learn. This sort of knowledge can be produced by machines in a more timely and accurate fashion than by the human brain. Our value-added, as humans, stems from our capacity to identify important questions and think through practical solutions to those questions in a creative fashion. This is the function of a broadly pitched course on methodology and it is what this text is designed to convey.

The present text is appropriate for use in introductory or intermediate methods courses at the undergraduate, master’s, or doctoral level. It is designed to assist those who are attempting to make sense of social science as well as those who are conducting original research. We assume no prior methodological knowledge, though we do presume that the reader has some background in at least one field of social science, e.g., anthropology, communications, criminal justice, economics (including business, finance, management), education, environmental policy, international development, law, political science, psychology, public health, public policy, social work, sociology, or urban planning.

1 Some disciplines begin methodological training early, and others delay this training until a substantive grounding in the subject has been conveyed. This means that the reader’s first substantial encounter with methods may occur during the course of an undergraduate, master’s, or PhD program.
We try to address key points of social science methodology in an applied fashion – so that readers can put these methods to work. Note that insofar as we can impact the societies we live in (in a conscious fashion) social science is indispensable. We can’t enhance economic growth, health, and education – or reduce poverty, crime, conflict, inequality, and global warming – without consulting the work of social scientists. To understand that work, and to conduct original research on these topics, an understanding of the methodological principles underlying this set of practices is indispensable. We hope that you will approach social science methodology not simply as a means for self-advancement (though there is surely nothing wrong with that!) but also as a set of tools for changing – and preserving – the world.

**A WIDE-RANGING APPROACH**

In many textbook markets the offerings are fairly similar. A standard format has been developed over the years that everyone adheres to (more or less), and the courses that utilize these texts bear a strong resemblance to each other. There is scholarly consensus in the field about how to teach a subject.

This does not describe the topic at hand. Gazing out across the social science disciplines one finds a wide range of methodological approaches, reflected in a wide range of textbooks. As a service to the prospective reader (and instructor) it may be helpful to indicate how *An Applied Guide* differs from other textbooks in this crowded field – and why.

Some methods texts limit their purview to a specific discipline, e.g., political science, sociology, or economics. This may seem reasonable, and it allows one to focus on a set of substantive problems that orient a field. However, few substantive problems are confined to a single discipline. In order to learn about crime, for example, you will probably need to read across the fields of sociology, psychology, law, political science, economics, and criminology. The same is true for most other problems, which do not observe neat disciplinary boundaries.

Of course, important differences in theory and method characterize the disciplines. But it does not follow that one is well-served by a text that offers only one view of how to conduct social science. A narrow methodological training does not prepare one to integrate knowledge from other disciplines. To understand the range of literature on a topic and to think creatively about methods that might be applied to that topic it makes sense to adopt an ecumenical approach. Hence, this book focuses broadly on the methodological principles of social science rather than on methods practiced within a single discipline.

Some texts are focused primarily on quantitative methods, i.e., statistics or econometrics. While these are important skills, this approach has a tendency to reduce methodology to mathematics. And this, in turn, presents a narrow and technical vision of social science that is not faithful to the way in which social science is practiced (or, at any rate, to the way it should be practiced). Statistics are the handmaiden of methodology, not the other way around.

Some texts are focused exclusively on qualitative methods. This is a hard topic to define, and these books are varied in their content and approach. A few are strongly anti-positivist, meaning that they reject the scientific ideal as it has been understood in the natural sciences. While we agree with the standard critique of a narrowly positivist approach to social science we also think the natural sciences and social sciences share a good deal in common. In any case, a book that treats only qualitative components of social science is missing a good deal of the action. Both qualitative and quantitative approaches are required as part of everyone’s social science education. Certainly, they are both required in order to make sense of the social science literature on a subject.
One way to handle this problem is to include both qualitative and quantitative methods within a single text but to keep them separate, with the idea that the tools are distinct and each draws on a different epistemology (theory of knowledge). In our opinion, this claim is difficult to sustain: “qualitative” and “quantitative” tools tend to blend together and their epistemological traditions are not as far apart as they might seem. More important, a segregated approach to knowledge is not helpful to the advancement of social science. If knowledge on a topic is to grow it must be based on a unified epistemology that encompasses both qualitative and quantitative methods. This is the approach taken in the present text.

The most distinctive feature of this book is its wide-ranging approach to the subject. The text is intended to encompass all of the social science disciplines, qualitative and quantitative methods, descriptive and causal knowledge, and experimental and observational research designs. We also address the nuts and bolts of how to conduct research, as laid out below.

Naturally, there are some topics that we do not have time or space to engage. However, relative to most methods texts this one qualifies as highly inclusive, offering an entrée to myriad aspects of social science methodology. To our way of thinking, these topics are all essential. And they are also closely linked. While there are many ways to do good social science these diverse approaches also share certain common elements. Only by grasping the full extent of social science’s diversity can we glimpse its underlying unity.

**OUTLINE AND FEATURES**

With a text of this size the reader may want to read strategically, focusing on chapters that are most relevant to your current work and interests, skipping or skimming chapters that cover topics about which you are already well-informed. A good textbook need not be read cover-to-cover.

However, readers should also be aware that the book is organized in a cumulative fashion, with later sections building on previous sections. Something may be lost if you peruse the text in a scattershot fashion.

Part I sets forth the basic **building blocks** of social science methodology. Chapter 1 introduces our topic, social science methodology, expanding on themes in the Preface and introducing several specific examples that will be referred to throughout the book. Subsequent chapters within this section focus on (2) arguments (including theories and hypotheses), (3) concepts and measures, and (4) analyses.

Part II focuses on **causal** arguments and analysis. This topic is broken down into chapters dealing with (5) causal frameworks, (6) causal hypotheses and analyses, (7) experimental research designs, (8) nonexperimental research designs, (9) case study research designs, and (10) diverse tools of causal inference.

Part III deals with the process of research and the presentation of results. This includes (11) reading and reviewing the literature on a subject, (12) brainstorming (finding a research topic and a specific hypothesis), (13) data gathering, (14) writing, (15) public speaking, and (16) ethics.

---

2 Philosophy of science is one such topic. This is not because we view it as unimportant. It is, rather, because we find it too important, and too complicated, to reduce to a concise format. The topic quickly becomes unwieldy in the context of an introductory text. Moreover, the abstract arguments generated by philosophers are often difficult to integrate in a meaningful way with the more prosaic tasks of applied social science methodology. Thus, we leave philosophy of science aside (see suggested readings at the end of Chapter 1).
Part IV deals with statistics. This is divided into several topics: (17) data management, (18) univariate statistics, (19) probability distributions, (20) statistical inference, (21) bivariate statistics, (22) regression, and (23) causal inference.

Every effort has been made to divide up these subjects in a way that makes logical sense and to avoid unnecessary redundancies. Of course, topics do not always neatly divide into separate chapters and sections. There is a holistic quality to social science methodology; diverse topics invariably bleed into one another. To assist the reader, we insert cross-references indicating where the reader might look for further elaboration of an issue. You may also consult the Detailed Table of Contents or the Index.

An objective of the book is to introduce readers to key terms of social science methodology. When a term is first introduced, or when it is formally defined, it is printed in bold. At the end of each chapter the reader will find a list of these bolded terms, which may be useful for purposes of review. In the Index, we indicate the page on which a term is defined by printing that number in bold.

At the end of each chapter the reader will find a series of questions and exercises under the heading Inquiries. These inquiries serve a review function, summarizing the main points of the chapter. Some questions are speculative, building on the material presented but also moving beyond it. Instructors may draw on these inquiries to structure class discussion, to construct quizzes or exams, or for assignments.

In posing questions and constructing exercises we are sensitive to the fact that readers of the book have diverse disciplinary backgrounds. Consequently, many of the inquiries are presented in a manner that allows for tailoring the questions to the reader’s particular field of expertise. Rather than imposing a particular concept or theory on a methodological issue we might ask readers to choose a concept or theory with which they are familiar and employ it to address a question in their course of study.

An introductory textbook of modest length must deal with topics in an expeditious fashion. Accordingly, we have omitted many qualifications, caveats, and citations to the literature in favor of a streamlined approach. Although the treatment in this text is somewhat more detailed than that found in many textbooks it is still highly selective when placed within the context of scholarly work on these subjects. This is the cost of writing a short book on a long subject. Readers who choose to continue in some branch of social science should view this book as a point of departure on their methodological journey. To that end, each chapter ends with a list of suggested readings and web sites related to topics broached in that chapter, under the heading Resources. Consider these references as an invitation to further study.
I. BUILDING BLOCKS

This part of the book is focused on fundamental elements of social science, elements that form building blocks for everything else. In Chapter 1, we lay out the rationale for a unified approach to our subject, social science methodology. In Chapter 2, we discuss social science arguments, with primary attention to descriptive and causal arguments. In Chapter 3, we turn to the topic of conceptualization and measurement. In Chapter 4, we discuss the generic features of empirical analysis.
1. **A Unified Framework**

The purpose of social science is to make a difference in the world by applying reason and evidence to problems of general concern. Every question of social science relates (or ought to relate) to normative concepts such as justice, democracy, human rights, prosperity, happiness, or equality.

What distinguishes social science from casual conversation, journalism, or political rhetoric may be summarized as follows. First, social science involves the systematic application of reason and evidence to problems with explicit attention to method and to possible sources of error. Second, social science is accompanied by realistic estimates of uncertainty with respect to whatever conclusions are drawn from the evidence. Third, social science attempts to provide a comprehensive treatment of a subject within whatever scope-conditions are defined by the study. All relevant information should be included; none should be arbitrarily excluded. Finally, social science adopts a disinterested posture with respect to all goals except the truth. Its purpose is to get as close to the truth as possible, in all its complexity, rather than to provoke, entertain, elucidate moral truths, or advance polemical claims.

These features render social science less stimulating than other media, where there is generally a premium on brevity, accessibility, provocation, righteousness, or humor. Social science is a sober profession. However, for those excited by the prospect of getting it right, and willing to expend some energy to get there, the practice of social science may be highly rewarding.

Consider the problem of crime, a topic that often evokes hot rhetoric and strong opinions. Most media reports and political speeches offer little useful information about the prevalence of crime, its sources, and its potential solutions. Instead, they exploit the public’s fascination with gruesome events and, in the process, provoke fear. From this perspective, the cold gaze of social science offers some relief.

Researchers have spent a good deal of time studying the rise and fall of violent crime in the United States and elsewhere. In the early 1960s, the US enjoyed a low homicide rate of 5 murders per 100,000 inhabitants. Over the next two decades this rate doubled – to 10 per 100,000 inhabitants – peaking in the late 1970s or early 1980s, at which point the US could claim the highest rate of violent crime of any advanced industrial country. Subsequently, the crime wave began to fall, and it now rests approximately where it was in 1960. What factors might explain this extraordinary rise and subsequent decline? What impact did the rise-and-fall of crime have on attitudes (e.g., towards immigrants and minorities) and on behavior (e.g., voting turnout and party affiliation)?

Those who study crime crossnationally also rely on murder rates to measure overall crime. Although crossnational statistics are prone to error, the greatest over-performers and under-performers are evident. At present, the highest violent crime rates in the world are found in Belize, Cote d’Ivoire, El Salvador, Guatemala, Honduras, Jamaica, and Venezuela – where there are 38-96 murder victims per 100,000 inhabitants every year. By contrast, murder rates in twenty-five countries are equal to, or less than, one per 100,000. This is an extraordinary range of variation, and it is only partly a product of economic development. Note that the murder capitals identified above are by no means the poorest countries in the world, and many relatively poor countries have murder rates of

---


2 Latzer (2016).
less than 3 per 100,000 – including Algeria, Armenia, Azerbaijan, Bangladesh, Bhutan, China, Egypt, Fiji, Iran, Jordan, Maldives, Micronesia, Nepal, Sao Tome, Tajikistan, Tonga, and Vanuatu. Another fascinating puzzle.

These questions are causal. But if we probe a bit we will quickly encounter issues of conceptualization and measurement. How shall we define criminal activity? Is murder a useful proxy for crime in general? What distinguishes murder from politically motivated acts of violence such as those accompanying terrorism or civil insurrection? (Is the Oklahoma City bombing, which claimed the lives of 168 Americans in 1995, a multiple homicide, or an act of domestic terrorism?) How has the definition of crime changed over time? How does it differ across countries or across regions within a country? How is crime understood within different communities?

These are the sorts of questions social science aims to address, and they are highly consequential. Improvements in our understanding of crime should help us to design better criminal-justice policies. Does community policing work? Does cleaning up visible manifestations of lawlessness in a neighborhood (e.g., fixing broken windows) affect the crime rate in that neighborhood? How effective are deterrents such as harsh jail sentences or capital punishment? How effective is the alternative approach based on rehabilitation of convicted criminals? Do features of our educational system affect the propensity of children to engage in criminal activity? Is crime rooted in socioeconomic deprivation? How is it affected by different social policies? Do different policy solutions work in different contexts, or for different sorts of criminal activity?

Those interested in questions like these should also be interested in social science methodology. The reason is that complex questions elicit debate among scholars. To understand this debate – to see why researchers agree and disagree and to make a determination about which is most believable – one needs to understand the nature of the theories and the evidence employed to evaluate theories and test related hypotheses.

Of course, most citizens and policymakers do not spend a great deal of time reading social science. Instead, they read journalistic accounts of social science research. There is surely nothing wrong with this. At the same time, one must bear in mind that newspaper articles and blog postings rarely explain the sort of background considerations that would allow one to informatively choose among rival conclusions about the same subject. This is not their fault; it is a limitation of the genre. The attraction of journalism is that it offers a brief account of a complex subject, suitable for consumption over breakfast, in the car, or on the train. If one wishes to go deeper – to read the reports upon which journalistic accounts are based – one must have a passing knowledge of social science methodology. (One would hope that journalists who offer pithy summaries of social science work also possess that deeper knowledge.)

Methodology should not be confused with a mastery of facts. While the latter is important, it is by no means sufficient to a determination of truth. Indeed, when experts disagree it is rarely over the facts of a case. It is, rather, over how those facts should be interpreted. An understanding of methodology involves an understanding of the logic of inquiry, i.e., the way in which one reaches conclusions from a body of evidence. This is what an informed consumer of social science must have if she is to decipher social science work on a subject.

For those who aim to become producers of social science the importance of methodology is even more apparent. Anyone who is dissatisfied with the field of criminology as it now stands would do well to design their own study. And designing such a study will require considerable training in the wiles of methodology if the result is to add anything to our knowledge of this complex subject. Methodology thus lies at the heart of contemporary political debates, providing the set of tools by which we might tackle social problems in a rational fashion.

\(^5\) UNODC (2011).
THE PURPOSE OF UNITY

This book embraces a broad view of social science. It encompasses work that is primarily descriptive as well as work that is primarily causal. It encompasses work that is experimental (involving a randomized treatment) and observational (i.e., nonexperimental). It encompasses quantitative and qualitative research. It encompasses a range of strategies of data collection, from standardized surveys to ethnography.

The book is also intended to encompass a wide range of disciplines, including anthropology, communications, criminal justice, economics (and subfields such as business, finance, management), education, environmental policy, international development, law, political science, psychology, public health, public policy, social work, sociology, and urban planning. Although these fields focus on different substantive problems, the methods they employ – and the methodological obstacles they encounter – are quite similar. Indeed, there is almost as much methodological diversity within a single discipline such as anthropology, sociology, or political science as there is across these disciplines.

Of course, there are many ways to do good social science. Sometimes, it makes sense to combine diverse methods in a single study – a multi-method approach to research (see Chapter 10). In any case, much depends on the nature of the evidence available and the nature of the question under investigation. It would be folly to propose a uniform method or theoretical framework for all of social science, or even for a single discipline. Methods pluralism is easy to justify. Indeed, it is impossible to avoid.

However, beneath the diversity of methods there is (or at least ought to be) a degree of methodological consensus. Note that if standards of truth are understandable only within the context of specific fields or theoretical traditions there is no way to adjudicate among contending views. Each truth becomes entirely self-reflective. Thus, while it is reasonable to cultivate a diversity of tools, it is unreasonable to cultivate a diversity of methodological standards. A discovery in sociology ought to be understandable, and appraisable, by those who are not sociologists; otherwise, it cannot claim the status of truth. Nor will it be of much use to anyone outside of sociology.

Moreover, as a matter of good scholarship, writers in the social sciences ought to be able to converse with one another. Economists interested in political economy should be cognizant – and should seek to incorporate, wherever possible – work in political science. And vice versa. Even arguments demand a common frame of reference. Without such shared ground they are merely statements of position. Here, science degenerates into a chorus of yea’s and nay’s reminiscent of Monty Python’s “Argument Clinic” sketch.

This is why the natural scope for the present volume is social science writ-large rather than a single field or subfield. Thinking about methodological topics in diverse settings forces us to think in new ways, to justify our choices on methodological grounds rather than on grounds of convenience or familiarity. It is not sufficient for sociologists to say that they do things in a certain way because that’s what they have always done. Likewise for economists, political scientists, and the rest of our quarrelsome band.

---

* Our perspective echoes the perspective of an influential book edited by Henry Brady and David Collier (2004), subtitled *Diverse Tools, Shared Standards*.
* This episode can be accessed via YouTube.
Accordingly, this book aims to provide a framework that reaches across the social sciences, providing common ground for those engaged in diverse topics and diverse research methods. We have looked to uncover the shared norms that govern activity – implicitly or explicitly – in the community of social scientists. What makes a work of social science true, useful, or convincing (“scientific”)? Why do we prefer one treatment of a subject over another? These are the sorts of ground-level judgments that define the activity of methodology. With these judgments, we hope to identify the threads that tie our methodological intuitions together into a relatively unified framework across the disciplines of social science.

Our approach centers on the identification of basic tasks of social science, strategies enlisted to achieve those tasks, and criteria associated with each task and strategy. These are laid out schematically in tables throughout the book.

Note that each task and criterion is viewed as a matter of degree. Achieving precision, for example, is not an either/or proposition. One tries to obtain as precise an estimate as possible, in full knowledge that there will always be some element of imprecision (variability). The same goes for other tasks and criteria. Note also that the tasks, strategies, and criteria laid out in the subsequent pages are sometimes in conflict with one another. For example, theories aim for both precision and breadth; however, achieving one may involve sacrifices for the other. Methodological tradeoffs of this sort are ubiquitous. This means that every task, strategy, or criterion must be understood with a ceteris paribus caveat. Precision is desirable, all other things being equal.

Although a relative and multidimensional standard may seem rather open-ended, this does not imply that anything goes. It means that the researcher must search for the theory and research design that maximizes goodness along a set of (relatively fixed) dimensions, reconciling divergent demands wherever possible. The goodness of a theory or research design is therefore judged only by reference to all possible theories or research designs that have been devised, or might be devised, to address the same research question. Best means best possible.8

This allows for all sorts of theories and research designs to enter the social science pantheon without shame or disparagement – but only if no better expedient can be found. It supposes that studies with weak theories or evidence answer a very difficult question: could an argument or research design be improved upon? What is achievable, under the circumstances?

If a research ideal is entirely out of reach – by virtue of lack of data, lack of funding sources, lack of cooperation on the part of relevant authorities, or ethical considerations – it is pointless to admonish an author for failing to achieve it. Perfection becomes the enemy of scientific progress. We must guard against the possibility that work adding value to what we know about a subject might be rejected even when no better approach is forthcoming. Standards must be realistic.

If, on the other hand, a better approach to a given subject can be envisioned and the costs of implementation are not too great, a study that chooses not to utilize that demonstrably better approach is rightly criticized. We must guard against the possibility that second-best approaches will drive out first-best approaches simply because the former adopt easier or more familiar methods. Mediocrity should not be the enemy of excellence. This is what we mean by best-possible, under the circumstances.

Equally important is to embrace the uncertainty of our enterprise, honestly and forthrightly. Weaknesses in design and analysis should be openly acknowledged rather than hidden in footnotes or obscured in jargon and endless statistical tests. This is important not just as a matter of intellectual honesty but also for the long-run development of the social sciences. The cumulation of knowledge in a field depends more on methodological transparency than on “statistically significant” results.

8 Lieberson and Horwich (2008).
The following chapters intersperse abstract methodological points with specific examples. While these examples vary, we draw repeatedly on three subjects that have played a key role in contemporary social science and in recent methodological debates: worker-training programs, social capital, and democracy. Readers who are unfamiliar with this terrain may use the following sections to acquaint themselves with these subject areas – though we do not pretend to offer anything like a comprehensive review (for further reading, see Resources).

While each has its disciplinary home turf – economics, sociology, and political science, respectively – it should be appreciated that these disciplinary categories are increasingly fluid. Economists, sociologists, and political scientists have worked on all three issue-areas. And these subjects are also important for cognate fields such as business, education, public policy, and social work. In this sense, our exemplars encompass the far reaches of social science.

Readers should also be aware that the three topics exemplify very different kinds of social science work. The first embodies a specific causal intervention – participation in a worker-training program – that operates on an individual level. We utilize this example frequently because many methodological principles are easier to discuss at the individual level. The other two topics embrace broader and more diffuse social and political institutions that are usually understood to operate at a societal level.

WORKER-TRAINING PROGRAMS

Unemployment is a problem not only for those who find themselves without a job but also for society at-large, who must bear the costs of supporting the unemployed (provided there are systems of relief, either private or public) and must bear the negative externalities brought on by unemployment (e.g., an increased tendency for criminal activity). The public policy question is how governments can best deal with this byproduct of capitalism.

One approach centers on worker-training programs. These programs enroll unemployed, or under-employed, persons with an attempt to boost their job-relevant skills. Programs may also seek to enhance morale and to educate participants in job-search strategies and workplace norms. Programs may be short in duration, or longer-term. They may be administered in conjunction with an apprenticeship. They may be accompanied by incentives for employers to participate. In short, there is great variety in the implementation of this category of social program directed at the unemployed.

The key question of interest is whether participation in such a program enhances a person’s probability of finding a job or enhances their long-term earnings. Insofar as there may be such an effect, we wish to know why – that is, the mechanisms through which the causal effect operates. Is it because participants are more persistent in their search for work? Is it because they have better skills, better morale, or better workplace behavior? Is it because employers view participation in a program as a sign of motivation? Many explanations might be offered.
For present purposes, what bears emphasis are the methodological properties of this field of research. There is, first of all, a key concept – the worker-training program, which seems fairly clear in most settings but is actually rather blurry around the edges. Does a one-day program focusing on advice for job-hunting qualify? How about a person who enlists government support to take classes at a community college? How about a program that emphasizes job placement with relatively little emphasis on training? There is, second, the hypothesis – that participation in such a program enhances employment and salary. There is, third, the theory, which concerns all the reasons that the hypothesis might be true (if indeed it is true).

SOCIAL CAPITAL

Our second example, centering on the concept of social capital, is considerably more complex. We shall define social capital as the benefits that derive from social networks that extend beyond family and clan. Where networks are intensive and extensive, societies should experience higher trust, lower crime, better public health, better governance, and as a result of these first-stage benefits, stronger growth. Likewise, individuals with more extensive networks should experience greater benefits (e.g., more economic opportunities) than individuals with circumscribed networks.

Indicators of social capital include membership in voluntary associations (e.g., unions, fraternal and soror organizations, neighborhood associations, and clubs) and political engagement (e.g., voter turnout). These may be explored separately or combined in a single index.

Some years ago, Robert Putnam discovered that many indicators of social capital in the United States showed a marked downturn beginning in the 1950s, suggesting a deep and far-reaching decline in social capital.9 (Similar patterns were found in some other advanced industrial countries, though not quite to the same degree.10) This spurred a good deal of hand-wringing about the state of the union, along with many social science studies. Some of these studies showed a mixed picture – decline in some areas but not in others, or a re-direction of activity from some areas to other areas.11 Another interpretation is that the decline is real but largely a function of the extraordinary high level of social capital found among members of the “greatest generation” – those who came of age in the 1930s and 1940s. From this perspective, the postwar decline represents a return to a normal level of social capital. The controversy has been difficult to resolve because most of the available measures of social capital stretch back only to the mid-twentieth century; thus, we have only a vague sense of the level of social capital existing in the United States prior to the 1940s.

Another set of controversies concern the causes of this decline. Are they the product of a general disenchantment during the turbulent 1960s, the entry of women into the labor force (pulling them away from social networking activities), migration, suburbanization, increasing diversity, or changing technologies (especially television and the internet)?

Still another set of controversies concern the possible effects of this decline. At first, the decline of social capital was linked to a rise in the crime rate. The rate of violent crime began to decline in the 1990s, however, casting doubt on a possible link between social capital and crime. The decline of social capital may also be linked to social and political instability, though evidence of such effects are thin. A third sort of effect may be decreasing concern for others, as manifested in lower public support for welfare programs intended to help less privileged members of society. Finally, one may conjecture that declining social capital imperils the willingness of citizens to support government, as manifested in anti-tax crusades and declining faith in political institutions.

---

Leaving aside various controversies that attend the “decline of social capital” thesis, let us take a moment to consider the possible impact of social capital on governance and economic development more generally. Putnam’s first book on the subject argued that differences in social capital between the northern and southern regions of Italy could account for differences in the quality of governance across the (well-governed) north and the (poorly-governed) south.12 Specifically, where reciprocity-relationships were extensive and social trust was high this boosted the quality of government. Where social networks were limited to the extended family and social trust was low, as it seemed to be in the southern regions of Italy, it was difficult to establish effective government. This had repercussions for growth and that is why, Putnam reasoned, we see a prosperous north and a much less prosperous south. One can also hypothesize that there might be direct effects from social capital to growth.13 For example, where networks are limited and trust is low, markets are more difficult to maintain, competition is likely to be limited, and transaction costs will be high. Indeed, scholars have argued that the strength or weakness of social capital is a key to long-term patterns of development around the world.14

In recent years, proponents of social capital have confronted the apparent fact that there are “good” and “bad” sorts of social capital. It is often noted that gangs are a voluntary network of individuals who prey on society. Likewise, neighborhood associations sometimes form in order to exclude social groups deemed threatening to the community. At the extreme, race riots may be understood as an expression of social capital. Indeed, Weimar Germany, which spawned the xenophobic ideology of Nazism, was a society rich in extra-familial social networks.15 In response, theorists now distinguish between “bonding” and “bridging” social capital. The first relates to social networks among people who are similar to each other – ethnically, socioeconomically, and so forth. The second refers to social networks that reach across social divides. The claim is that these two types of social capital have divergent effects on a variety of outcomes. In this fashion, a significant modification of the original theory is introduced.

Of course, these matters are complicated. What we have offered above is a brief review of a large and complex literature. Our purpose is not to represent the entirety of these debates but merely to illustrate several key elements of social science argumentation. Note, first, the key concept, social capital, and various indicators that have been used to measure it. Note, second, the descriptive hypothesis that social capital has declined in recent decades in the United States (and perhaps elsewhere). Note, finally, various hypotheses about the causes and effects of that decline and theoretical expectations about why (i.e., the mechanisms by which) social capital might lead to enhanced governance and economic development.

DEMOCRACY
Democracy refers generally to rule by the people. Below this level of abstraction, there is great debate about how to best define this key concept. Most definitions include the idea of electoral contestation. That is, in order to be considered democratic a polity must allow free and fair elections with a broad electorate; those elected must be allowed to take office; and elective bodies must not be constrained by unelective bodies such as a military tribunal or monarch. Additional attributes such as constraints on the exercise of power, civil liberty, political equality, deliberation, and full participation might also be included in a definition of this key concept.

15 Berman (1997).
There are a variety of crossnational indicators of democracy. However, most of these empirical measures focus on the electoral component of the concept, as set forth above. Most also regard democracy as a matter of degrees, stretching from autocracy (i.e. dictatorship, authoritarian rule) to full democracy. This includes the widely-used indices produced by Polity (“Polity2,” a 21-point scale) and Freedom House (“Political Rights,” a 7-point scale).16

Sometimes, however, it is important to divide up the world of polities into those that are (predominantly) autocratic and those that are (predominantly) democratic. The most widely-employed binary indicator (0=autocracy, 1=democracy) is the Democracy-Dictatorship (DD) index developed by Adam Przeworski and colleagues.17 Accordingly, a regime is a democracy if leaders are selected through contested elections. To operationalize this conception of democracy the authors identify four criteria:

1. The chief executive must be chosen by popular election or by a body that was itself popularly elected.
2. The legislature must be popularly elected.
3. There must be more than one party competing in the elections.
4. An alternation in power under electoral rules identical to the ones that brought the incumbent to office must have taken place.18

All four conditions must be satisfied in order for a polity to be considered democratic.

With respect to democracy, it is helpful to distinguish several sorts of research questions. First, what is the empirical pattern of democratization throughout the world? Samuel Huntington discerns three democratic “waves” in the contemporary era – the first beginning in the early nineteenth century, the second after the conclusion of World War II, and the third beginning in the 1970s.19 Criticism of this account centers, first of all, on Huntington’s definition of democracy. If it is broadened to include female suffrage and informal impediments to suffrage for males (e.g., “Jim Crow” laws in the American South), historical patterns of democratization look rather different.20 A second issue concerns the denominator – the total number of countries under examination. Conventionally, these are defined as sovereign nation-states. However, it will be noticed that the number of sovereign states expands rapidly over the observed period, especially in the 1960s when most of Africa is liberated from colonial rule. This means that an apparent downturn in the rate of democracy may be due to a statistical artifact: the momentary increase in the number of countries considered as part of a global sample.21 A final issue concerns how to regard the concept of a “wave.” Does it refer to changes in the global level of democracy (as measured by Polity, for example), to net-transitions to democracy, or to linkages among cases of democratization (diffusion)?22

---

16 Freedom House employs two indices, “Political Rights” and “Civil Liberties” (sometimes they are employed in tandem, sometimes singly) each of which extends back to 1972 and covers most sovereign and semi-sovereign nations (see www.freedomhouse.org). Polity IV (Marshall and Jaggers 2007) also provides two aggregate indices, “Democracy” and “Autocracy,” usually used in tandem (by subtracting one from the other), which provides the Polity2 variable. Coverage extends back to 1800 for most sovereign countries with populations greater than 500,000 (www.cidcm.umd.edu/inscr/polity).
17 Cheibub, Gandhi, Vreeland (2010).
21 Doorenspleet (2000).
Causal questions begin with democracy’s rise. What might account for the pattern of
democratization that we see across the world over the past two centuries? Structural (distal)
explanations are grounded variously in geography, colonial history, religion, ethnicity,
modernization, and particular types of authoritarian rule. Proximate causes include features of the
transition itself such as whether liberalization (civil liberties, constraints on executive power) occurred
prior to an electoral opening or whether a pact was established among important political players.
None have been conclusively established.23

A separate set of causal questions concern democracy’s causal effects. Does a transition (or
improvement in the quality of democracy) bring with it improvements in the quality of governance
(e.g., less corruption, more provision of public goods)? Does it foster higher levels of education,
health, and infrastructure, or greater equality across the sexes and across ethnic groups within a
society? These outcomes might be summarized in the phrase, does democracy bring development?
Again, we find vast disagreement.24

A somewhat separate question concerns whether regime-type influences the conduct of
foreign policy. According to the well-known theory of the democratic peace two countries that are
democratic should never fight wars with one another.25 A softer version of this thesis interprets the
matter probabilistically: two countries that are democratic are less likely to fight wars with each other
than any other pairing of countries.

Again, it must be stressed that our purpose is not to offer a comprehensive overview of this
immense subject but simply to display some of the methodological properties of the debates that
occur around the subject. In particular, we have outlined debates over the key concept (democracy),
various indicators for that concept (including Polity, Freedom House, and Dictatorship-Democracy),
an influential descriptive account (Huntington’s three waves of democratization), and two general
causal questions, one pertaining to the causes of democratization and the other to its effects.

**Conclusions**

The examples introduced above were chosen because they are prominent and also highly
contentious. Scholars have differing views of the definition and measurement of key concepts, the
descriptive features pertaining to the phenomenon, and/or the causes or effects of that
phenomenon. This is especially the case for social capital and democracy, and somewhat less so for
worker-training programs.

Some may feel that these ongoing debates are an indication of the weakness of social science
– its failure to reach closure, even with questions as old as democracy. While this is certainly an
abiding characteristic of social science one must also bear in mind that our difficulties in reaching
consensus arise primarily from the nature of the problems themselves. If we were to choose simpler
problems we would no doubt arrive at greater consensus with respect to their answers. Indeed, the
simplest problem among our three examples – worker-training programs – is also the one that has
gained the greatest scholarly consensus.

It is not clear that society would be better served if social science narrowed its focus to
tractable questions, ignoring the macro-level features that – presumably – operate over long periods,
at macro levels, and sometimes beneath the surface of social life. This includes social capital and
democracy, along with other similarly diffuse topics.

---

23 Coppedge (2012).
24 Carbone (2009).
In any case, we hope that these examples serve as an invitation to our topic – social science methodology. For, it is only by understanding the methodological properties of these topics that we can hope to understand these debates – and, perhaps, over time, to attain greater consensus.
**KEY TERMS**

- Social science
- Methodology
- Multi-method
- Worker-training programs
- Social capital
- Democracy

**INQUIRIES**

*Note: The following questions, some of which are touched on in the chapter, may be fruitfully discussed after reading this chapter and/or after completing the book – at which point the reader may have a better idea of how they apply to the work of social science.*

1. What is social science? How is it distinguished from natural science (the study of non-living things and of the non-decisional or biological aspects of human beings)? (Should it be distinguished from natural science, or considered as a branch of natural science?) How is it distinguished from the humanities (the study of history, philosophy, literature, and the arts)? How is it distinguished from journalism or casual conversation?

2. Can social life (e.g., crime, worker-training programs, social capital, democracy) be studied objectively? What does it mean to study something “objectively”? 

3. Should social science attempt to study single events, classes of events, or try to encompass both of these objectives? For example, should we study the 9/11 bombing of the World Trade Center or terrorism in general? Should we study World War I or wars in general? Should we study the transition to democracy in Tunisia or democratization (in general)? Are these goals in conflict with one another, or in harmony?

4. Is social science a misnomer? Can decisional aspects of human behavior be studied scientifically? What does it mean to study phenomena like voting, buying/selling, learning, friendship in a scientific manner?

5. Is the current fragmentation of social science a problem? Or is it a sign of a healthy level of specialization and diversity?

6. What is the difference between methods-pluralism and methodological pluralism?

7. Does social science matter? Does it influence politics and policy? Is it more influential in some countries than in others (and if so, why)?
8. Tradeoffs are endemic in social science. For example, the same features of a research design that might assist in coming up with new ideas about a subject may also impede effective testing of that new idea. Think of some other examples of conflicting methodological goals. Is it a cop-out to say that something is desirable, “all other things being equal” when in fact other things are rarely equal? Is the “trade-ups” notion fanciful?

**RESOURCES**

**WEB SITES**
- Consortium on Qualitative Research Methods (CQRM) - [http://www.maxwell.syr.edu/movnihan/programs/cqrm/](http://www.maxwell.syr.edu/movnihan/programs/cqrm/)
- Methods Coordination Project [On-line bibliography of sources covering many areas of social science methodology.] [https://qdr.syr.edu/mcp](https://qdr.syr.edu/mcp)
- OPOSSEM – Open Portal for Social Science Education in Methodology - [http://opossem.org/](http://opossem.org/)
- Society for Political Methodology - [http://polmeth.wustl.edu/](http://polmeth.wustl.edu/)

**HISTORY AND PHILOSOPHY OF SOCIAL SCIENCE**

**WORKER-TRAINING PROGRAMS**
DEMOCRACY


SOCIAL CAPITAL

2. **Arguments**

Social science is organized around arguments. This is what we hope to prove, or disprove, by an empirical analysis. Science proceeds as arguments are advanced, revised, and proven or disproven.

A **theory** is the reasoning behind an argument. It explains a relationship, its mechanisms (if the relationship is causal), scope-conditions, background conditions, and any additional information needed to interpret the argument. One might theorize that democratization enhances economic growth because it holds leaders accountable, and accountable leaders are more likely to adopt policies that serve the public good.

If a theory is presented in a formal manner, perhaps with a set of mathematical expressions, it may be referred to as a **model**. At present, however, we do not distinguish between an argument, theory, or model; these terms will be used more or less synonymously.

A **hypothesis** is the specific, testable element(s) of any argument. One might hypothesize that as a country becomes more democratic its GDP growth increases. Our use of the term "argument" is thus meant to encompass both theory and hypothesis.

This chapter begins by distinguishing various types of social science argumentation. Arguments central to social science are generally classifiable as either descriptive or causal, so we spend considerable time on these genres. Next, we contrast these genres with other sorts of arguments such as those that are predictive, normative, or prescriptive. Finally, we lay out the characteristics of a good descriptive or causal argument. (Additional criteria specific to causal arguments are postponed until Part II of the book.)

**Descriptive Arguments**

A **descriptive argument** describes some aspect of the world. In doing so, it aims to answer **what** questions (e.g., **when**, **whom**, **out of what**, **in what manner**) about a phenomenon. Descriptive arguments are about what is or what was.

Note that many features of the world are intrinsically important. We want to know which countries are democratic and which are not, and in what ways they are democratic or un-democratic. We want to know how many Jews, gypsies, gays, and socialists were killed in the European Holocaust. We want to know the intellectual origins of individualism, and how it evolved through time and across cultures. We want to know whether the media in a country presents a biased view of news events, and in what direction that bias runs.

Descriptive arguments that are general in nature (they apply to a large number of events) may be classified as **associations, syntheses, or typologies**. Typologies, in turn, may be understood as **simple, periodization, matrix, or taxonomy**. This is how social scientists carve up nature. These are the patterns that we look for when attempting to describe events in the social world.

These various genres may be arranged within a taxonomy, as shown in Figure 2.1. However, the taxonomy probably won’t make much sense until after these concepts have been explained, so the reader is advised to return to this figure when s/he has finished this section of the chapter.
Figure 2.1: Taxonomy of Descriptive Arguments

- **Descriptive arguments**
  - **Non-grouping**
    - Associations
  - **Grouping**
    - One category
    - Multiple categories
      - Syntheses
      - Typologies
        - Simple
        - Periodization
        - Matrix
        - Taxonomy
**Associations**

A descriptive argument that involves a relationship among several factors is **associational**. Many studies have focused, for example, on the degree to which involvement in politics is skewed toward the middle and upper classes. That is, they are focused on the association between social class and political engagement.26

We might hypothesize that where the value of one factor, \( X \), is high, the value another factor, \( Y \), is high and correspondingly where \( X \) is low, \( Y \) is low, as depicted in Panel (a) of Figure 2.2. Or we might claim that they are inversely related – where \( X \) is high, \( Y \) is low, as depicted in Panel (b). We might also hypothesize that \( X \) and \( Y \) are related in a curvilinear (non-linear) pattern such as an inverted \( U \), as illustrated in panel (c). There are an infinite number of ways in which two factors may be related, and things naturally become more complicated if additional factors (beyond \( X \) and \( Y \)) are integrated into the argument.

Figure 2.2: Possible Relationships among Two Factors

(a) Direct relationship

(b) Inverse relationship

(c) Curvilinear relationship
Trend analysis seeks to discover a relationship between a phenomenon and the passage of time. For example, Robert Putnam finds that social capital in the United States has declined precipitously since the 1930s and 1940s. This is demonstrated by examining patterns in political engagement (turnout and other forms of political activity) and membership in unions, fraternal and sororal organizations, churches, and neighborhood associations such as bowling leagues, school associations, and clubs. Many of these indices show a marked downturn since the mid-twentieth century, suggesting a widespread decline in social capital.27

Network analysis focuses on interrelationships among many units (which may be understood in spatial, temporal, or functional ways). Researchers have studied political networks, corporate networks, networks extending across business and politics, terrorist networks, transportation networks, as well as social networks, including those built around web sites such as Facebook (see Chapter 5). A common focus of network analysis is the centrality of various nodes, i.e., the extent to which an individual (or group) serves as a communicative hub within a network of individuals (or groups). Scholars also attempt to determine what factors predict an individual’s position within that network, and what implications the structure of a network might have on various outcomes (e.g., the probability of finding a marriage partner or influencing a collective decision).28

Syntheses
Some multidimensional arguments attempt to group together diverse dimensions into distinct categories. If there is only one category of interest (others are not well-defined), this style of argument may be called synthetic. The claim here is that diverse attributes of a topic revolve around a central theme that unifies the attributes, lending coherence to an otherwise disparate set of phenomena. Synthetic arguments are therefore holistic endeavors, emphasizing similarities rather than differences among the chosen sample of cases.

This type of argument is often applied to cultural analysis. For example, Louis Hartz (building upon Tocqueville) argues that American political culture is “liberal” – a term he understood as an amalgam of individualism and antistatism. This large-order concept helps to organize many features of American politics and society going back (he argues) to the Revolution, including the absence of redistributive measures, the weakness of labor and socialism, the minimal presence of the national state, and the contrasting robustness of civil society. It is also consistent with a longstanding narrative about American life, visible in political speech and in other popular media, that praises individual freedom and equal opportunity while decrying the corruption of politics.29 Whether this ideology explains political outcomes such as the failure of socialism is, for present purposes, not important. It is sufficient to note that a large number of relatively distinctive features of American history fall into place when placed within the rubric of the liberal tradition.

A second example of social-scientific synthesis is provided by Orlando Patterson’s acclaimed study of slavery. Patterson begins by showing that through history slaves have occupied diverse roles in society. They have been entrusted with the management of large commercial enterprises and the leadership of empires, though they are routinely allocated to menial tasks. Accordingly, slaves have been rich and they have been poor. What all of these experiences have in common – indeed, about all they have in common, according to Patterson – is that slaves are deprived of a social identity distinct from their owners. As property, their identity is not unique to them. It is this feature of

28 Jackson (2010).
29 Hartz (1955), Tocqueville (1945).
“social death” that constitutes the distinctive (and therefore defining) feature of slavery throughout history.  

These are very different sorts of descriptive syntheses. But they both aim to satisfy the same methodological goal, i.e., to summarize many attributes and many phenomena in a single concept or phrase. Of course, the attempt to synthesize is also, at the same time, an attempt to differentiate. For example, the liberalism of American culture is contrasted with the non-liberal cultures of Europe. The social death of slaves is contrasted with the social life enjoyed by free people. Insofar as these distinctions are explicit and insofar as they provide the grist for extensive empirical analysis a synthesis begins to look more like a typology – our next topic.

SIMPLE TYPLOGIES

Where multiple discrete categories are defined, the result is a typology. Here, the goal is to sort phenomena into categories that are mutually exclusive and exhaustive on the basis of a consistent categorization principle (or principles). Let us begin with some examples.

Max Weber argues that political authority draws upon three alternate forms of legitimacy: traditional, charismatic, or rational-legal. Traditional authority derives from conformance with established custom. Thus, a king in a country with a longstanding monarchy is regarded as legitimate because his position accords with the accepted practice of rulership (assuming, that is, that he is the legitimate heir to the throne). Charismatic authority derives from a leader’s personal attributes, which may be imbued with special and perhaps supernatural powers. Leaders who elicit extraordinary devotion and whose position stems not from birth or from their position within a hierarchy are often regarded as charismatic. Rational-legal authority is derived from a leader’s position within a constitutional (rule-of-law) system. The leader’s power is contingent upon the office that s/he holds, and is bound by the strictures pertaining to that office. Once s/he steps outside the office, or steps beyond the legal bounds of the office, s/he loses authority. In contrast to charismatic authority, rational-legal authority is impersonal. In contrast to traditional authority, it is rational in the sense of being constituted within a web of rules that are logically consistent rather than arbitrary or ad hoc.

Simple typologies of this nature are ubiquitous. Aristotle classifies polities as monarchies (rule of one), oligarchies (rule of a few), or democracies (rule of many). Samuel Finer, working over the same subject many years later, organizes polities according to the nature of their rulers: Palace, Church, Nobility, or Forum. Albert Hirschman argues that the influence of constituents on organizations may be felt through exit and/or voice. Gosta Esping-Andersen divides the world of welfare regimes into those that are liberal, corporatist, or social democratic.

More complex typologies may be understood as periodizations, matrix typologies, or taxonomies.

---

30 Patterson (1982).
31 In a later study, Hartz (1964) develops a “fragment” thesis to explain political cultures in settler societies such as the U.S., Canada, Australia, and South Africa.
32 Confusingly, three words are often used semi-synonymously: typology, classification, and taxonomy. In my adopted usage, “taxonomy” refers to a specific kind of typology.
34 Finer (1997).
PERIODIZATION

A typology that is temporally ordered may be understood as a periodization. For example, it is argued that several waves of democratization have advanced across the world over the past two centuries, each with distinctive features.37 The first wave began in the early nineteenth century in the United States, where white males were granted the right to vote in the first mass-suffrage elections. This wave crested after World War One, when the rise of fascism in Europe temporarily beat back the tide of democracy. The second wave began after World War Two with the defeat of fascism and the rise of many newly independent nations – some of which were (at least briefly) democratic. The third wave began with the overthrow of dictatorship in Portugal in 1974, extending to Latin America and the Asia-Pacific in the 1980s, and Eastern Europe and Africa in the 1990s.

Grand periodization schemes have absorbed many thinkers. Tocqueville proclaimed the beginning of a democratic age sometime in the late-eighteenth century, which may be compared with the previous feudal or aristocratic ages. Along these lines, Marx proposed to typologize recorded human history into feudal, capitalist, and communist stages.

Other periodization schemes focus on a single country. For example, many students of American political history are convinced that fundamental political changes have occurred only episodically, during “realignment” periods.38 Others defend an older tradition, dividing American political history into “eras” (Revolutionary, Jacksonian, Civil War, Reconstruction, et al.). Still others argue that the topic is best approached through an even more differentiated periodization defined by presidencies.39

Each attempt to establish a periodization appeals to the same general desiderata, i.e., to identify key points of change within a historical topic such that the resulting periods are mutually exclusive and exhaustive (along whatever dimensions are of interest to the study).

MATRIX TYPOLOGIES

When a typology is formed by the intersection of several categorical variables the result is a matrix typology. Following Robert Dahl, we can posit that there are two core components of democracy, contestation and participation. The intersection of these two factors produces four types, which Dahl labeled (a) closed hegemony, (b) inclusive hegemony, (c) competitive oligarchy, and (d) polyarchy, as illustrated in Table 2.1.40

Matrix typologies may contain any number of factors, resulting in any number of compound types (cells). However, the two-by-two matrix is still the most common – presumably because adding a third (or fourth) dimension does not usually result in discrete and recognizable types.

---

38 Sundquist (1983).
39 For discussion, see Cochran (1948), Zelizer (2002).
40 Dahl (1971: 7). Another example of a matrix typology is Aristotle’s ancient typology of regime types (Lehnert 2007: 65). Here, the number of rulers (one, a few, or many) is cross-tabulated with the rulers’ goals (self-interest or the greater good) to produce six categories: tyranny, oligarchy, democracy, monarchy, aristocracy, and polity. Additional examples of matrix typologies related to the concept of democracy can be found in Almond and Verba (1963/1989: 16), Weyland (1995).
Table 2.1: A Matrix Typology: Regime-types

<table>
<thead>
<tr>
<th>Participation</th>
<th>Law</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Closed Hegemony</td>
<td>Inclusive Hegemony</td>
</tr>
<tr>
<td>High</td>
<td>Competitive Oligarchy</td>
<td>Polyarchy</td>
</tr>
</tbody>
</table>

**TAXONOMIES**

A typology that stretches in a hierarchical fashion across several levels of analysis may be referred to as a *taxonomy*. For example, one might stipulate that there are two basic polity types: autocracy and democracy. Among democracies, some are direct and others representative. Among representative democracies, one finds electoral, liberal, majoritarian, participatory, and egalitarian varieties. The nested quality of this family of terms may be illustrated in tabular format (see Table 2.2) or in a tree diagram (see Figure 2.3).41

Note that each subordinate level of the taxonomy possesses all the attributes of the superordinate category, plus one (or several). Each concept within a taxonomy may therefore be defined by specifying its superordinate category plus its differentiating attribute or attributes – its *genus et differentium*. (Concepts so defined are sometimes described as “classical” in reference to their Aristotelian lineage and their venerable place within the field of logic.)

---

41 As a second example, one might consider Reynolds and Reilly’s (2005: 28) taxonomy of electoral systems. The still-classic example of a taxonomy is the Linnaean system of biological classification (Linsley and Usinger 1959).
Table 2.2: Regime Taxonomy in Tabular Format

<table>
<thead>
<tr>
<th>TERMS</th>
<th>ATTRIBUTES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Gov. form</td>
</tr>
<tr>
<td>I. Polity</td>
<td>●</td>
</tr>
<tr>
<td>A. Autocracy</td>
<td>●</td>
</tr>
<tr>
<td>B. Democracy</td>
<td>●</td>
</tr>
<tr>
<td>1. Direct</td>
<td>●</td>
</tr>
<tr>
<td>2. Representative</td>
<td>●</td>
</tr>
<tr>
<td>a. Electoral</td>
<td>●</td>
</tr>
<tr>
<td>b. Liberal</td>
<td>●</td>
</tr>
<tr>
<td>c. Majoritarian</td>
<td>●</td>
</tr>
<tr>
<td>d. Participatory</td>
<td>●</td>
</tr>
<tr>
<td>e. Deliberative</td>
<td>●</td>
</tr>
<tr>
<td>f. Egalitarian</td>
<td>●</td>
</tr>
</tbody>
</table>
Figure 2.3: Regime Taxonomy in Tree-Diagram Format
CAUSAL ARGUMENTS

The reader may have noticed that many of the descriptive relationships introduced in the previous section could be reformulated as **causal arguments**. Instead of saying that $X$ and $Y$ are associated (correlated) with each other, for example, we might claim that $X$ *causes* $Y$. While the shift may appear subtle, this is more than a minor change of terminology. It is a fundamentally different argument.

A causal argument (or theory) involves at least two elements: a *cause* and an *outcome*. To say that a factor, $X$, is a cause of an outcome, $Y$, is to say that a change in $X$ generates a change in $Y$ relative to what $Y$ would otherwise be (the counterfactual condition), given certain background conditions (ceteris paribus assumptions) and scope-conditions (the population of the inference).

As an example, let us focus on the causal role of a worker-training program (as introduced in Chapter 1). A reasonable hypothesis is that participation in the program ($X$) will enhance a participant’s earnings ($Y$). If the relationship is causal, her earnings should be higher than they would be if she had never participated in the program, all other things being equal.

A causal argument implies a **counterfactual**. The counterfactual in this example is that the worker does *not* participate in the worker-training program. The counterfactual outcome is that she does not realize a gain in earnings (or experiences less of an increase than she would had she participated in the program).

Importantly, when one asserts that $X$ *causes* $Y$ one is asserting that the actual (ontological) probability of an event is increased by $X$, not simply a theory’s fit with the available data. This is what distinguishes a causal argument from description or prediction. To be causal, the factor in question must *generate*, *create*, or *produce* an effect.

Of course, it is not always possible to specify precisely why $X$ generates a change in $Y$. Yet, in identifying $X$ as a cause of $Y$ one is presuming the existence of a causal *mechanism* – understood here as the pathway or process or chain of intermediary variables by which $X$ affects $Y$. In our example, the mechanism might be the factual information about a particular occupation that is imparted in the worker training program. Or it could be the imparting of job search skills such as how to construct a resume. Or it could be general workplace comportment (how to relate to one’s boss and co-workers). Lots of potential mechanisms might be identified, and any combination of them may be at work if the relationship is causal.

Causal relationships occur against a **background** of other factors. These are conditions that make any causal relationship possible. One would not expect worker training programs to have any causal impact in a country where jobs (and promotions) are allocated by clientelistic networks or political patronage. Likewise, one would not expect a worker training program to affect earnings in a collapsing economy. So we might stipulate that the background conditions of the argument include an intact economy and meritocratic appointment and promotion practices. Unless otherwise specified, these background conditions are presumed to hold constant: they do not vary. This is known as the *ceteris paribus* (all else equal) assumption, and is implicit in all causal arguments. Often, the ceteris paribus conditions of an argument are obvious enough. But when they are not, the writer should spell things out.
**OTHER ARGUMENTS**

Description and causation are the dominant modalities of social science argumentation. However, they do not exhaust the universe of arguments. Since other styles of argument occasionally enter into social science (and always inform the work of social science in a tacit fashion) it is important that we survey this field, even if only very briefly. Our survey includes arguments that are **predictive**, **normative**, and **prescriptive**.

**Predictive arguments** (forecasts) attempt to tell us something about the future. That is, a factor, X, is thought to help predict an outcome, Y. This looks a lot like a causal argument; indeed, most causal arguments serve a predictive function. However, in saying that X predicts Y we do not presume that X causes Y. For example, it has been argued that a rise in infant mortality is a strong predictor of state failure.\(^4^2\) This does not imply that killing babies will bring down a state – a rather absurd, not to mention gruesome, idea. It seems more likely that some other factor is causing both X and Y. Infant mortality rates thus serve a useful predictive function without possessing causal attributes.

**Normative arguments** are about the moral goodness/badness of a phenomenon. For example, studies of foreign aid might be couched in positive or negative terms, implying that development assistance granted by developed countries to less developed countries is a good or bad thing. Frequently, the normative freight of an argument is fairly subtle, as revealed in an author’s choice of topic and his or her choice of words. But even when an author is careful not to tip her hat, it may be impossible to avoid certain normative preferences because they are embedded in our language. Imagine discussing topics such as human rights or fascism in a neutral fashion.

**Prescriptive arguments** enjoin us to take, or to refrain from taking, certain actions. They are often wedded to normative arguments. For example, a normative argument might be that foreign aid is benevolent. The implicit prescriptive argument is that we ought to increase development assistance.

We have now surveyed five types of arguments, as summarized in Table 2.4. The question naturally arises why we give preference in this book to description and causation. A first point to note is that normative and prescriptive arguments are not empirical in the usual sense of the term. One cannot prove or disprove a statement like *Foreign aid is good*, or *We ought to be more generous in our assistance to the developing world*, by appealing to facts. Of course, one could build a case for such arguments by appealing to facts. This, one might say, is the job of social science. But the case itself would be constructed with descriptive and causal arguments – and perhaps with predictive arguments as well. Thus, we might count up the amount of money devoted to foreign aid and to which countries, and which goals, it is allotted – a descriptive account. We might try to determine whether this assistance fostered stronger economic performance in the countries in which it was distributed – a causal account. And we might try to ascertain what effects additional aid will have in the future – a predictive account. Each of these arguments might be summarized in a few key hypotheses, and these hypotheses could be subjected to empirical proof, on the basis of which they could be proven right or wrong. (Of course, things are rarely so simple; but this is the general goal, at any rate, of an empirical approach to science.)

The result of this extensive empirical investigation might influence our views of the normative and prescriptive aspects of foreign aid. But the latter are not empirical matters in the same way as descriptive, causal, and predictive arguments. Because normative and prescriptive arguments are nonempirical they are generally downplayed in the work of social science.

\(^{4^2}\) Goldstone et al. (2010).
In this light, one might wonder about our choice to omit a detailed consideration of predictive arguments from the remainder of this book. Predictions of course can be empirically verified and are a bona fide part of the arsenal of social science. Our reason for excluding them from the book is more pragmatic than principled. At this point in time, models whose primary goal is predictive have not played a large role in most of the social science disciplines. (Prediction: this may change.)

In any case, it is useful to be able to parse arguments, i.e., to distinguish elements of an argument that are descriptive, causal, predictive, normative, prescriptive, or some combination thereof. This is important as a matter of clear expression and logical argumentation. For example, if normative arguments are disguised as description it will be difficult to figure out what an author is arguing – not to mention if s/he is right or wrong. Likewise, one must be careful when inferring prescriptive conclusions from a causal argument. The causal argument that foreign aid contributes to higher rates of child vaccination is consistent with a prescriptive argument that we should (nonetheless) curtail foreign aid. Only in parsing these arguments will you be able to tease these things apart. Thinking logically means, in the first place, distinguishing among different styles of argumentation.

Because social science is grounded in empirical arguments (descriptive, causal, predictive) it often assumes a rather austere tone. Consider the prospect of military intervention in order to support democracy abroad. An editorial or political speech on this topic is likely to foreground the normative and prescriptive elements of the subject. The writer or speaker who supports intervention is likely to dwell on the evils of autocracy in that country, on how many people have suffered and how they have suffered. S/he will advocate a course of action, and will prognosticate about what is likely to happen if we refuse to intervene. A social science study, by contrast, will likely highlight descriptive and causal aspects of the topic. What is the history and the current situation? What causes are at work? What is the general relationship between intervention and democratic transition? These are things that can be empirically verified, even if they are somewhat less exciting. The mark of a good social scientist is an ability to hold normative and prescriptive passions in check – harnessing them to the wheel of science.

### Table 2.4: Typology of Arguments

- **Descriptive.** An argument about patterns in the world.
- **Causal.** An argument about what generates change in an outcome.
- **Predictive.** An argument about what predicts future variation in an outcome.
- **Normative.** An argument about what is right or wrong, good or bad.
- **Prescriptive.** An argument about what actions one should take.

### GOOD ARGUMENTS

Having outlined different sorts of arguments we turn to the qualities of a good argument – descriptive and causal arguments, that is. What makes a theory (and its associated hypotheses) satisfactory?
Social science arguments strive to achieve a number of virtues, including *precision, generality, boundedness, parsimony, logical coherence, commensurability, innovation, and relevance*, summarized in Table 2.5. Whatever the topic might be, a social science theory is more useful if it is precise, general, wellbounded, parsimonious, coherent, commensurable, innovative, and relevant. Now, let us clarify these key objectives.

**Table 2.5: Theorizing: General Criteria**

<table>
<thead>
<tr>
<th><strong>Precision</strong> (specificity)</th>
<th>• Is it precise?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Generality</strong> (breadth, domain, population, range, scope)</td>
<td>• How broad is the scope? How many phenomena does a theory describe/explain?</td>
</tr>
<tr>
<td><strong>Boundedness</strong> (scope-conditions)</td>
<td>• How well-bounded is it?</td>
</tr>
<tr>
<td><strong>Parsimony</strong> (concision, economy, Occam’s razor, reduction, simplicity)</td>
<td>• How parsimonious is it? How many assumptions are required?</td>
</tr>
<tr>
<td><strong>Logical coherence</strong> (clarity, consistency; <em>antonym</em>: ambiguity)</td>
<td>• How logically coherent is it?</td>
</tr>
<tr>
<td><strong>Commensurability</strong> (consilience, harmony, logical economy, theoretical utility; <em>antonym</em>: adhocery)</td>
<td>• How well does it cumulate with other inferences? Does it advance logical economy in a field?</td>
</tr>
<tr>
<td><strong>Innovation</strong> (novelty)</td>
<td>• How new is it?</td>
</tr>
<tr>
<td><strong>Relevance</strong> (everyday importance, significance)</td>
<td>• How relevant is it to issues of concern to citizens and policymakers?</td>
</tr>
</tbody>
</table>

**Precision**

Theories strive for precision. The more precise a claim, the more useful it is, in the sense of providing more information about a putative phenomenon.

It also makes the argument more amenable to testing. If you are unconvinced by this, consider the opposite: a perfectly imprecise statement about the world, e.g., “Municipal governments in Africa are democratic, autocratic, or somewhere in between.” This sort of statement includes all possibilities since all polities must be classifiable somewhere on this spectrum. At the limit, a statement of no precision whatsoever says absolutely nothing about the world, and therefore is completely unfalsifiable.

**Generality**

If the fundamental purpose of social science is to tell us about the world then it stands to reason that a theory informing us about many phenomena is, by virtue of this fact, more useful than a theory pursuant to only a few phenomena. We will refer to this goal as *generality*, though it might also be called generalizability, breadth, or scope.
One wishes for a theory to encompass as many phenomena as possible. The more one can explain with a given argument the more powerful that argument is. Theories of great breadth tell us more about the world by explaining larger portions of that world. Thus, a theory of democracy that satisfactorily describes or explains regime-types across all nation-states is superior to one that applies to only a single region of the world or a single historical epoch. And a theory or theoretical framework describing or explaining different types of phenomena, such as evolution, is more useful than one pertaining to only a single outcome.

**Boundedness**

A theory should be well-bounded. That is, the boundaries of the theory – where it holds true (or is thought to hold true) and where it does not – should be clarified. These boundaries in time and space are sometimes referred to as the scope-conditions of a theory. Thus, if a theory about worker-training programs applies only to unemployed workers (and not to the under-employed), this is important to clarify. Theorizing on this subject is incomplete unless boundary conditions are established. Note that even “universal” theories have some boundary conditions, implicit in the concepts with which they have been articulated.

**Boundedness** is not established simply by carefully articulating boundaries. It is essential that those boundaries make sense – that they be true, or at least plausibly true (for they are not always easy to test). In framing an argument the researcher’s objective is to identify those phenomena that properly belong within the scope of a theory, excluding those that do not. Inferences should be appropriately bounded – neither too big nor too small. An arbitrary bounded inference, one that follows no apparent logic, is not convincing.

For example, if we stipulate that our theory of democratization applies to Sweden but not to Norway the reader would probably not be convinced. This scope-condition seems highly arbitrary, since these countries are very similar to each other. However, we may be able to justify this distinction with a more general scope-condition. For example, we might argue that our theory of democratization applies only to non-oil producing countries, i.e., to countries that are not subject to a “resource curse.” Since Norway is a large oil producer it might make sense to exclude it from the boundaries of the inference.

**Parsimony**

A fourth general goal of science is reduction, i.e., reducing the infinite plenitude of reality into a carefully framed argument from which unnecessary dross is removed. To the extent that a theory achieves this goal it is parsimonious. Like a lever, it lifts heavy weights with a moderate application of force. It is efficient, and its efficiency derives from its capacity to explain a lot with a minimal expenditure of energy.

The goal of parsimony, sometimes expressed as *Ockam’s razor*, is not assessable by the length of a study. Indeed, lengthy analyses may be required to provide evidence for a pithy argument. But it does call for a summary statement of key propositions.

Note that when a theory is parsimonious it is easier to test because there are fewer moving parts. More importantly, there are probably fewer assumptions necessary in order to sustain the argument. This later point deserves some discussion.
Theories build upon what we know already – or think we know – about the world. Nothing starts entirely from scratch. A good theory requires fewer departures from common sense, fewer leaps of faith, fewer a priori assumptions. A poorly constructed theory, by contrast, asks the reader to accept a great deal about the world upon the authority of the author. This sort of inference does not build on solid foundations. It is stipulative. The point, then, is not to do away with assumptions but to limit them – or at least the most questionable of them – as much as possible. This is perhaps the most important implication of parsimony as a criterion of good theorizing. The best sort of theory has only one empirical question at issue – the main hypothesis – all else is already firmly established and likely to be accepted by the reader.

**Logical Coherence**

In order to be meaningful an argument must demonstrate some degree of *logical coherence*. If there are many moving parts, as in a broad and abstract theory, they ought to hold together. A complex theory should revolve around a single core and the precepts should be logically linked. One facet should imply the others. Indeed, if the parts of an argument are inconsistent, the argument itself is virtually meaningless, and surely untestable.

Examples of logical inconsistency are not hard to find. While Marxist theory is premised on the idea that social classes advance their self-interest, the final stage of socialism seems to take as its premise that people are no longer motivated by self-interest – or have a very different idea of what that is.

**Commensurability**

Theories assume meaning within a field of pre-existing concepts and theories, typically a field or subfield of study. If a theory fits comfortably within a body of work we shall say that it possesses *commensurability*. This might be achieved by subsuming other theories within a new theoretical framework, or by offering a clarification or refinement of an existing theory. In any case, the theory advances logical economy in the field, thereby assisting the cumulation of knowledge about a subject.

If it does not – if it sits by itself in a corner and does not relate productively to other theories – then it is likely to be dismissed as “ad hoc,” or “idiosyncratic.” It does not fit with present understandings of the world. It has little theoretical or conceptual utility.

Of course, deviant theories and neologisms (novel concepts) may be extremely useful in the long run. Indeed, the first sign of breakdown in a broad theory or paradigm is the existence of findings that cannot easily be made sense of. Yet, until such time as a new theory or paradigm can be constructed (one that would gather the new findings together with the old in a single overarching framework) the wayward proposition is ad hoc, idiosyncratic, and apt to be ignored.

**Innovation**

A novel theory presents a new way of thinking about some arena of activity. This, by itself, may constitute a significant contribution to our knowledge of the world, even if it cannot easily be demonstrated or proven (though it must at some point be falsifiable). It might mean the discovery of a new topic, never previously described or reconceptualized. (Recall that all topics familiar today were once brand spanking new.) It might mean a new explanation for an outcome of general concern. Or it might mean the reformulation of an existing theory. There are many ways to achieve theoretical *innovation*. 
A recent example of theoretical innovation is the theory of social capital – the idea that social networks have far-reaching consequences for governance and for economic performance (see Chapter 1). This idea is a relative newcomer to the theoretical toolkit of social science, having been developed over the last several decades. Note, however, that it was not launched all at once. Social capital theory was the work of many scholars across the fields of economics, political science, and sociology. And it continues to evolve. Arguably, any influential work on a subject alters the general theory on that subject in some fashion and can therefore claim the mantle of innovation.

**RELEVANCE**

A theory, finally, should be relevant to things that matter to non-social scientists (“real people”). Relevance does not mean that lay citizens need to understand its intricacies (for some theories can be quite technical). But they should care about what it describes or explains. It is worth reminding ourselves that social science is not a disembodied body of knowledge – the way pure mathematics is sometimes described. It emanates from the needs and concerns of human beings. It is for, and about, society.

Indeed, most social scientists aim to affect public opinion and/or the course of public policy with their work, in some fashion. And it seems fair to judge the theories that possess a strong claim to relevance as superior (ceteris paribus) to those that do not. And it seems fair to ask writers to justify the reader’s potential expenditure of time, effort, and money with some sort of pay-off. Readers are not likely to be carried very far on the strength of a writer’s method or prose if they do not feel that there is something important at stake in the investigation. They must care about the outcome.

Arguably, truth-claims are enhanced when a writer frankly proclaims her preferences at the outset of the work. This way, possible inaccuracies in evidence or presentation are easier to detect, and to evaluate. Hidden prejudices probably do more harm than those that are openly avowed. Yet, it must be stressed that the value of a work of social science derives from its value-added as descriptive or causal propositions about the world, not its normative point of view or prescriptions for action. To say, “Y is good” or “We should do Y” is to say very little.

Social science is most powerful when the normative and prescriptive angles of a work are handled delicately. The most compelling arguments for increased spending on social welfare programs, for example, are those that demonstrate causal relationships, e.g., that particular programs aid in alleviating conditions of poverty and do not have negative externalities. Such studies do not proclaim baldly “Poverty is bad,” or “We should increase social welfare spending,” although there is no question that these views undergird a good deal of research on poverty and social policy. So long as the author’s research is sound, one need not concern oneself with her normative position on the matter.

Finally, it seems appropriate to observe that the vast majority of social science analysis has little to do with what is good or bad. No one – or virtually no one – argues against the virtues of peace, prosperity, democracy, and self-fulfillment. What is relevant (in the larger sense of the word), is any knowledge that might help us to achieve these desiderata. Here is where social science matters, or ought to matter.

**CONCLUSIONS**

This chapter began by distinguishing various types of social science argumentation. The most important of these are descriptive and causal, each of which has many subtypes. These dominant
forms of argument were then contrasted with predictive, normative, and prescriptive arguments. In the final section, we set forth the characteristics of a good descriptive or causal argument, namely precision, generality, boundedness, parsimony, logical coherence, commensurability, innovation, and relevance (see Table 2.5).

A number of these goals relate to a theory’s testability, or falsifiability. Karl Popper, who invented the term, pointed out that in order to be true, a theory must also have the potential to be false. Unless a theory admits this possibility it has no empirical content; it is neither true nor false. Of course, some very general theories do not translate easily into testable hypotheses. Nonetheless, the process of empirical testing must occur at some point, and in this sense every theory embraces Popper’s project.43

It should be acknowledged that the eight criteria in Table 2.5 are by no means easy to satisfy. Indeed, there are many situations in which one or more of these demands are not met, or are only partially met. There are also plenty of situations in which attaining one objective entails sacrificing another. The more precise a theory, the less general it may be. The more commensurable an argument the less innovative it may be. And so forth. Thus, when we say that it is important to strive for a goal we do not mean to imply that other goals should be neglected. We mean, rather, that the researcher should attempt to balance these divergent goals in a satisfactory manner, achieving the best possible compromise among various criteria – or, at the least, acknowledging the sacrifices made on one or more dimensions in order to achieve success along other dimensions.

In the following chapter we proceed to the components of an argument, namely its key concepts and the indicators chosen to measure those concepts.

43 We do not mean to imply that all wrong theories can be decisively falsified; this goal of Popper’s seems unrealistic.
Key Terms

- Theory
- Model
- Hypothesis
- Descriptive argument (association, synthesis, typology)
- Trend analysis
- Network analysis
- Periodization
- Matrix typology
- Taxonomy
- Causal argument
- Predictive argument
- Normative argument
- Prescriptive argument
- Precision
- Generality
- Boundedness
- Parsimony
- Logical coherence
- Commensurability
- Innovation
- Relevance
- Falsifiability

Inquiries

1. Consider a topic of your choice and use it to illustrate different forms of descriptive argumentation. That is, show what an association, a synthesis, and a typology of that topic might look like. (This exercise need not be fully developed; we are using it here for heuristic purposes, i.e., to practice various modes of argumentation.)

2. Consider a wellknown synthesis in your field. What makes it powerful, persuasive? What is its appeal to scholars in that field? What are its weaknesses? Are there disagreements over how to define and apply the concept? Could it be improved upon? Is a better synthesis possible? Or is the attempt at synthesis doomed? (If so, is there a better approach?)

3. Consider a wellknown typology in your field. Why has it proven attractive to scholars? What are its uses and its strengths? What are its limitations and weaknesses? How successfully does it divide up the phenomena (i.e., how exclusive and exhaustive are its categories)? Does it force hard-and-fast distinctions on a fluid subject? Could it be improved upon? Or is typological thinking inappropriate for the phenomena?
4. There are no “universal” laws of social science; all theories have some boundary conditions (scope-conditions). Explain.

5. Take a theory that you are familiar with and discuss its scope-conditions. What do you think is the most logical way to bound this theory? What would be an inappropriate (arbitrary) boundary?

6. Theories should be **commensurable**. What does this mean?

7. What makes an argument good (from the perspective of social science)? Compare and contrast an article drawn from a newspaper with an article (perhaps on the same general subject) drawn from a social science journal. What are the similarities and differences? Do the criteria of a good social science argument – precision, generality, boundedness, parsimony, logical coherence, commensurability, innovation, and relevance – also apply to works of journalism? (Should they?)

8. The story is told of a hapless physics student who was defending his thesis before a committee of professors. One of the committee members was unimpressed, declaring: “This is the worst thesis I have ever seen. It is not even wrong!” What is the point of this harsh critique? What does the professor mean to say? (How can something be worse than wrong?)

9. Take an article drawn from a newspaper or academic journal and try to classify each argument in the 5-part typology represented in Table 2.4. If an argument doesn’t fit, or doesn’t seem to fit, explain why, and perhaps propose a new category for the typology.

**Resources**

**DESCRIPTION, GENERALLY CONSIDERED**


**TYPOLOGIES, CLASSIFICATION**

CAUSAL ARGUMENTS

- See sources listed at the end of Chapters 5 and 6.
3. **Concepts and Measures**

Arguments are articulated with the use of key concepts. Indeed, the argument of a study is inseparable from its key concepts since the latter are the linguistic tools with which an argument is formulated. Any study of democracy, for example, must wrestle with the problem of how to define this key term – which will guide our discussion in this chapter.

Concepts, in turn, receive empirical meaning through the indicators chosen to measure them. Any study of democracy must be concerned not only with how to define democracy but also with how to **operationalize** (measure) this abstract concept.

Conceptualization and measurement are thus closely linked. This is why we have chosen to present them together in this chapter, which begins with concept formation and proceeds to measurement.

**Concepts**

The key concepts of social science are never fixed and, regrettably, not always clear. Many abstract concepts – such as democracy or social capital – are employed in a variety of ways and thus mean different things in different contexts. This is true even of more specific concepts such as worker-training programs. (Does a one-day program focusing on advice for job-hunting qualify? How about a person who enlists government support to take classes at a community college, or an apprenticeship program?)

The persistent ambiguity of key concepts makes it difficult for the reader, who may struggle to figure out what a term means in a given context and how it connects with other work (using the same or similar terms). It also makes it difficult for writers, who must identify which of several terms they should adopt in their own work and how they should define the chosen term.

Sometimes, the task of forming concepts seems highly arbitrary. And this, in turn, may prompt readers to adopt a skeptical attitude toward the subject. At the same time, the choice of concepts is never entirely arbitrary. Some choices are usually better than others, and a few are patently absurd.

In this spirit, we offer the following criteria, intended to guide the process of concept formation in the social sciences. A good concept, we shall argue, is **resonant**, **internally coherent**, **externally differentiated**, **theoretically useful**, and **consistent** in meaning, as summarized in Table 3.1.
Table 3.1: Criteria of Concept Formation

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resonance</td>
<td>(familiarity, normal usage; antonym: idiosyncrasy, neologism, stipulation)</td>
</tr>
<tr>
<td></td>
<td>• How faithful is the concept to extant definitions and established usage?</td>
</tr>
<tr>
<td>Internal coherence</td>
<td>(depth, essence, fecundity, fruitfulness, natural kinds, power, real, thickness)</td>
</tr>
<tr>
<td></td>
<td>• How many attributes do referents of a concept share?</td>
</tr>
<tr>
<td>External differentiation</td>
<td>(context, contrast-space, perspective, reference-point, semantic field)</td>
</tr>
<tr>
<td></td>
<td>• How differentiated is a concept from neighboring concepts? What is the contrast-space against which a concept defines itself?</td>
</tr>
<tr>
<td>Theoretical utility</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• What utility does a concept have within a larger theory and research design?</td>
</tr>
<tr>
<td>Consistency</td>
<td>(antonym: slippage)</td>
</tr>
<tr>
<td></td>
<td>• Is the meaning of a concept consistent throughout a work?</td>
</tr>
</tbody>
</table>

**Resonance**

Consider the following definition of “democracy”: *a furry animal with four legs*. This is nonsense, of course. And it is nonsense precisely because it breaks radically with common usage. This is not what anyone thinks of when they hear the word democracy and it doesn’t share any elements of prevailing definitions of the term.

Forming a concept involves choosing a term, and a definition for that term, that *resonates* with standard usage in ordinary speech or in a specialized language region (i.e., a field of inquiry). Idiosyncratic concepts (such as the foregoing example) should be avoided, wherever possible.

Of course, sometimes there is no concept that defines a topic in precisely the way that you wish to define it. In this case, some departure from ordinary usage is required. But this neologism should deviate from ordinary usage as little as possible. Otherwise, a term loses touch with its semantic roots.

**Internal coherence**

Concepts strive to identify those things that are alike, grouping them together, and contrasting them with things that are different. Apples with apples, and oranges with oranges, as the phrase goes. A concept that accomplishes this possesses *internally coherence*. It may also be described as fecund, fruitful, illuminating, informative, or insightful.

Consider three conceptualizations of regime-type. One differentiates between democracies and autocracies;44 another distinguishes pure democracies, competitive authoritarian states, and pure autocracies;45 and a third establishes a twenty-one point index that is intended to function as an interval scale.46 Which of these is most satisfactory? Evidently, each may be satisfactory for different purposes. However, for descriptive purposes the utility of a schema hinges largely upon its fecundity. In the present instance, this means: which schema best describes the subject matter? More specifically, which schema most successfully bundles regime characteristics together, differentiating them from other bundles? Is the natural break-point among regimes to be found between autocracies and democracies (a two-part classification), among pure democracies, competitive

---

44 Alvarez et al. (1996).
45 Levitsky and Way (2002).
autocracies, and pure autocracies, or is there instead a continuum of characteristics with no clear “bundles,” justifying a continuous dimensional space?

Naturally, many other options might also be considered. Some might argue that regime-types are multidimensional, and therefore inappropriate for a unidimensional scale (see discussion below). But all such arguments appeal to the ideal of internal coherence.

**EXTERNAL DIFFERENTIATION**

A concept cannot be internally coherent unless it is distinguishable from other concepts. If apples are indistinguishable from oranges, the coherence of “apple” is called into question. The twin desiderata of coherence and differentiation correspond to “lumping and splitting” operations in social classification (Zerubavel 1996) and to “similarity and difference” judgments in cognitive linguistics (Tversky and Gati 1978).

How, then, does a concept establish clearly demarcated borders? A key element is to specify carefully how a concept fits within a larger semantic field composed of neighboring concepts and referents. We shall refer to this as the background context or contrast-space of a concept.

Concepts are defined in terms of other concepts – boys in terms of girls, nation-states in terms of empires, parties in terms of interest groups. These neighboring terms (synonyms, near-synonyms, antonyms, and superordinate/subordinate concepts) give meaning to a concept. Precisely because of the interconnectedness of language, the redefinition of a term necessarily involves some resettling of its semantic field. It is impossible to redefine one term without also, at least by implication, redefining others. Any redefinition of democracy changes our understanding of authoritarianism.

It follows that a new concept should unsettle the semantic field as little as possible, leaving other concepts as they were (more or less). Indeed, a new term or re-definition that poaches attributes from neighboring concepts is laying the ground for future conceptual anarchy. It may resonate on first reading, but is likely to foster confusion in that field or subfield over the long-term. “Crowded” semantic fields are an example of this. Consider the many terms that have been developed over the past several decades to refer to citizen-based groups—civic association, voluntary association, civil society organization, citizen sector organization, nongovernmental organization (NGO), interest group, and grassroots organization. While subtle differences may be established among these terms, one must also recognize that the endless propagation of terms is not productive for this field of study. Often, neologisms are a sign of conceptual disarray rather than of theoretical fecundity.

In any case, it is incumbent upon writers to clarify how their chosen concept(s) differ from neighboring concepts sharing the same semantic and phenomenal space. This requires establishing clear contrasts with what lies outside the boundaries of a concept. It should be clear how democracies differ from dictatorships, and how social capital differs from ordinary networks.

---

48 The twin desiderata of coherence and differentiation correspond to “lumping and splitting” operations in social classification (Zerubavel 1996) and to “similarity and difference” judgments in cognitive linguistics (Tversky and Gati 1978).
THEORETICAL UTILITY
We have said that concepts play a key role within a theory (or theories). So it stands to reason that we should think about their function within that theory – its theoretical utility – when choosing terms and definitions.

One important function of a concept is to help define the limits of a theory, its scope-conditions. If one is arguing that a theory of electoral politics operates only within democracies, one must define what a democracy is – and this definition will be partly driven by what one perceives as the key elements of the theory. Perhaps the theory of electoral politics requires multi-party competition but not equal access to media and campaign finance. The definition of “democracy” should reflect that.

Within a causal theory, key concepts define the causal factor(s) and the outcome. Typically, concepts designed for use as dependent variables group together many attributes. Here, a diffuse definition may be fruitful. By contrast, concepts designed for use as independent variables are generally smaller, more parsimonious. Additionally, concept formation in the context of causal models must be careful to employ concepts that differentiate a cause from its effect, so that circularity in the argument is avoided. (These points will become clearer after a reading of Part II, focused on causality.)

CONSISTENCY
Having formed a concept, one should employ it with consistency throughout a study. And it should mean the same thing in all contexts that are relevant to a study. If not, the concept will generate unwanted ambiguity.

Inconsistency – where a term means something different in different contexts – creates a problem of conceptual stretching. Suppose, for example, that democracy is initially defined as the existence of multi-party elections. Later on, the author uses the term in a way that suggests that civil liberties are essential to the achievement of democracy. Clearly, the term is being used inconsistently, and this cannot help but generate confusion.

One way to avoid this sort of inconsistency is to distinguish among several key terms that share common attributes by the addition of a qualifier. In this example, one might want to distinguish electoral democracy from liberal democracy.

Consistency does not imply that all users of a concept should understand it in the same way. With multivalent concepts such as democracy this is clearly impossible. Regardless of how “democracy” is defined by an author, some people will continue to use the term in different ways. There is no harm in this (so long as the chosen term resonates with at least one common meaning of the term). Indeed, the progress of social science depends upon the development of specialized terms, which coexist with a variety of everyday or specialized meanings. The criterion of consistency imposes restrictions only on the author of a study (who must use terms in a consistent fashion), not on other authors or speakers.

STRATEGIES OF CONCEPT FORMATION

Having surveyed general criteria pertaining to concepts, we turn to strategies that may help to achieve these goals.

Concept formation generally begins with a survey of potential concepts. As one is formulating a theory, it is important to gather together the various terms that might be used to describe this phenomenon and commonly used definitions of these terms. Thus, if you are contemplating social capital you might also want to think about related terms such as civic engagement, civic associations, civil society, social networks, participation, and voluntary associations. Which of these terms expresses most accurately and precisely what you are trying to say? Which encompasses your subject, but does not extend beyond it? A well-chosen concept helps to establish the boundaries around an argument, as discussed above.

Once you have settled on a key term, begin to consider various ways of defining that term. This may be done by consulting scholarly books, dictionaries, and usage patterns. Think to yourself, what do people mean when they say “X”? If there are a number of plausible definitions it may be worthwhile to break these down into their component parts. This approach is usually more useful than a listing of definitions. For example, there are an infinite number of definitions of democracy. However, these definitions tend to repeat standard elements, which can be organized into a single table, as shown in Table 3.2.51

Table 3.2: Democracy: Fundamental Attributes

**Electoral** (elite, minimal, realist, Schumpeterian)
- **Ideals:** contestation, competition.
- **Question:** Are important government offices filled by free and fair multiparty elections?
- **Institutions:** elections, political parties, competitiveness, turnover.

**Liberal**
- **Ideals:** limited government, horizontal accountability, individual rights, civil liberties, transparency.
- **Question:** Is power constrained and individual rights guaranteed?
- **Institutions:** independent media, interest groups, judiciary; written constitution with explicit guarantees.

**Majoritarian** (responsible party government)
- **Ideals:** majority rule, centralization, vertical accountability.
- **Question:** Does the majority (or plurality) rule?
- **Institutions:** consolidated and centralized, with special focus on the role of political parties.

**Consensual**
- **Ideal:** power sharing, multiple veto-points.
- **Question:** How numerous, independent, and diverse are the groups and institutions that participate in policymaking?
- **Institutions:** multiparty system, proportional electoral laws, supermajorities, oversized cabinets, federalism.

**Participatory**
- **Ideal:** government by the people.
- **Question:** Do ordinary citizens participate in politics?
- **Institutions:** election law, civil society, local government, direct democracy.

**Deliberative**
- **Ideal:** government by reason.
- **Question:** Are political decisions the product of public deliberation?
- **Institutions:** media, hearings, panels, other deliberative bodies.

**Egalitarian** (social)
- **Ideal:** political equality.
- **Question:** Are all citizens equally empowered?
- **Institutions:** socio-economic and political factors that generate conditions for political equality.
Table 3.2 provides a fairly encompassing set of traits associated with democracy. Yet, because of the number and diversity of these attributes, this table of attributes does not take us very far towards a final definition. In order to create a more tractable—and internally coherent—empirical concept, one must go further. This next step—toward a specialized definition—is crucial. To achieve it, two approaches will be reviewed: minimal and maximal.

**Minimal**

One strategy seeks to identify the bare essentials of a concept, sufficient to differentiate it from neighboring concepts without excluding any of the phenomena generally understood as part of the concept. The resulting definition should be capable of substituting for all (non-idiomatic) uses of the term without too much loss of meaning. This means of course that it should not conflict with any (non-idiomatic) usages.

You might try this out by substituting various definitions of democracy for the term “democracy” in a sentence. That is, think of a sentence including the word “democracy.” Now substitute the term for one of its definitions, as summarized in Table 3.2 or elsewhere. Does it make sense? Or does it seem strained, partial?

Each attribute that defines a concept minimally is regarded as a necessary condition; all entities must possess this attribute to be considered a member of the set. Collectively, these attributes are jointly sufficient to bound the concept. Minimal definitions thus aim for crisp borders, allowing for the classification of entities as “in” or “out.”

Sometimes, minimal concepts are crafted around an abstract core principle such as “rule by the people.” In this instance, the core meaning satisfies the criterion of resonance, for all invocations of democracy revolve in some way around this idea. However, such an abstract definition does not achieve crisp borders for the concept; indeed, it scarcely identifies borders. In this respect, it is problematic. (The rulers of North Korea would probably claim that their system of government enshrines rule by the people.)

A more common approach is to identify a specific component of the term that nearly everyone agrees upon. If we are limiting ourselves to representative democracies (excluding direct democracies), one might argue that free and fair elections constitutes a necessary condition of democracy. This attribute suffices as a minimal definition for it is sufficient to bound the entity empirically. That is, having free and fair elections makes a polity a democracy; no other attributes are necessary. And this definition is sufficient to distinguish democracy from neighboring concepts such as dictatorship.

The caveat, of course, is that we are defining democracy in a very minimal fashion, leaving other attributes often associated with the concept in abeyance. This imposes some costs in resonance. The stripped down meaning of democracy sounds hollow to those attuned to the concept’s many nuances. It ignores a lot.
MAXIMAL

Maximal definitions aim for the inclusion of all (non-idiosyncratic) attributes, thereby defining a concept in its purest, most “ideal” form. This would of course include the attribute(s) that defines the concept minimally – its necessary condition(s). As Weber describes it, “an ideal-type is formed...by the synthesis of a great many diffuse, discrete, more or less present and occasionally absent concrete individual/phenomena, which are arranged according to those one-sidedly emphasized viewpoints into a unified analytical construct.”[52]

Following this recipe, one might create an ideal-type definition of democracy that includes most or all of the dimensions listed in Table 3.2. Of course, some might be excluded if it could be argued that they detract significantly from the coherence of the overall concept. Blatantly contradictory elements – such as the majoritarian and consensus components of democracy in Table 3.2 – should be avoided.

Ideal-types, as the term suggests, need not have a specific real-life empirical referent. Perhaps no extant polity achieves democracy in its purest form. To be of service, however, an ideal-type must approximate real, existing entities so that the latter can be scored according to how closely they resemble the ideal. Ideal-types are always matters of degree.

CODA

Not all definitions follow a minimal or maximal logic. Others are shaped primarily by theoretical utility, i.e., by the functions the concept is expected to perform within a larger argument and an associated research design. All definitions should, in any case, respect the general criteria of resonance, internal coherence, external differentiation, theoretical utility, and consistency, as summarized in Table 3.1.

Remember that the over-arching goal of a concept is to help an argument along, to clarify what the author is trying to articulate. A key concept should not overshadow an argument. Although it is important to be selfconscious about the terms and definitions you choose, it is also important not to spend too much time defining terms; otherwise, the conceptual discussion will detract from the main point. So, define your terms carefully, making as few departures from everyday usage as possible. And then move on.

MEASURES

Having defined a concept, the critical issue is how to recognize it when we see it. Can democracy be distinguished from autocracy? Can a situation of high social capital be distinguished from a situation of low social capital? What do these concepts mean empirically?

Operationalizing a concept involves choosing an indicator to represent that concept empirically. This may also be referred to as a dimension, measure, or variable. Whatever the terminology, an indicator is the primitive empirical proposition underlying all other generalizing arguments about the world. An indicator of democracy might be the existence of multi-party competition. An indicator of human development might be the infant mortality rate (the rate of deaths between the ages of 0 and 1 as a share of 1,000 live births). Alternatively, a number of indicators may be combined into a single index in order to better represent a concept.

[52] Weber (1905/1949: 90). See also Burger (1976). In citing Weber, we do not claim to be using the concept of an ideal-type in precisely the way that Weber envisioned.
The problem of measurement stems from the fact that most (and perhaps all) important social science concepts are not directly observable. They are latent. All abstract concepts fall into this category. We cannot “see” democracy or social capital. Many terms in the social science lexicon suffer this shortcoming. Alienation, anomie, charisma, civil society, collective conscience, crisis, culture, democracy, dogmatism, equality, false consciousness, hegemony, ideology, legitimacy, mass society, national character, petty bourgeoisie, rationalization, sovereignty, state, and status anxiety are all “fuzzy” concepts. We may be able to define them in a general way, but we have immense difficulty locating their referents in empirical space. These are the sorts of measurement problems that social science is at pains to resolve.

In the following sections we review various strategies of measurement, along with the challenges that each strategy entails. These strategies involve: (a) levels of abstraction, (b) scales, (c) aggregation techniques, and (d) basic objectives, as summarized in Table 3.3.

**Table 3.3: Measurement Strategies**

<table>
<thead>
<tr>
<th>Levels of abstraction</th>
<th>Scales</th>
<th>Aggregation</th>
<th>Objectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>• High</td>
<td>• Categorical (Nominal, Ordinal)</td>
<td>• Boolean</td>
<td>• Discrimination</td>
</tr>
<tr>
<td>• Medium</td>
<td>• Numeric (Interval, Ratio)</td>
<td>• Additive</td>
<td>• Grouping</td>
</tr>
<tr>
<td>• Low</td>
<td></td>
<td>• Multiplicative</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Factor-analytic</td>
<td></td>
</tr>
</tbody>
</table>

** Levels of Abstraction **

All empirical concepts of interest to social science encompass multiple levels of abstraction. At the very least, one can distinguish between the attributes that define a concept and the indicators that operationalize it, generating two tiers: (a) conceptualization and (b) measurement. This is probably sufficient for a small-order concept like political parties or elections.

For more abstract concepts like democracy multiple tiers may be required to adequately represent all the levels of analysis implicit in the concept, and to fully operationalize it – to bring it down to earth, so to speak. Consider the following hierarchy:

1. Democracy (the latent concept of theoretical interest)
2. Electoral (a conception of democracy)
3. Free and fair elections (a key component of electoral democracy)
4. Validation of an election by international election observers (an indicator of free/fair elections)

Here, four tiers of a concept are illustrated. Naturally, one might add levels at the bottom of the hierarchy (e.g., a more specific and operational definition of international election observers determine the freeness and fairness of elections and how disagreements among them should be resolved), at the top (e.g., a superordinate category that subsumes democracy such as governance), or somewhere in the middle. The tiers on a ladder of abstraction are in some sense arbitrary and can be expanded or contracted like an accordion.

Sometimes, problems of measurement can often be resolved, or at least mitigated, by moving down the ladder of abstraction, e.g., from #1 (Democracy) to #2 (Electoral) to #3 (Free and fair elections) to #4 (Validation of an election by international election observers). Small, concrete things are often easier to measure than large, abstract things.

Naturally, at a certain point micro-level phenomena become less observable, and more difficult to measure. This is the situation faced in fields like biology and physics, where molecules and subatomic particles are at the frontiers of measurement. In the social sciences, however, the individual (i.e., the whole human being) is usually regarded as the most disaggregated unit of analysis. In these fields, problems of measurement are generally the product of abstraction, not of specificity.

At the same time, as one scopes down from “democracy” to low-level indicators one may find that the connection between the concept and the phenomena becomes highly attenuated. A chosen indicator may be highly precise but of questionable validity with respect to a high-order concept of theoretical interest. For example, if one attempts to measure the existence of democracy solely by looking at the decisions of international election observers one may feel that important elements of the concept are being ignored. This is the tradeoff encountered when moving along a ladder of abstraction: precision is usually enhanced as one moves down while conceptual validity is enhanced as one moves up. (These terms are formally defined in Chapter 4.)

**SCALES**

Measurement is an inherently comparative venture. It presumes a scale — i.e., a standard metric by which heterogeneous things can be systematically and precisely compared. Hence, all attempts at measurement face a problem of equivalence or consistency across contexts. A chosen measure must mean the same thing, and must adequately represent the concept of theoretical interest, across all contexts to which it is being applied. The challenge is therefore to find a way to compare things across diverse contexts without too much loss of meaning or distortion.

Some scales are categorical (qualitative), by virtue of the fact that the distance between categories is undefined. Other scales are numeric (quantitative) by virtue of the fact that the distance between categories is defined and measured along a numeric scale. Within this two-part classification other sub-types fall, as indicated in Table 3.4.

Among categorical scales, those that are nominal define members of the same class (they are examples of something) but are un-ranked. For example, apples, oranges, and grapes are all correctly classified as fruit; but they are not more or less of anything relative to each other.

**Ordinal** scales are members of the same class and also ranked: very sweet is sweeter than sweet. But one does not know the true distance separating each level in the scale. It is unclear, for example, how much sweeter “very sweet” is relative to “sweet.”

Among numeric scales, those that are interval are characterized by a consistent measure of distance between categories. For example, the distance between 3 and 4 on a temperature scale
(Celsius or Fahrenheit) is the same as the distance between 25 and 26, and is defined by a formal rule, consistently applied across the scale.

**Ratio** scales are interval scales with a true zero, indicating the absence of whatever quantity is being measured (a null set). In the case of money, 0 signals no money. In the case of temperature on the Kelvin scale, 0 indicates the absence of all thermal energy.

Frequently, interval and ratio scales fulfill the requirements of a numeric scale only within certain bounds. For example, life-span is bounded on the lower end at zero (arguably, it is also bounded at the upper end, though this boundary is more difficult to define).

The table below is a typology of scales:

<table>
<thead>
<tr>
<th>Categorical</th>
<th>Numeric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>x</td>
</tr>
<tr>
<td>Ordinal</td>
<td>x</td>
</tr>
<tr>
<td>Interval</td>
<td>x</td>
</tr>
<tr>
<td>Ratio</td>
<td>x</td>
</tr>
</tbody>
</table>

Because scales are defined for specific purposes the same phenomena may be differently classified. For some purposes, it may be sensible to consider varieties of fruit as nominal categories. For other purposes, it may be sensible to consider them as part of an ordinal scale (more or less acidic) or a ratio scale (using a ratio measure of acidity).

For many topics, it is correct to regard higher-level scales as more informative. Thus, we would ordinarily interpret an ordinal scale for temperature ("hot," "medium," "cold") as less precise (and therefore less informative) than an interval scale (e.g., Celsius) or ratio scale (e.g., Kelvin). However, this is true only with reference to that particular phenomenon. It would not be true for sex, for example, since this dimension admits of only two categories – male and female (as usually understood). Here, an interval scale reduces to a nominal scale.

Note also that while more precise indicators promise more, they also demand more. Specifically, they require a greater number of assumptions about the nature of the underlying data. If any of these assumptions are false, or only partially true, any inference building upon that indicator will be cast in doubt.

Let us explore some examples.

For many purposes, it is essential to distinguish polities in a dichotomous fashion, as democratic or authoritarian (autocratic).\(^5^3\) This produces a nominal scale with two categories, also known as a binary scale or dummy variable. These categories might be Yes/No, Present/Absent, High/Low, Male/Female, or, in our example, Democratic/Autocratic.

For other purposes, one may require a more finely graded indicator of democracy. Gerardo Munck defines a four-level ordinal index of electoral contestation including categories for authoritarianism, semi-authoritarianism, semi-democracy, and democracy.\(^5^4\) Here, each category is

---


\(^5^4\) Munck (2009: 45).
distinguishable and clearly ranked relative to the concept of theoretical interest. Defining attributes for each category are elaborated in Table 3.5. If one is willing to accept the additional assumption that categories in the four-point scale are equidistant, one may regard this index as an interval scale. And insofar as the first category comprises a true zero – no contestation whatsoever – the index may also be understood as a ratio scale.

This example nicely illustrates the fact that the same set of categories may be different interpreted, according to different assumptions about the underlying empirical phenomena and different uses for which the same indicator may be enlisted.

Choices among scales are often driven by their relative tractability. For some purposes, binary scales are more useful than interval scales, and vice-versa. The point to keep in mind is that these decisions, while analytically convenient, often involve either a loss of information and/or the introduction of bias in the variable of interest. Although there are no “natural” scales, some interpretations of reality are more plausible than others.

| Table 3.5: A Single Index with Multiple Interpretations |
| Key concept: Electoral contestation |
0. Authoritarianism: No elections or elections with only one party or candidate.
1. Semi-authoritarianism: Elections in which more than one party or candidate runs but not all parties and candidates face the possibility of losing.
2. Semi-democracy: Elections in which more than one party or candidate run and all parties and candidates face the possibility of losing but not all parties or candidates are allowed to participate.
3. Democracy: Elections in which only anti-system extremist groups are banned and all parties and candidates face the possibility of losing.

AGGREGATION
While concrete concepts such as infant mortality may be measurable with a single indicator, more complex concepts such as democracy usually require the inclusion of several elements into a single index. In these settings, researchers must grapple with the problem of aggregation.

Perhaps the simplest way to construct an index is with the deterministic conditions prescribed by Boolean (set theory) logic. In this setting, all indicators must be coded in a binary (0/1) fashion. For purposes of illustration, let us return to the attribute of democracy known as “free and fair elections.” This attribute may be understood in one of three ways.

(1) As a necessary condition of democracy, a polity must embody this attribute, though there may be other membership conditions as well. A polity must have free and fair elections in order to qualify as democratic, but having free and fair elections is not sufficient to qualify as democratic.

(2) As a sufficient condition, this feature is sufficient by itself to qualify a polity as democratic, though there are understood to be other conditions that would also qualify a polity as democratic. For example, one might take the position that democracy can be achieved either through free and fair elections or through deliberative decisionmaking. In one version of democracy people rule through chosen representatives and in another version of democracy the people rule directly (on the

55 Drawn from Munck (2009: 45).
model of the Athenian polis). In this situation, two attributes are substitutable for the other. The possession of either attribute is sufficient to define a concept.

(3) As a necessary-and-sufficient condition of democracy, this is the only characteristic that matters. A polity with free and fair elections is a democracy; one without is an autocracy. That is all there is to it.

Boolean conditions aggregate in an explicit and clear-cut fashion. For example, if democracy is operationalized with three necessary conditions, all must be satisfied in order to code a case as democratic. If democracy is operationalized with three sufficient conditions, one or more conditions must be satisfied in order to code a case as democratic. If democracy is operationalized with a single necessary-and-sufficient condition, this condition must be satisfied in order to code a case as democratic; no other conditions are relevant.

Set-theoretic conditions may be combined in various ways to form a single aggregation formula (so long as they don’t contradict each other). In principle, there is no limit to the number of necessary or sufficient conditions that can be accommodated. Set-theoretic conditions are thus easy to aggregate.

Unfortunately, it is often difficult to reduce the complexities of social life to traits that can be coded in a binary fashion without losing important information or making arbitrary judgments. Likewise, attributes do not always combine in a crisp, Boolean fashion to capture the concept of interest. Thus, we are led to consider a variety of alternatives.

An additive index combines indicators by addition, with the assumption that each component is equal (in importance) and independent (its contribution to the concept of interest is independent of the values of other indicators). Suppose four attributes – $X_1, X_2, X_3, X_4$ – are considered relevant to measuring democracy. An additive index could be constructed in two steps. First, one would need to adjust the scales of these variables so that they align with one another. Note that if one indicator is scored from 0 to 1 and another from 0 to 100 the latter will carry 100 times the weight of the former in an additive index. Interval scales may be aligned by finding a common denominator (e.g., 100) or by converting each variable to standardized (“Z”) scores, as explained in Chapter 17. The second and final step is to add each indicator together ($X_1 + X_2 + X_3 + X_4$).

A multiplicative index combines indicators by multiplication, with the assumption that each component is equal but not independent (its contribution to the concept of interest depends on the values of other indicators). Here, too, it is important to examine the impact of diverse scales – although the correct solution is not always to align them to a common scale. For example, it may be important to maintain a binary scale (0/1) rather than to convert it to an interval scale.

Zero plays an especially important part in multiplicative indices. Note that if any indicator is equal to zero the resulting index will take on the value of zero, regardless of the value of other indicators. An example of this can be found in Tatu Vanhanen’s index of democracy, which combines two factors: (a) the competitiveness of an election (100 minus the share of the vote garnered by the largest party) and (b) the rate of participation in that election (turnout, understood as a ratio of the total population, from 0 to 100 percent). These are multiplied together and divided by 100 to form an index of democracy. Note that if only one party is allowed to compete in elections, factor (a) will take the value of zero. When multiplying this with (b) the participation rate, the result will also be zero – regardless of how many people show up to vote. This underscores an assumption, namely that participation does not count for democracy unless some degree of

---

57 Vanhanen (2000).
multiparty competition is allowed. North Korea is not more democratic than Cuba even if it has higher turnout elections.

A *factor-analytic* index is generated by factor analysis, with the assumption that each component is equal and independent. This approach uses an algorithm — usually *principal components* — to find the common dimension(s) among a number of indicators. Typically, researchers are only interested in the first dimension, the one that explains the most variance across the chosen indicators. But sometimes, they will retain other dimensions to be used as indices for concepts of theoretical interest.

In this fashion, Michael Coppedge, Angel Alvarez, and Claudia Maldonado explore empirical patterns across fifteen measures of democracy, including the well known Freedom House and Polity indices. They discover that about seventy-five percent of the variance in these measures is reducible to two relatively distinct components: *contestation* (competitive elections and associated institutions) and *inclusiveness* (e.g., broad suffrage and high turnout). Since these components have strong grounding in democratic theory, there are good reasons to regard them as more than empirical artifacts.

Having surveyed a few of the most common methods of aggregation it is important to point out that approaches to aggregation are much more variegated than is suggested by this short list. Researchers may combine different approaches. They may weight some factors more than others, with the notion that certain factors are a priori more important. Or they may construct aggregation approaches that bear little resemblance to any of the foregoing approaches.

By way of conclusion, we note that aggregation matters. The same indicators, put together in very different ways, may produce very different indices. Whatever solution to aggregation is chosen, the researcher should be sure that it is clear, explicit, and replicable. That is, it should be possible for another researcher to follow the choices made by the original researcher, thereby reconstructing the index.

**OBJECTIVES**

The construction of an indicator may aim to achieve maximum *discrimination* among entities or optimal *grouping* among entities. The first will likely utilize numeric scales and the second will likely enlist categorical scales, as discussed above. One or the other of these fundamental objectives governs all measurement instruments. Needless to say, a single instrument is unlikely to serve both goals at once.

Discrimination refers to the ability of an instrument to reveal finely graded differences of degree — usually unidimensional but occasionally multidimensional — in some latent trait possessed by a sample of people, objects, or events. Accordingly, a test of educational achievement should provide the basis for maximally sensitive scale (measuring differences in knowledge or ability in a subject among test-takers) with a minimal number of standardized questions. This requires that each question on the test be independent of all others and that each reflect different levels of the latent trait of interest (knowledge/ability in some subject area), thus adding to the information provided by the other questions. If two individuals with different levels of knowledge/ability give the same answer to the same question that question is not helping to discriminate between them (i.e. it is redundant). Likewise, if two questions are interdependent — such that an answer to question #2 depends (in some logical fashion) upon the answer given to question #1 — then no new information is learned from question #2. The result of a well-crafted measurement tool (constructed for the

---

58 While Coppedge, Alvarez, and Maldonado (2008) employ principal components analysis, other recent studies have enlisted Bayesian techniques (Pemstein, Meserve, and Melton 2010; Treier and Jackman 2008).
59 Dahl (1971).
purpose of maximum discrimination) is a finely-graded scale with no bunching, that is, scores are evenly distributed across the sample of respondents.

Grouping, on the other hand, refers to the ability of an instrument to sort items into discrete categories on the basis of similarities and differences in some latent trait(s). Common techniques include factor analysis, principal component analysis, cluster analysis, and Q-sort analysis. Note that the goal of crisp categories may not always be fully achievable. But it is the guiding objective. The success of a technique is its ability to sort items into discrete categories, apples with apples, oranges with oranges. If phenomena are not grouped naturally into categories ("apples," "oranges,"...), then the job of the measurement instrument is to discern break-points in numeric scales. This may be accomplished in an inductive or deductive manner – which brings us to our next topic.

**CONCLUSIONS**

Arguments take shape with concepts, and concepts take empirical shape through indicators. This is the succession of topics we have followed through Chapters 2 and 3. It is worth remembering that these topics are interwoven. Concepts are built for use in arguments; they don’t always make sense outside of that particular context. Indicators are inextricably linked to the concept they are intended to measure. They have no intrinsic meaning.

Nonetheless, in order to understand each component of social science methodology we need to take these components apart. That is what the foregoing chapters have attempted to do. In the next chapter, we look at the task of empirical analysis, i.e., how arguments are tested.
KEY TERMS

- Operationalization
- Resonance
- Internal coherence
- External differentiation
- Theoretical utility
- Scope-condition
- Dependent variable
- Independent variable
- Consistency
- Minimal definition
- Maximal definition
- Indicator
- Level/ladder of abstraction
- Conceptualization
- Measurement
- Scale (categorical, numeric, nominal, ordinal, interval, ratio)
- Binary
- Aggregation
- Boolean
- Necessary and sufficient conditions
- Additive
- Multiplicative
- Factor-analytic

INQUIRIES

1. Identify a concept of interest to you and then identify various indicators (or indices) that are used, or might be used, to measure that concept. What are the strengths and weaknesses of each indicator? Are these indicators likely to be highly correlated, or weakly correlated? (Will they identify different phenomena as members of the concept?) Are there any controversies surrounding that measurement instrument? How successful is it in representing the concept of theoretical interest? Are there parts of that concept that are left out or downplayed? Are there other problems of measurement associated with it? Could it be improved upon? What would your approach to measurement be?

2. Identify an abstract concept and figure out what the different levels of abstraction are within that concept, as we did with “democracy.” Are the lower levels of this concept easier to measure than the higher levels? (Is this always the case?) Is it possible to replace higher-order concepts with their lower-level components? Will this solve problems of conceptualization and measurement?
3. Pick a concept that is of interest to you and construct different sorts of scales to measure that concept, following the typology in Table 3.4. What are the pros and cons of different scales for that particular concept? (Which sorts of scales would be useful, and for what purposes?)

4. With abstract concepts, the biggest obstacle to measurement is usually aggregation – putting various indicators together into a single index. Why is this so difficult?

5. Identify a concept and try to operationalize it using each of the following aggregation principles: (1) boolean, (2) additive, and (3) multiplicative. If you are familiar with factor analysis, you may also attempt this approach to aggregation.

6. The purpose of an indicator may be to discriminate fine-differences among similar entities or to group together entities that share lots of characteristics. In what circumstances might one want to adopt one or another approach to a concept such as democracy? (What’s the point of discriminating, and what’s the point of grouping?) Consider a specific task such as disbursing money from development agencies to poor countries. One agency wishes to allocate money to countries that surpass a threshold of democracy. Another wishes to allocate money to countries that show improvement in the quality of democracy. What sort of approaches should one take to measurement, given these two divergent objectives?

7. Consider the question of media bias as a problem of conceptualization and measurement. Review several studies that take differing views on this vexed subject. See if you can explain why they come to different conclusions. Is there a good way to measure media bias? Here is an illustrative list of studies – both journalistic and social-scientific – that you can use to examine the topic (you needn’t include all of them):

8. How should we define and measure prejudice? Is it possible to solve this problem with experimental approaches, such as those exploited by Bertrand and Mullainathan (2004), Page (1998), and Sniderman and Carmines (1997)? Does observational data – such as that collected by Kirschenman and Neckerman (1991) – have a role to play?


9. Should concepts be measured by people’s own self-understandings or by their behavioral attributes? Consider the concept of partisanship in American politics. For many years, survey researchers have been asking respondents whether they identify as a member of a major political party (Democratic or Republican), or whether they consider themselves to be independent. If they answer the latter, the respondent is then asked whether he or she “leans” toward one of the major parties. This research shows that the number of people identifying as Independent has grown astronomically over the past several decades, suggesting a massive decline in partisanship. However, it has also been shown that those who declare themselves independent but lean towards one of the major parties – the “leaners” – behave much like partisans when we examine their voting behavior (Keith et al. 1992). This suggests that the major development over the past several decades is that Americans are less inclined to call themselves partisans but still behave like partisans. Should we ignore self-professed partisanship in favor of a “behavioral” measure of this concept?

10. Surveys seem to show that in the United States and in many other countries around the world citizens have lost trust in institutions. A recent crossnational poll discovered that “less than one in five respondents…believes a business or governmental leader will actually tel
the truth when confronted with a difficult issue.” At the same time, behavioral manifestations of trust seem to show a quite different picture. People invest in the stock market; they purchase products over the internet using a credit card; they exchange a great deal of personal information over the internet or by cell phone; they call the police when a crime has been committed; they follow directions from government officials; they are genuinely shocked when government officials behave in a corrupt fashion. In short, nothing about personal behavior patterns would suggest that there is a massive loss of trust in institutions of government or business. How should we measure “trust”? What are the consequences of different choices in measurement?

11. Identify a term that is central to your field of study. For present purposes, it probably makes sense to choose a term about which there is some disagreement among scholars. If you are having trouble finding such a term you might consider one of the following: accountability, affirmative action, balance of power, bureaucracy, civil liberty, clientelism, colonialism, corporation, corruption, culture, democracy, discrimination, ethnic group, fascism, federalism, genocide, globalization, ideology, media bias, minority, politics, power, prejudice, property rights, public goods, racism, religion, revolution, rule of passage, rule of law, slavery, social capital, social movement, social class, sovereignty, state formation, terrorism, transaction costs, or war.

(a) Collect a series of definitions drawn from the academic literature on the chosen subject. (Try to identify different kinds of definitions so there is some variety in your collection.) Why have different authors chosen to define the same term in different ways? Is there logic to each author’s choices?

(b) Next, try to identify the core attributes of the term—those attributes that are nonidiosyncratic—and arrange these attributes in a table (along the lines of Table 3.2).

(c) Now try to construct minimal and maximal definitions for the chosen term. Are they satisfactory? Is one more satisfactory than the other?

(d) Finally, decide which of the many possible ways of defining this term is most useful. This might involve selecting one of the definitions in your initial survey of definitions, or it might involve generating your own definition. Explain why you think this way of defining the topic is better than others.

RESOURCES

CONCEPTS


---

Examples:


Measures


Examples:

4. ANALYSES

We began (in Chapter 2) with arguments and proceeded (in Chapter 3) to conceptualization and measurement. In this chapter, we turn to the problem of how to analyze an argument empirically. This may be referred to variously as appraisal, assessment, corroboration, demonstration, empirics, evaluation, methods, proof, or testing. Pursued in a self-conscious fashion, this stage of research involves a research design, i.e., an explicit method of selecting and analyzing data.

We begin by introducing a set of key terms that are necessary to understand the construction of a research design. We proceed to a discussion of the general issues that all analyses encounter. This includes precision and validity, internal and external validity, sample representativeness, sample size, probability and non-probability sampling, and missing-ness. The terms and topics introduced in this chapter will enter the narrative in later chapters repeatedly. This chapter therefore plays a foundational role in the book.

DEFINITIONS

A standard empirical analysis involves a number of components, which must be clarified before we continue. Much of this vocabulary is borrowed from survey research. Nonetheless, the concepts are helpful in all styles of research, whether quantitative or qualitative, and are illustrated in Figure 4.1.

The most basic unit in any analysis is an observation. Observations are the pieces of evidence deemed relevant for demonstrating an argument. In a standard matrix (rectangular) dataset, an observation is usually represented as a row. Each row in Figure 4.1 represents a single observation.

Each observation should record values for all relevant variables. In causal analysis, this includes $X$ (the causal factor of theoretical interest) and $Y$ (the outcome of interest), along with any other variables deemed essential for the analysis. In a rectangular dataset, variables are usually represented with vertical lines. There are three variables in Figure 4.1: $X$, $C$, and $Y$.

An observation is drawn from a unit or case – bounded entities such as individuals, organizations, communities, or nation-states, which may be observed spatially and/or temporally (through time). The terms unit and case are more or less equivalent. (While a unit is bounded spatially, a case may also have implicit or explicit temporal boundaries.)

Collectively, the observations in a study comprise a study’s sample. The size of a sample is the total number of observations, often denoted as “$N$” ($N$ may also refer to the number of units or cases, which may be considerably less than the number of observations. This should be clear from the context.)

A population is the universe of phenomena that a hypothesis seeks to describe or explain. It usually remains unstudied, or is studied only in a very informal manner, e.g., through the secondary literature.

The sample (the observations that are actually studied) is drawn from the population, and is usually much smaller. Hence, the notion of sampling from a population. Note, however, that the term sample, as used here, does not imply that the studied observations have been chosen randomly from a population. This ideal is rarely followed in practice, as discussed below.

78
Occasionally, a set of observations includes the entire population of interest. This is known as a **census**. A population census includes all persons residing within a country (though of course coverage is never entirely comprehensive). Likewise, a census study of nation-states would include all nation-states. Since census studies (where \( N_{\text{sample}} = N_{\text{population}} \)) are rare, we leave them aside in the following discussion.

These inter-related concepts are illustrated in Figure 4.1, where we can see a fairly typical time-series cross-section research design in a rectangular dataset format. Recall, observations are represented as rows, variables as columns, and cells as their intersection. Note that cells are nested within observations, observations are nested within units (aka cases), units are nested within the sample, and the sample is nested within the population.

Hypothetically, let us imagine that the population of the inference includes all worker-training programs in the United States and the sample consists of eight programs, observed annually for five years, yielding a sample of forty observations (\( N=40 \)). The **units of analysis** (the type of phenomena treated as observations in an analysis) in this hypothetical example are program-years.

All these terms are slippery insofar as they depend for their meaning on a particular proposition and a corresponding research design. Any changes in that proposition may affect the sorts of phenomena that are classified as observations and units, not to mention the composition of the sample and the population. Thus, an investigation of worker-training programs might begin by identifying **programs** as the principal unit of analysis but then shift to a lower **level of analysis** (e.g., **participants**) or a higher level of analysis (e.g., **state**) at different points in the study. Sometimes, different levels of analysis are combined in a single study. This is common in **case study** work (see Chapter 9) and is the defining feature of **hierarchical (multi-level)** statistical models.
Figure 4.1: Time-Series Cross-Section Dataset

Population \( (N) \) = indeterminate   Cases/Units \( (N) \) = 8   Sample/Observations \( (N) \) = 40   Cells \( (N) \) = 120   Time-periods \( (T) \) \( (N) \) = 1-5   Variables \( (N) \) = 3
Before leaving this discussion of basic terms we must address the important distinction between quantitative and qualitative analysis. This contrast is ubiquitous, and will no doubt be familiar to the reader. But it is also ambiguous since these terms can be defined in many different ways, and sometimes they are not defined at all. In this text, we adopt the following definitions.

**Quantitative** analysis is a formal analysis of matrix-based observations. A matrix observation is the conventional sort, represented as a row in a rectangular dataset (illustrated in Figure 4.1. Each observation is coded along a number of dimensions understood as columns in the matrix and as variables in an analysis. All observations are regarded as examples of the same general phenomenon and are presumed to have been drawn from the same population. Each is regarded as comparable to all the others (with some degree of error) with respect to whatever analysis is undertaken. The analysis is “formal” insofar as it rests on an explicit framework of inference such as logic/set theory, Bayesian statistics, frequentist statistics, or randomization inference.

By contrast, **qualitative** analysis refers to an informal analysis of non-comparable observations. Non-comparable observations cannot be arrayed in a matrix format because they are examples of different things, drawn from different populations. The analysis is “informal” insofar as it is articulated with natural language and is unconnected to an explicit and general framework of inference. When applied in the context of causal inference this sort of evidence may be referred to as causal-process observations, cues, colligation, congruence, genetic explanation, narrative analysis, or process tracing.

There is a strong elective affinity between quantitative analysis and large samples, as well as between qualitative analysis and small samples. One would be hard-pressed to apply informal styles of analysis to a sample of 1,000. Likewise, one would be hard-pressed to apply a formal analysis to a sample of 2. The size of a sample thus influences the style of analysis. However, it does not determine it. This is apparent in the middle range. A sample of 20 may be analyzed formally or informally (or both). Thus, when we use the terms quantitative and qualitative the reader should understand that the former usually (but not always) corresponds to large samples and the latter usually (but not always) corresponds to small samples. The qual/quant distinction is not solely a matter of \( n \).

So defined, there is no epistemological gulf separating quantitative and qualitative analysis. Indeed, any qualitative analysis can be quantified – with the cost of reducing complex, multifaceted data into matrix observations. Many quantitative datasets began life, one might say, as a series of qualitative observations. These were then coded in a systematic fashion to generate a set of observations that could be arrayed in a dataset. Often, this is useful. But not always; nor is it always possible. That is why qualitative analysis continues to play an important role in social science research – and especially in case study research, as discussed in Chapter 9.

That said, quantitative analysis is dominant in many fields and also has a more developed methodology. Consequently, when we use the term “observation” or “analysis” in this book the reader can assume that we are talking about the quantitative variety, unless stated otherwise.62

---

**PRECISION & VALIDITY**

Social science analyses strive for **precision** (aka **reliability**) and **validity** (aka the absence of **bias**). **Precision** refers to the closeness of repeated estimates of the phenomenon of interest when using

---

61 We do not mean to imply that these terms are identical, merely that they are overlapping.

62 For further discussion, see Gerring (2016b).
the same measurement instrument or causal model. **Validity** refers to the closeness of an estimate to the true (often unknown) value. To explore these concepts we employ a hypothetical example.

In recent years, survey researchers have employed a number of techniques to obtain representative samples of the general public. These include (a) hardline telephones, (b) hardline phones and cell phones, and (c) door-to-door canvassing.

In order to evaluate their precision and validity, we shall imagine employing each of these recruitment techniques to conduct 20 surveys of the general public in the United States. Each survey includes 2000 participants, who are chosen in whatever fashion the pollster believes will result in the most representative sample. There is only one question on the survey: What is your family income? For each sample, we calculate the **median** (that value for which there are equal numbers of values above and below). These results are plotted on the graphs in Figure 4.2.

We shall assume that census results represent the true value of household income in the country. In 2006, the United States Census Bureau reported that the median annual household income was roughly $50,000. This value is represented by a blue vertical bar in Figure 4.2, and is the quantity of theoretical interest.

Results for the first recruitment technique – hardline telephones – are presented in panel (a) of Figure 4.2. They suggest that this method is fairly precise, as the estimates cluster tightly around the sample mean. However, the sample mean falls far from the true value (the median value for the population, as revealed by census data), suggesting that the method is not valid.

Results for the second recruitment technique – hardline phones and cell phones – are presented in panel (b). They suggest that the method is imprecise, as estimates vary widely around the sample mean. However, the sample mean falls close to the true value, so the technique may be regarded as valid (in repeated sampling).

Results for the third recruitment technique – door-to-door canvassing – presented in panel (c), is both precise (tightly clustered) and valid (close to the population value of interest). Thus, on the basis of this set of tests, door-to-door canvasing is superior to the other sampling techniques. Of course, these results are entirely hypothetical. You should also note that many other factors – including cost – may figure into a pollster's decision to employ a method of sampling. Nonetheless, the illustration is effective in demonstrating a crucial distinction between precision and validity.
Methods of sampling: (a) hardline telephones, (b) hardline phones and cell phones, (c) door-to-door canvassing. Vertical line: true value, according to US census.
Let us now elaborate a bit on these concepts and their employment in social science research.

Precision, or reliability, refers to level of stochastic (random) error, or noise, encountered in measurement or some other feature of estimation. Precision in measurement is often assessable through **reliability tests**, where the same technique of data gathering is employed multiple times in order to ascertain its reliability. Reliability tests might focus on a survey technique, as in our example, on experts who code data from a primary or secondary source, or on any aspect of data collection. If multiple applications of the measurement instrument reveal a high level of consistency one may regard the chosen instrument or model as reliable (precise). This is typically calculated as the inverse of the **variance** (i.e., dispersion around the mean). Greater variance means less reliability. The same logic applies to precision in causal inference, i.e., when one is comparing models employed to assess a causal relationship between X and Y.

If the opportunity to test multiple iterations of an indicator or model is not present then the issue of reliability remains at the level of an assumption. But it is nonetheless crucial. A high probability of random error may doom even the simplest generalization about the world.

Validity, by contrast, refers to systematic measurement error, error that – by definition – introduces bias into the resulting analysis. One often finds, for example, that the level of accuracy with which an indicator is measured varies directly with some factor of theoretical interest. For example, in constructing international indicators for human development (e.g., life expectancy, literacy) and economic performance (e.g., inflation, GDP) we rely on surveys conducted in countries throughout the world. Sometimes, these surveys present a more favorable picture of the quality of life or the strength of the economy than is warranted. And there is some reason to imagine that autocratic governments engage in this practice to a greater extent than democratic governments. If so, these indicators suffer from systematic bias. However, because we cannot be sure of this bias, and have no estimate of its extent, we have no easy means to correct it.

While we can usually assess reliability we rarely have a fix on the problem of validity. Note that our hypothetical example is highly unusual in one important respect: we know the true value of the measure of interest — $50,000. In the world of social science it is rare to possess a definitive measure of anything of great substantive importance. Indeed, our hypothetical example might be challenged on the ground that population censuses are never entirely comprehensive; some citizens escape the prying eyes of government surveyors. In the 2000 US Census, for example, despite elaborate advertising and outreach, the overall mail response rate was only 64%. Worryingly, these non-respondents may be quite different than respondents, leading to systematic bias, a problem the Census Bureau is aware of and attempts to evaluate. For example, homeless people are less likely to be contacted in a census and since homeless people have much lower family income than people with stable addresses it is reasonable to suppose that all census-based data (unless adjusted to correct for this deficiency) is biased to some extent. Various sociodemographic groups and geographic areas have been shown to be underrepresented, especially Blacks and Hispanics.

All of this is to say that validity, unlike reliability, is very difficult to test in a definitive fashion. Even our toy example is open to dispute. And with more complex concepts such as democracy or Gross Domestic Product (GDP) the points of potential dispute are multiplied. For causal inferences, which build on descriptive inferences, the problem is magnified even further.

---

63 Stackhouse (2002).
64 Griffin (2002).
INTERNAL & EXTERNAL VALIDITY

Typically, researchers examine only a small number of instances of a phenomenon of interest. If one is trying to ascertain the median income within a large population, as in the example explored above, one might sample only 2000 respondents. On this basis one would hope to infer median income across 320 million people. Even more striking is the attempt to learn something about human nature from experiments conducted on a sample of college students drawn from a single classroom, a common practice in psychology. Here, a sample of several hundred may be the basis for generalizations about the entire human race.

In either case, social science must grapple with a crucial question: how to relate findings drawn from a sample to a larger population of interest.

This problem engages researchers in a two-level game. The first part of the game concerns reaching conclusions about the sample. The second part of the game is about extrapolating those conclusions to a larger population, sometimes referred to as a problem of inference.

To distinguish these two spheres of truth social scientists invoke a distinction between internal and external validity. The first refers to the validity of a hypothesis with respect to the studied sample. In our previous example, a problem of internal validity may arise if respondents lie about their family income, perhaps in response to perceived norms. If so, a calculation of mean family income for a single sample – or a group of samples – may be too high or too low.

The second issue arises when we try to extend sample-based results to a larger population. Naturally, if there are problems of internal validity there are likely to be problems of external validity. But even if our analyses of the sample are correct they may not be correct across a larger population. This is the problem of external validity, toward which the rest of the chapter is directed.

With respect to external validity, two characteristics of a sample are especially relevant: sample representativeness and sample size.

SAMPLE REPRESENTATIVENESS

The external validity of a study is grounded largely on the representativeness of a chosen sample. Is the sample similar to the population with respect to the hypothesis that is being tested? Are we entitled to generalize from a given sample to a larger universe of cases?

In the case of research into income one must consider whether the sample exhibits the same income distribution as the general population. In the case of research into cognitive properties of human nature one must consider whether college students are similar in these respects to other humans (and how far back in time a result might be generalizable)? And with respect to studies of worker-training programs one must consider whether the chosen program sites are representative of a larger population of programs that one wishes to understand.

Note that questions about representativeness are also questions about how to define the population. Consider the study of worker-training programs that focuses on programs in the state of New York. It could be that results from this study are generalizable (a) to that state (only), (b) to the United States, (c) to advanced industrial societies, or (d) to all societies in the contemporary era. Likewise, it could be that the results are generalizable (a) to native-born unemployed persons between the ages of 20 and 50 without disabilities, (b) to unemployed people between the ages of 20 and 50 without disabilities, (c) to people between the ages of 20 and 50 without disabilities, (d) to people without disabilities, or (e) to all people.
The point is that any consideration of external validity forces one to be very specific about the population and the scope-conditions of an inference. What is the population, exactly? What is it, exactly, that is generalizable to that population?

Unfortunately, these questions are often difficult to answer in a definitive fashion, for reasons already discussed (see Boundedness in Chapter 2). However, they must be addressed, even if only in a speculative fashion.

**SAMPLE SIZE (N)**

The second characteristic that impacts a study's external validity is the size of the sample upon which the study is based. More observations are better than fewer because they provide a more precise estimate of the quantity of interest.

Suppose one is trying to figure out the effect of a worker-training program on employment prospects or earnings but one has available information for only one program. Under the circumstances, it will probably be difficult to reach any firm conclusions about the matter. Of course, one case is a lot better than none. Indeed, it is a quantum leap. Yet, empirical research with only one case is also highly indeterminate, and apt to be consistent with a wide variety of competing hypotheses. Conclusions about a broader population are hazardous when one considers the many opportunities for error and the highly stochastic nature of most social phenomena.

One sort of problem stems from problems of sampling error encountered when the sample is very small. Note in order to make accurate inferences about a larger population one must have a sample that is similar to that population in relevant respects. The chances of finding such a sample when the sample is small are considerably reduced. One's chances of achieving a representative sample increase with sample size – if the sample is chosen randomly from the population. We provide a more detailed look at the importance of sample size in probability sampling in Chapter 21.

Uncertainty associated with sampling variability is captured in a statistic known as a **confidence interval**. A confidence interval indicates the bounds within which a true value is likely to fall at a chosen level of probability. A 95% confidence interval means that our confidence interval captures the true value 95% of the time. A 90% confidence interval means that our confidence interval captures the true value 90% of the time. We describe how to calculate confidence intervals and provide a more precise treatment of their interpretation in Chapter 21.

A larger sample is advisable if everything else is equal. Of course, this is sometimes not the case. For example, sometimes increasing the size of a sample decreases its representativeness. Consider a sample that is representative. Now add cases non-randomly. Chances are, the larger sample is less representative than the smaller sample. Likewise, sometimes one can gain greater leverage on a question with a carefully chosen small sample than with a large sample; this is the justification for purposive case-selection procedures, discussed in the context of case-study research (Chapter 9). In particular, an empirical study whose sole purpose is to disprove an invariant causal or descriptive law can achieve this purpose with a single observation – so long as it contradicts the hypothesis.

The point remains, obtaining a large sample is a noble objective so long as it doesn’t interfere with other goals. When it does, the researcher needs to decide which goals to prioritize – or, alternatively, adopt a multi-method research design that incorporates large- and small-N components.

How large is large enough? What is an appropriate size for a sample? There is no easy answer to this question. It is not the case that sample size should be proportional to the size of the
population. Sampling error rests primarily on the size of the sample not the sample/population ratio. Following precedent, i.e., what other scholars have done, may be appropriate – but only if the goals of your analysis are similar to theirs. Occasionally, one may calculate an appropriate sample size by stipulating the goals of the analysis and an acceptable confidence interval for the variable of theoretical interest. However, this only works if there is a single hypothesis (not a multitude of hypotheses) and if it is possible to identify a benchmark confidence interval (which is not always possible).

Requisite sample size depends, in general, in the relative strength of the ‘signal’ (the variable of theoretical interest) relative to background ‘noise’ (factors that might muffle the signal). Let us say that instead of seeking to estimate the height of individuals within a population we were interested in estimating the impact of a daylong worker-training program on subsequent earnings over the succeeding two years. As one might imagine, the impact of a single day’s training on subsequent earnings is likely to be fairly minimal, and many factors other than training affect earnings. In this setting, one would presume that the ratio of signal to noise is pretty low, requiring a great many observations in order to discern a causal effect (if indeed there is one).

**Probability Sampling**

The preferred approach to sampling is to choose cases randomly from the population. Because cases are chosen randomly, one knows the probability that any given case will be chosen as a member of the sample. This approach to sampling is therefore referred to as **probability sampling**.

In **simple random sampling**, each case within the population has an equal chance of being selected for the sample. (This is sometimes referred to as an *equal probability sample.*) The mechanism of selection might be drawing balls from an urn, as in raffles. More commonly, random-selection is achieved with a random-number generator from a computer program. (These may be found on-line or as part of a software package.) The statistics we introduce in subsequent assume random sampling.

In **systematic sampling**, members of a population are chosen at equal intervals, e.g., every 10th or every 1,000th. This assures equal probability sampling if the chosen interval is not associated with any particular feature of the population, a matter that may be difficult to discern.

In **cluster sampling**, members of a population are divided into clusters (groups), clusters are chosen randomly (using some random-selection mechanism), and then each member of the cluster is automatically included in the sample. This approach is usually taken for logistical reasons, i.e., when it is easy to include all members of a naturally occurring cluster such as a school, neighborhood, census tract, or family.

In **stratified sampling**, each member of the population is assigned to a stratum and cases are chosen randomly from within each stratum. If the number chosen from each strata does not reflect the proportional size of that strata within the population, cases will need to be re-weighted so that the resulting sample is representative of the population. For example, in a sample of 2,000 individuals drawn from the United States it may be important for theoretical reasons to identify strata composed of a various minority groups. While some minorities like Hispanics and African-Americans are quite large, others such as Jewish-Americans are quite small (roughly 3% of the general population). A stratum composed of 3% of 2,000 yields a sub-sample of only 60 individuals – too small to allow for precise estimates of the actual population of Jews in the US. Under the circumstances, it probably makes sense to *over-sample* among Jews, granting Jews a greater probability of entering the sample than members of other social groups. Let us say that the researcher decides...
to select twice as many Jews as their share of the population would allow, raising their share of the sample to 6%. Representativeness can then be restored to the sample by down-weighting Jews in the sample – in this case, granting half the weight to Jewish respondents as to other respondents. This approach is generally cheaper to implement than the alternative approach – doubling the size of the entire sample (e.g., from 2,000 to 4,000) – and achieves the same results. Naturally, it requires that one identify those strata which are of theoretical interest prior to drawing the sample.

Various approaches – simple, systematic, cluster, stratified – may be combined in multi-stage sampling. For example, one might begin with clusters, stratify within clusters, and then sample randomly within each cluster/stratum. The key point is that whatever system of probability sampling is employed, disproportionalities should be corrected (by weighting) so that the sample is representative of the population of interest.

A key advantage of probability sampling is that one can estimate sampling variability (from sample to sample), thus providing estimates of precision to accompany whatever inferences one wishes to draw. It is not enough to say that a sample is “large” and therefore “precise.” One wishes, as well, to know how precise a given sample estimate is, that is, how close it is likely to be to the population parameter.

**Non-probability Sampling**

A very different approach to sampling is to select cases non-randomly from a population. A small number of cases may be selected with specific purposes in mind, as in case-study designs (see Chapter 9). Cases may be chosen in a snowball fashion. This approach is common in interview-based research, when one relies on each respondent to suggest other possible respondents – creating a snowball effect (the ball of respondents gets larger as each new snowflake joins the ball). Cases may also be chosen for logistical reasons, e.g., because they are accessible, cheap, or for some other reason easy to study. This is sometimes known as convenience sampling.

In all of these approaches the researcher has no way of assessing the probability, for each case in the population, of being selected into the sample. Accordingly, these approaches are sometimes referred to as non-probability samples. Such approaches produce samples of uncertain representativeness. We don’t know how similar they are to the population of interest; likewise, we may be uncertain about what that population is. If the sample is small, as it is with case-study designs, then the study also faces a problem of reliability (imprecision).

From the perspective of external validity there is little to be said in favor of non-probability sampling. However, external validity is not the only goal of social science research, and other goals sometimes require non-probability approaches to sampling.

At an early stage of investigation, when not much is known about the phenomenon of interest and before one has identified a specific hypothesis, it is common to focus on a small number of cases so that one may observe those cases in an intensive fashion. Likewise, if one already has a clear sense of a relationship but one does not know why it obtains, one might prefer to focus on a small number of cases, intensively observed. These are classic justifications for a case-study research design.

While it may be feasible to select a single case, or several cases, randomly from a population, our earlier discussion – of sample size – shows how unreliable such tiny samples can be. For this reason, one is generally advised not to use a probability-based method for choosing a very small sample. The exception would be a situation in which cases found within the same stratum are all equally informative. Here, one may elect to choose cases within that stratum randomly. While this
introduces an element of random-ness, it applies only to the chosen stratum or strata. Presumably, not all strata would be included so that the resulting sample remains un-representative of the population.

Even where studies incorporate a large sample it still may be undesirable, or impossible, to implement probability-based sampling procedures. Worker-training programs cannot draw randomly from the universe of unemployed people because many people will refuse to participate. As such, the sample of subjects analyzed in such a study are not likely to be representative of the larger population of unemployed – though they might be considered representative of a smaller population vaguely defined as “those who are willing to participate in a worker-training program.”

Wherever random sampling techniques are inapplicable researchers must struggle to define the representativeness of a sample, and hence the plausible generalizability of results based on that sample. This is true regardless of whether the sample is very small (i.e., a case study format) or very large.

**MISSING-NESS**

In the discussion so far we have assumed that all cases in the population can be accessed through probability sampling procedures and that the chosen cases can be included in the sample, i.e., they can be studied. Unfortunately, this does not always hold.

There may be slippage between the population and the **sampling frame**, those members of the population who are accessible to the probability sampling procedure. For example, if a survey is conducted by telephone the only persons who can become members of the sample are those who have telephones. Thus, if one has access to all telephones – through random-digit dialing – one can obtain a representative sample of telephone owners.

Another source of bias occurs when a chosen case cannot be studied, or can be studied only partially. This might be because a respondent refuses to participate in a study (non-response). It might be because the respondent completes only part of a survey or does not adhere to the protocol of an experiment (non-compliance). It might be because a chosen case is especially sensitive, for political or ethical reasons, and therefore cannot be included in the sample. In a historical study, it might be that a chosen case does not offer the archival records that would be required in order to conduct the study. Lots of things may intervene to thwart the goals of a sampling procedure.

We shall refer to these problems generically as **missing-ness**, that is missing data. What is meant by missing data is that a sample lacks observations for some units that should (by some probability-based principle of selection) be included.

If the pattern of missing data is random it causes little harm. Suppose that those who own phones and agree to conduct a survey are no different (in relevant respects) to those who do not. The survey researcher need only increase the number of calls in order to obtain the desired sample size, which will in any case be representative.

If, however, the pattern of missing-ness is systematic then the sample will be biased. For example, if telephone owners are different from those who don’t own telephones in ways that are relevant to the analysis, estimates will be biased away from the (true) population parameters. (A good deal of research has gone into this question, with inconclusive results.)

A potential solution is to fill in missing data, creating a full sample that is larger and – one hopes – more representative than the truncated sample. If one has a good guess about the nature of the missing data one may develop a simple decision rule for filling in missing observations. For example, if one knows (from other sources) the mean income of persons without telephones, and
their share of the general population, one might use this value for all such phantom respondents, thus rectifying the non-representativeness of the sampling frame.

Another approach is to employ a statistical model (an algorithm) to estimate missing values based on patterns in the data that has been gathered. This requires knowing something about the cases that have missing values. Let us say that we know their telephone prefix, and that we can safely assume people with the same prefixes share certain characteristics (because they live in the same area or were assigned their cell phone number at the same time). On this basis, we might estimate the income of these non-respondents based on information that we have already collected from respondents with the same prefix who answered their phones and completed the survey.66

CONCLUSIONS

In this chapter, we have introduced the core elements of empirical analysis in social science, applicable to both descriptive and causal analyses. After reviewing key terms, we discussed the twin goals of precision and validity. Next, we distinguished between internal and external validity. The final sections of the chapter deal with the attempt to achieve external validity. This quest involves sample representativeness and sample size. Specific techniques include probability sampling, missing-ness, and non-probability sampling.

By way of conclusion, it is worth pointing out the obvious: all empirical knowledge is to some extent uncertain. This stems, arguably, from stochastic features of the world at a subatomic level – a matter debated by scholars of philosophy of science and physics. It stems, in any case, from our inability to attain perfect understanding of the complex world around us. Consequently, there is always a degree of uncertainty about any statement we might make about the world, even if the degree of uncertainty is judged to be quite small. Note that as the topic increases in importance the level of uncertainty usually rises in tandem. We can make mundane statements about the world with a high level of certainty, but we cannot pronounce upon the causes of democratization, or the causes of war and peace, with such assurance. It follows that the most relevant work in social science – in the sense of addressing issues that ordinary people care deeply about – is often accompanied by a high degree of uncertainty.

One of the features that distinguishes science from other modes of apprehending the world – for which “journalism” is our convenient foil – is the attempt to represent uncertainty in a forthright and realistic manner (rather than sweeping it under the rug, so to speak). To that end, we must address a common misunderstanding.

Measures of precision, as discussed in this chapter, usually encompass only one source of uncertainty – that associated with sampling from a population. Other sources of uncertainty such as that associated with measurement (discussed in Chapter 3) or causal inference (discussed in Part II) are not generally incorporated into sampling-based statistics. Thus, when you encounter terms that purport to measure uncertainty – e.g., variance, standard error, t statistic, confidence interval – it is important to bear in mind that these statistics are probably taking account of only one threat to inference. Other threats to inference, although more common and more problematic, are harder to measure in an objective fashion, and hence go unreported or are dealt with in prose. This is another reminder that methodological adequacy is often not summmarizable in handy statistical formats. You have to slog through the details of a research design to see its strengths and weaknesses.

66 Allison (2002).
KEY TERMS

- Research design
- Observation
- Variable
- Unit/case
- Sample
- Population
- Census
- Unit of analysis
- Quantitative
- Qualitative
- Precision
- Validity
- Median
- Reliability tests
- Variance
- Internal/external validity
- Representativeness
- Sample size
- Confidence interval
- Probability sampling
- Simple random sampling
- Systematic sampling
- Cluster sampling
- Stratified sampling
- Sampling frame
- Missing-ness
- Nonprobability sampling
- Snowball sampling
- Convenience sampling
- Inference
- Level of analysis
- Noise

INQUIRIES

1. Take a close look at a specific study and try to answer the following questions. What is the main hypothesis (or, if there are several, what is a main hypothesis)? What is the population of
this inference? What is the sample of observations that is being studied? What are the cases or units of interest?

2. Sometimes, the foregoing terms are difficult to define. Why? What makes them slippery?

3. Occasionally, a study will employ several research designs, each of which operates at a different level of analysis. What does this mean?

4. Think of some examples of work that is high in internal validity but low in (suspected) external validity. Is it possible for a work to be high in external validity but low in internal validity?

5. If you select a very small sample randomly from a population, any analysis conducted upon that sample is regarded as unbiased but highly imprecise. Explain.

6. Increasing the size of a sample increases its precision but not necessarily its representativeness. Explain.

7. What is the difference between random error and systematic error? Explain.

8. What are the different methods of drawing a representative sample from a population?

9. Why is it not always feasible to draw samples randomly?

10. Suppose you wish to identify a random sample of songs in a particular genre (e.g., hip-hop or contemporary jazz). How would you go about doing this? What would be the obstacles?

**RESOURCES**

**SAMPLING**

II. CAUSALITY

Causation is the central explanatory trope by which relationships among persons and things are established – the cement of the universe, in Hume’s words. Without some understanding of who is doing what to whom we cannot make sense of the world that we live in, we cannot hold people and institutions accountable for their actions, and we cannot act efficaciously in the world. Without a causal understanding of the world it is unlikely that we could navigate even the most mundane details of our lives, much less plan for the future. This is obvious in the policy world, where causal understanding undergirds any rational intervention. One must have some sense of what impact a policy is likely to have in order to support its adoption.

Even where causal understanding does not relate to future changes in the status quo we are likely to be reassured when we can order events around us into cause-and-effect relationships. “When we have such understanding,” notes Judea Pearl, “we feel ‘in control’ even if we have no practical way of controlling things.” Causality is not just a methodological preoccupation. It is also a way of relating to the world. That said, there are important differences between causal inference in everyday contexts and in social-scientific contexts.

Chapter 5 introduces a variety of causal frameworks that are widely employed in social science today. They may be viewed as the building blocks of a causal explanation. Chapter 6 defines the topic of causality, lays out the attributes of a good causal hypothesis and the core components of causal analysis. Chapter 7 discusses experimental research designs, where the causal factor of interest is randomized across groups. Chapter 8 discusses large-N observational designs (i.e., nonexperimental designs). Chapter 9 deals with case study designs, where the number of units is limited to one or several. Chapter 10 reviews and reflects on various aspects of causal inference, serving as a coda for this part of the book.

---

67 Hume (1888).
5. CAUSAL FRAMEWORKS

There are many sorts of causal factors and many sorts of causal mechanisms. Indeed, the variety of causal frameworks is so great that no overview can hope to be comprehensive. Even so, certain explanations are common enough across the social sciences to deserve our attention.

Three general causal frameworks are ubiquitous. Motivational frameworks include interests, norms, and psychology. Structural frameworks include material factors, human capital/demography, and institutions. Interactive frameworks include explanations grounded in adaptation, coordination, diffusion, networks, and path dependence. These varied approaches to explanation are summarized in Table 5.1.

In this chapter, we lay out each framework along with a few illustrative examples. The final section of the chapter adds important clarifications to this scaffolding and discusses how the various frameworks might be put together to provide a coherent overall explanation. Citations to the literature are located, as usual, at the end of the chapter.

By providing an overview of these frameworks we are providing an overview of the theories that inform the social science disciplines, as well as an entrée to the topic of causality. Although this chapter is not essential to those that follow, it will help you to have a sense for the substantive theories that inform later discussions of causal inference.
Table 5.1: Causal Frameworks

**Motivational frameworks**
- **Interests** (that which benefits an actor)
- **Norms** (values, beliefs, ideology, culture, identity)
- **Psychology** (cognition, emotion, personality, genetics)

**Structural frameworks**
- **Material factors** (financial resources, geography, technology)
- **Human capital/Demography** (health, education, skills, migration, life-cycle events)
- **Institutions** (formal rules and organizations)

**Interactive frameworks**
- **Adaptation** (feedback, competition, selection, evolution, efficiency, functionalism)
- **Coordination** (collective action, externalities, credible commitment, transaction costs)
- **Diffusion** (demonstration effect, exposure, ideas, information, knowledge, exposure)
- **Networks** (informal associational structures, social capital)
- **Path dependence** (contingent moments [critical junctures] leading to fixed trajectories)

**MOTIVATIONAL FRAMEWORKS**

One set of explanations center on motivations, i.e., what impels people to behave in certain ways or to think about things in certain ways. Three **motivational frameworks** predominate in the work of social science: interests, norms, and psychology.

One need not worry about which of these frameworks is “rational” or “irrational.” (Whatever conclusions one might draw would depend upon how one chooses to define that complex term.) The point is that they motivate, and thereby help to explain, human attitudes and behavior.

**INTERESTS**

Attitudes, preferences, and behavior are often the product of interests, i.e., what it is in the interest of a person or group to think or to do. That includes material factors (e.g., money), power, or status.69

Interest-based explanations play a key role in Aristotle’s *Politics* and are also a hallmark of Marxist theory and of studies that adopt an economic interpretation of human behavior, including work by Charles Beard, Gary Becker, and the collaborative team of Daron Acemoglu & James Robinson.70 Accordingly, if one wishes to explain why elites generally resist revolutionary change, and why members of the landless proletariat occasionally support fundamental change, one might begin with the notion that each group has something to gain by the positions they take in this

---

69 A broader definition, including whatever individuals define as worthwhile, becomes so diffuse as to lose its explanatory power or merges into a preference-based analysis, as discussed at the end of the chapter.

70 E.g., Acemoglu, Robinson (2012), Beard (1913), Becker (1976).
struggle. Likewise, if one wishes to explain why rich people generally support conservative parties and poor people generally support liberal or left-wing parties, one might suppose that these well-established partisan preferences are grounded in the divergent interests of these social classes – in the one case for property rights and limited taxes and in the other case for redistribution of wealth. By the same token, a reasonable starting point for any analysis of inter-state relations is that each state has a set of interests – e.g., land, resources, power – that it wishes to protect, and if possible expand. This is the foundation of the realist paradigm of international relations. And finally, in understanding the behavior of politicians a reasonable starting-point is the assumption that politicians desire to attain public office and, once in office, desire to maintain their position.

Interest-based explanations are often self-evident, a primal explanatory trope. When we are asked to explain a phenomenon we are apt to look first to self-interest as an explanation. However, this sort of explanation is not always as clear-cut as it may seem.

First, the various goals assembled under the rubric of interests – money, power, and status – are not always in synch. For a politician, staying in office may maximize her power while leaving office may enhance her pocketbook (if more remunerative jobs are available in the private sector). Here, two interest-based explanations collide.

Second, short- and long-run interests may conflict. For example, it may be in one's short-term interest to purchase consumption goods and in one's long-term interest to invest. To take another example, one may maximize short-term gain by stealing but long-term financial gain is probably maximized by working hard and following prevailing rules and norms. Different temporal frames thus suggest different courses of action, and the choice between the two will depend upon other (non-interest based) factors.

Third, actors often find themselves with conflicting interests based on different social identities. To simplify, let's think of interests as referring solely to money. If a person sacrifices to pay for her child's education she is sacrificing her interest as an individual but satisfying her interest as a parent. Likewise, one's financial interest as a member of one group (say, an ethnic group) may conflict with one's financial interest as a member of another group (say, a social class).

Fourth, since interests are often unclear, actors in the same objective position may nonetheless have different ideas about their (true) interests. For example, some members of the working class may believe their interests are best served by a socialist program while others believe their interests are best served by a conservative or liberal program. Karl Marx regarded members of the working class as victims of “false consciousness” if they did not support the communist cause. Others might argue that working classes rarely stand to benefit from communist revolutions, and thus their true interests lie with capitalism.

To sum up, it is often possible to infer attitudes and behavior from an individual’s (or group’s) position by assuming that people are acting to maximize their perquisites, power, or status (according to some definition of what that is). There is almost always a self-interest explanation for a persistent pattern of behavior. However, the plasticity of the concept of self-interest – the multitude of things that seem to satisfy this explanatory framework – should give one pause. There is a sense in which self-interest explains everything, and nothing.

NORMS

Interest-based explanations are often contrasted with those based on norms. By this, we mean to include values, ideologies, beliefs, belief-systems, philosophies, worldviews, religions, cultures, identities, and other related concepts. When we use these terms we suppose that the norm in question is not simply a reflection of interests (as described above) or of factual information about the world (as described under the rubric of diffusion below). We also suppose that a norm carries
special force. It is what we ought to do, i.e., what is expected of us (by some reference group whose opinions we value). Norms follow a logic of appropriateness.

Norms may be regarded as specific to a particular action such as the norm against incest. They may also be understood as part of a broader ideational system—an ideology, culture, belief-system, or value-system. For those who define themselves in religious terms the norm against incest would be understood as an element of their (Christian, Muslim, Buddhist, et al.) belief-system rather than as an isolated proscription. Identification with a social group (e.g., an ethnic group, national group, professional group, or college cohort) entails sharing the beliefs and values of that group. It is this set of norms that help to define the collectivity.

Norms may be set out in a formal manner, e.g., the Old Testament’s Ten Commandments, or they may take the form of informal understandings. The point is that they are shared and that they command obedience for reasons that reach beyond the interests of the actor. One feels a sense of obligation to a norm; if there is no such obligation then it is not a norm.

However, our use of the term presupposes that there is no explicit sanction for misbehavior or organization charged with maintaining that sanction. This is what distinguishes a norm from an institution (discussed below). For example, an electoral norm discourages politicians from lying to constituents in order to get elected. An electoral rule forbids politicians who are not citizens of a country from running for office. Infringement of the first brings no specific sanctions. Infringement of the second brings disqualification. This, in a nutshell, is the difference between a norm and a rule or institution, as we use these terms.

PSYCHOLOGY

In addition to interests, norms, and information, a person’s attitudes and behavior may be affected by the way in which stimuli are processed. We shall refer to this set of explanatory factors as the realm of psychology.

Arguably, all individual-level behavior rests on some sort of psychological explanation. Sometimes, this explanation is so simple that we don’t bother to delve into “psychology.” When a consumer purchases a product after the latter is discounted this sort of behavior is consistent with an interest-based explanation and requires little comment. However, we could imagine a psychological explanation that would construct a micro-level foundation for this behavior. Someday, it may be possible to trace these micro-foundations with indicators of brain activity, as we can already do in a crude fashion with some emotions and cognitive faculties.

When writers invoke psychology it is usually to explain behavior that doesn’t seem to be based solely on a calculation of self-interest, the logic of appropriateness, or the inherent truth of an idea (as revealed by some diffusion process). For example, in understanding conflict of a violent nature it may be important to understand the emotions of the actors in that conflict, emotions such as hatred, fear, or resentment. In understanding the role of information it may be important to understand how individuals calculate uncertainty. Research has shown, for example, that most people count losses more than gains, that they discount the future heavily, and that information is more believable if it is propounded by someone from their own social or political group or from a trusted authority.

Some psychological factors are widely shared among members of the species; they are universal, or nearly so. Other psychological factors may be specific to men or women or to persons of a specific age group. Others may be specific to individuals who occupy certain roles (e.g., leaders), inherit specific cultural traits (e.g., individualism), or have cognitive predispositions that set them apart from others (e.g., risk aversion). The notion of a personality type is intended to distinguish categories of people according to their way of processing information and reacting to the world.
Psychological characteristics may be hard-wired in a person’s genetic makeup. These features are often referred to as cognitive or innate. A lively research agenda within the social sciences focuses on cognitive features that influence attitudes, preferences, and behavior that we observe in the world today. Other psychological features are the product of a person’s culture, family, neighborhood, or some other aspect of their lived experience (perhaps in interaction with their genetic endowments). This is learned.

The nature/nurture distinction is an important one. However, we should be wary of treating it as a dichotomy. Research suggests that our genetic makeup is not entirely fixed at conception, and may evolve over time through lived experiences. In some respects, our nature is nurtured.

**STRUCTURAL FRAMEWORKS**

A second set of explanations center on structural features that condition the perceptions, motivations, and capacity of actors. Three structural frameworks are common in the work of social science: material factors, human capital/demography, and institutions.

**MATERIAL FACTORS**

Material factors enable and constrain what humans can do. They include financial resources (e.g., income or capital), geography, and technology. An explanation based on material factors may be described as materialist. While materialist structures often work in tandem with interest-based motivations, as discussed above, they are not identical, as will become clear.

Individuals with lots of financial resources have the capacity to attain many goals that are beyond the reach of those without such resources. Moreover, those with and without money may see things differently, and behave differently, by virtue of their social class position. The poor usually vote differently from the rich; they are more likely to support redistributive policies; and they are more likely to commit violent crimes. Social class matters, as researchers since Karl Marx have reminded us. The primordial question of social science, Marxists would argue, is who has what?

Material explanations may also hinge upon the distribution of a good within a population, i.e., who has more. Sometimes, the distribution of a good matters more than its absolute level. For example, happiness may depend more upon one’s standing relative to others than one’s actual (absolute) income. Here, income becomes a marker for success and thus determines one’s social status. It is not having a car that matters but rather whether others in the neighborhood also possess cars, or whether they possess fancier cars than you do.

Just as rich people are different from poor people, rich societies are different from poor societies. Work in the modernization paradigm may be understood as an exploration of all the ways in which the advance of material development affects society, culture, and politics. An intriguing (but also controversial) example is “mother love” – the bond of attachment between a mother and child. Nancy Scheper-Hughes argues that the strength of this bond is to some extent determined by infant mortality rates within a society. As rates fall and infants are expected to survive through the perilous period of childhood, parents begin to invest emotional energy in their children at an earlier age. If so, mother love – seemingly, the most natural emotion of all – is explained as a product of material conditions.

Geography provides another type of material explanation. If one lives in a mountainous or heavily forested area that is distant from the sea or a navigable river one is likely to be relatively

---

isolated. Over time, such isolation may prevent development and preserve long-established cultural practices. Marvin Harris’s book, *Cows, Pigs, Wars & Witches*, explores the role of geography and other material factors in the development of cultural practices around the world.\(^{72}\) Jared Diamond’s more recent synthesis, *Guns, Germs, and Steel*, explores the ramifications of geography for long-run development.\(^{73}\)

Technology is a third type of material factor, referring to tools that help produce or achieve an outcome. A generous reading of “tool” includes everything from the invention of writing to biogenetic compounds and prophylactics (birth control). As such, technology represents a key factor in the arts, in the economy, and in all aspects of human life. Some technological factors operate over long periods of time and in ways that are scarcely apparent to those of us who inherit their historical effects. Alberto Alesina, Paola Giuliano, & Nathan Nunn argue that “the descendants of societies that traditionally practiced plough agriculture today have lower rates of female participation in the workplace, in politics, and in entrepreneurial activities, as well as a greater prevalence of attitudes favoring gender inequality.”\(^{74}\) If the authors are correct, the timing of a technological advance in agriculture – namely, the plough – has an impact on gender roles many centuries later.

**HUMAN CAPITAL/DEMOGRAPHY**

*Human capital* and *demographic* explanations focus attention on individual attributes such as education, skills, health, migration, age, and life-cycle events such as birth, marriage, and death. These attributes may also be aggregated up to describe a population. For example, one can speak of the educational attainment of an individual or the (average) educational attainment of a population.

Education is generally acknowledged to be a strong predictor of success, at all levels. Individuals with more education earn more than individuals with less education. Businesses with a lot of smart, well-trained people are likely to out-perform businesses with weaker employees.

Countries with high human capital are thought to enjoy a significant advantage over countries with high illiteracy. Economists have also found that measures of health show a strong relationship to individual- and country-level achievement. Demographers have explored the various causes of high and low fertility across societies, and the impact of fertility on gender roles, infant survival, and cultural attributes such as individualism. The so-called **demographic transition** – a point in time when mortality rates fall and birth rates fall – is thought to have vast implications for society and is widely regarded as a key element of economic and social development.

It seems clear that human capital/demography explains a lot about human attitudes and behavior, at individual, group, and societal levels.

**INSTITUTIONS**

*Institutions* refer to humanly created rules and the organizations that establish and maintain those rules. Together, they create incentives that condition preferences and behavior – and may, over time, affect the construction of norms.

A *rule* is understood as formal, explicit, and (usually) written, and is connected to sanctions or rewards for good/bad behavior. For example, electoral law establishes the context for elections, affecting the behavior of candidates, parties, and voters. Property law establishes the context for the operation of markets, affecting the behavior of investors, producers, and consumers. Those who do not abide by these rules are subject to prosecution.

---

\(^{72}\) Harris (1974).  
\(^{73}\) Diamond (1992).  
\(^{74}\) Alesina, Giuliano, Nunn (2013).
Rules are enforced by organizations, which is why we incorporate both elements into our definition of an institution. An organization might be religious (e.g., the Catholic church), economic (e.g., a firm), political (e.g., a government), or social (e.g., a kin-group). In order to qualify as an organization, a group must have relatively clear boundaries, explicit criteria for membership, a set of rules, and officials who make and enforce the rules. A nation-state is an organization because it possesses all of these attributes. A nation without a state, by contrast, is less organizational. It is not clear for example who is a member of the Kurdish nation, how one gains or loses membership in this body, or who is responsible for making and enforcing the rules.

Our definition of institutions is a fairly narrow one, intended to help distinguish among rival explanatory frameworks. A broader definition of the subject would encompass many of the other explanatory frameworks explored in this chapter. (“Institutions” is one of the more malleable words in the social science lexicon.) However, for present purposes a narrow definition is more useful. However defined, social science work often focuses on the role of institutions – so much so that we are at a loss to provide specific examples for this section of the chapter, as they would encompass a majority of work in economics, political science, sociology, and related fields.

The popularity of institutions is not because institutions are necessarily more important than other explanatory factors. It is, rather, because institutions are the one explanatory factor that is solidly within our power – collectively – to change. Insofar as we want to improve society, we are likely to focus on institutional theories about how society works, for they are more relevant than factors over which we have little direct influence. When societies seek to change factors like resources, human capital, or motivational features of human existence, they do so by changing institutions that structure these outcomes. To change the distribution of resources we might propose a change in the tax code, for example. To improve the quality of human capital we might propose an increase in spending on educational programs. In this respect, institutional factors can claim primacy over other explanatory frameworks.

**INTERACTIVE FRAMEWORKS**

Sometimes, human attitudes and behavior is best explained as a dynamic interactive process rather than as the product of a single factor. Interactive frameworks include adaptation, coordination, diffusion, networks, and path dependence. Here, outcomes are understood as the product of interactions among people over time. Interactive factors thus combine elements already introduced as motivational or structural, building upon previous sections.

**ADAPTATION**

Adaptation focuses on the way in which individuals and groups adapt to a given environment. The notion is that, over time, the most efficient (fully adaptive) form will be achieved, with efficiency understood according to the preferences of the actors or by measuring success or survival. Likewise, the attitudes and behavior of individuals and groups within a society may be understood according to the roles they play, i.e., the functions they perform within that larger society. This functionalist view of social relations has a long history in anthropology and sociology stretching back to Emile Durkheim.

Sometimes, the adaptive (efficient) result is obtained through a process of feedback. When an individual pursues a strategy (e.g., to get elected or to sell a product) she receives feedback (from voters or from consumers). Based on this feedback she may decide to change her strategy.
Sometimes, the adaptive result is obtained through a process of competition. In a market economy, firms compete with one another for profits and market share. Competitive pressures—Adam Smith’s invisible hand—may prompt firms to lower costs, improve their product, or adjust their sales strategy. Likewise, in a political market, parties (and politicians) compete for votes. Competitive pressures may prompt parties to alter their platforms, change their behavior, or adopt different campaign strategies. At a national level, one might regard nation-states as engaged in a similar battle for survival, prompting adaptive changes in organization. The implicit selection mechanism at work here—with firms, parties, or states—is survival. Organizations that do not adapt may disappear. This explanatory trope is modeled on Charles Darwin’s theory of evolution, in which a process of natural selection rewards species that successfully adapt to their environment and punishes (by extinction) species that do not.

Sometimes, adaptation happens because individuals choose outcomes that benefit them. Here, adaptation maximizes utility. Efficient outcomes might take the form of a formal contract (e.g., between a buyer and a seller) or an informal understanding (e.g., between a patron and a client). One may also regard larger societal outcomes as efficient if they maximize the preferences of members in that society. For example, it might be argued that the spread of the nation-state throughout the world in the past several centuries is due to the fact that this particular form of political organization is the most efficient way to resolve conflict and provide public goods (though not everyone agrees with this premise). In any case, adaptive explanations presume a group—which may be as small as a family or as large as a country—for whom the arrangement is efficient.

**Coordination**

Coordination explanations focus on the failure to achieve efficient outcomes. Specifically, they focus on situations where the pursuit of individual preferences leads to suboptimal outcomes for those concerned. In this sense, all coordination problems may be described as collective action dilemmas.

One common coordination problem occurs when individuals find it in their interest to free-ride on others. For example, citizens may prefer not to pay taxes even though they benefit from the services that government provides. Likewise, states may prefer not to lower tariffs on foreign goods even though they benefit from global free trade.

Another sort of coordination failure is the prisoner’s dilemma—where the payoffs from a situation give each player an incentive to defect from a cooperative outcome that would benefit both parties. Consider the following stylized scenario. Two prisoners arrested for the same crime are placed in separate cells and offered the same choice by the district attorney: confess (and implicate the other defendant) or remain silent. The payoffs for each set of decisions (length of sentence in prison), as stipulated by the district attorney, are listed in Figure 5.1. Note that the optimal outcome for both prisoners—considered collectively—is to stay silent, which will mean they serve only a short, one-year term in prison. However, each prisoner is better off confessing than remaining silent, regardless of what the other prisoner decides. Defecting (confessing) is therefore the rational choice in a situation where prisoners are unable to coordinate with each other.
This hypothetical game has spawned a thousand others, each with modifications to the payoff structure, the number of players, the degree of communication allowed, the sort of rewards or punishments offered, and so forth. For our purposes, the key insight is that many social settings seem to exhibit coordination failures – where a suboptimal outcome is achieved because individuals (or groups) cannot successfully coordinate their activities.

A third sort of failure involves inter-temporal coordination failures such as those arising from the inability to make a credible commitment to a policy. Here, one party wishes to make a promise to another party but has no means of convincing the other party that the commitment will be upheld. An oft-cited example is the dictator who promises to respect property rights. In this scenario, investors have no assurance that the dictator will remain true to his/her promise in the future because a dictator – by definition – can change the rules whenever it suits him or her to do so.

A fourth sort of coordination failure (or impediment) involves transaction costs among individuals seeking to reach an agreement (or contract). These refer to any sort of barrier that increases the costs (on either party) of exchanging a good or service. This might involve information costs (relative to the substance of the agreement or the product being purchased), bargaining costs, or enforcement costs. While the transaction-cost framework was originally developed to understand the behavior of markets it has also been applied to politics, social institutions, and other venues.

Problems of coordination bedevil all social settings, and are especially valuable in explaining the formation of institutions and in designing new institutions. Indeed, for some writers the concept of an institution is inseparable from the coordination problem it is designed to solve. It is often argued, for example, that the primary function of government is to solve collective action problems, e.g., by internalizing costs and benefits and preventing free-riding. Likewise, some state structures may enjoy an advantage because they solve a particular coordination problem. For example, since democratically elected rulers are constrained by a constitution and an independent judiciary, as well as by an electorate (who can vote out the incumbent), democracies may be better situated to solve problems of credible commitment – assuring investors that their property will be respected in the future.

**Diffusion**

Some phenomena are so irresistible that their spread can be explained simply by exposure. This process is often labeled diffusion. It might apply to especially appealing ideas, new information that affects the way we think about a topic, or new technologies that promise great returns. Here, a
demonstration effect of prowess is sufficient. Since information expands knowledge and facilitates learning, knowledge- and learning-based explanations are integrated into this discussion.

Consider the choice between dropping out of school and remaining in school. This choice is presumably affected by one’s understanding of the payoffs to each course of action. One might know with a high degree of certainty that one can get a job and the wages that job would bring, and one might be highly uncertain about whether one could get a higher-paying job if one continued one’s education. Changes in this information – gained, say, by talking to recent graduates or reading a study following the fortunes of recent graduates – are likely to affect one’s decision. The same is true for other decisions, e.g., purchasing a good, casting a ballot, joining an insurrection, committing a crime, or procreating.

Not surprisingly, attempts to change behavior often take the form of information campaigns. To combat the spread of HIV/AIDS, people are informed about how the disease is spread, the consequences of contracting the disease, and methods of protecting themselves. To combat the use of tobacco and other drugs (especially among youth), people are informed of their highly addictive qualities and the consequences of long-term use. To promote a candidate, voters are informed of the candidate’s background, her achievements to date, and her proposed program of action, which may be directly contrasted with her opponents’ qualifications and policy commitments. Likewise, a promotional campaign for a new product may convey important information about that product.

Granted, information campaigns are generally not limited to information. There may also be a good deal of extraneous material, some normative appeals, and perhaps some downright falsehoods. Moreover, discerning information (truth) from lies is not always easy. This is why writers sometimes prefer to speak of ideas rather than information. In any case, the point remains that information (or ideas) about a subject often affects what an actor thinks and does.

The diffusion framework is so simple that it requires little elaboration. Suffice to say that causal explanations usually have a second focus – on the factors that encourage or discourage exposure. For example, diffusion of an idea might be enhanced by a network of communications, which in turn may be affected by communications technology (e.g., the availability of phone, email, or web-based connections). This leads to our next topic.

Networks

Wherever a stable pattern of communication among individuals (or groups) exists the resulting network may be mapped, i.e., drawn in a diagram or represented in a matrix. The network configuration may include additional information concerning the frequency of contact within a dyad, the duration of the event or the inter-connection, the substance of the communication, or attitudes towards others (inside and outside the network). Armed with this information, it may be possible to characterize ties as strong or weak and nodes in the network as central or peripheral. Figure 5.2 shows a typical social network (using data from Facebook), consisting of individuals and the ties between them.

The premise is that one can explain individual-level outcomes by examining their network position. For example, those with greater centrality may be more likely to receive information, to influence the behavior of others, or to fall victim of a communicable disease. At a system level, it may be possible to explain the relative speed or extent by which ideas, institutions, germs, technology, or some other factor of interest diffuses through a network.
Figure 5.2: A Facebook Social Network
Social capital explanations focus on the relative density of social networks. In a community where individuals have lots of interconnections that extend beyond their immediate family and workplace, there is high social capital. This sort of interaction is thought to foster social trust (and is in turn the product of social trust), political engagement, and employment opportunities. At an aggregate level, social capital is often seen as a key feature in economic development and good governance. Since these arguments were already introduced at some length in Chapter 1 we shall not expatiate further.

Path dependence

A critical juncture is an event that is in some respect stochastic (random), i.e., it could have been otherwise and cannot easily be explained by some other factor. Sometimes, an event of this sort has enduring consequences that flow in a structured way from the event. Here, we may refer to the resulting trajectory as path-dependent. This is the loose sense in which the terms critical juncture and path dependence are usually employed.

A narrower definition of path dependence refers to a situation in which an initial event is followed by a positive feedback loop (“increasing returns” or “lock-in”) that works to entrench the original event. Unfortunately, it is often difficult to tell when a weak or strong version of the framework applies. And many writers do not have this more restrictive condition in mind when they use the term. Thus, we adopt the looser, more encompassing definition.

An instance of critical juncture/path dependence (by any definition) is the location of the keys on a standard English keyboard, which spell out the sequence QWERTY. The reason for this arrangement is largely accidental. An early typewriter assumed this format and later keyboards followed suit, for the simple reason that consumers (it was assumed) would not wish to learn a new system. It is, however, not highly adaptive (see discussion above), as typing speed would increase if the keys were arranged in a different fashion (once people learned the new system). But we are unlikely ever to realize that new arrangement because of sunk costs, i.e., path dependence.

Institutions usually involve a degree of path dependence. Consider that every public policy involves a set of beneficiaries and a larger set of actors whose incentives are affected by the law. For example, the passage of the Social Security Act in 1935 established a new class of government beneficiaries (old people) and a larger class of individuals (including Americans of all ages) who base their retirement and investment decisions on the existence of this policy. Once a society has adjusted to this reality it becomes very difficult to change. Other policy solutions become less feasible.75

Large and abstract institutions are also heavily path-dependent. Once democracy has become established in a country people accustom themselves to this form of government. Leaders gain material stakes in the system. Everyone’s behavior and norms change. Consequently, democracy develops staying-power. This is one way to explain the fact that few countries revert to autocracy once they have experienced several decades of democratic rule.

Importantly, path-dependent explanations are not always efficient (in the sense of maximizing the long-run preferences or interests of all concerned). Sometimes, a low-equilibrium outcome will persist even though it is not serving anyone very well. An example would be the QWERTY keyboard. Some might argue that in the context of an aging society the shape of social security policies also qualifies as an example of a non-efficient outcome.

75 Derthick (1979).
BUILDING A THEORY

Having introduced a typology of explanatory frameworks, summarized in Table 5.1, we need to clarify several points that have been left dangling.

First, the typology is not intended to be exhaustive. Indeed, the notion of a comprehensive set of explanatory frameworks would presume a comprehensive typology of human behavior – an unlikely prospect.

Second, each of the frameworks listed in Table 5.1 is broadly defined so as to encompass a wide variety of specific theories. Readers should not suppose that there is only one sort of materialist framework or that writers invoking material explanations agree with one another about the way the world works. There is a good deal of diversity within each framework, as we have sought to demonstrate.

Third, most of the frameworks listed in Table 5.1 can be applied at the individual or group level. For example, a person may be subject to material constraints as well as a social group, organization, or an entire nation. A group, after all, is a collection of individuals. Likewise, different explanations may apply at different levels. For example, individual behavior may respond to material incentives while group behavior is best explained by institutions.

Fourth, the outcomes of a causal theory – the phenomena one is ultimately concerned to explain – may be variously defined. Specifically, a theory may be designed to explain attitudes (what people think or feel), preferences (how people prioritize across options), behavior (what people do), or conditions (how people live or die).

Fifth, any number of possible interrelationships may be envisioned among the various frameworks. For example, a materialist theory may be enlisted to explain an institutional outcome. In this case, one framework serves as the cause and another serves as the outcome.

Finally, these frameworks often play complementary roles. Indeed, a complete causal explanation for any social phenomenon generally incorporates more than one of the frameworks listed in Table 5.1. An explanation might enlist coordination logic to explain a set of institutions; these institutions might establish material incentives; and these incentives, in turn, may serve to motivate behavior. Here, diverse frameworks are woven together into a single narrative. So, causal frameworks are not always rivals. Sometimes, they are building blocks that combine to form a coherent causal explanation.

AN EXAMPLE: BIRTH

As an example of how different explanatory frameworks inter-weave let us consider the decision to have a child, which we have classified under the rubric of demography in Table 5.1. At group levels, this decision is typically measured as the total fertility rate, the average number of children born to each woman within a group or society during her childbearing years.

Some societies have high fertility rates. In sub-Saharan Africa, several countries have fertility rates of 7 or more. Advanced industrial societies generally have fertility rates of 2 or less, below the rate at which a population can sustain its current level (without in-migration). Fertility rates in Hong Kong and Macao are now at about 1 child per woman, an astounding fact in light of human history. Only a century ago, fertility rates in most countries were close to those found today in the developing world. Indeed, fertility is an area where enormous change has occurred over a relatively small period of time, with immense consequences for humankind and for the natural environment.

A range of theoretical frameworks have been enlisted to explain fertility. Interest-based explanations regard procreation as a choice based a calculation of costs and benefits. Fertility goes
down when the costs of bearing a child go up relative to the associated benefits. This, in turn, is
affected by the material situation in which adults find themselves, calling forth materialist
explanations. Urban dwellers are less likely to need the labor provided by their children than those
living in the countryside; likewise, space is scarcer and more expensive. Both of these factors alter
the calculus of childrearing as the process of urbanization moves forward. Likewise, social insurance
policies designed to care for adults in their old age lessen the need for children to serve as caretakers;
in this fashion, the long-term payoff provided by children is lowered. Finally, as child mortality rates
fall (as they have virtually everywhere in the world), the number of births required to achieve a given
family size also declines. Here, mortality alters fertility.

However, varying fertility rates around the world do not seem to be solely the product of
material considerations. Indeed, families are much larger in some parts of the world – especially in
sub-Saharan Africa – than would be explainable on the basis of a narrow conception of self-interest.
One possibility is that people are not correctly calculating the costs and benefits of having a child
under current conditions; instead, they may be following norms of behavior that were established
some time ago, under very different conditions (e.g., when child mortality was high). Perhaps norms
lag behind objective circumstances. If so, we may think of material factors as the ultimate cause and
norms as the proximate cause of behavior. If so, a change of norms may be required to affect
fertility rates in a society. Another possibility is that adults are not thinking about procreation in a
cost-benefit framework at all; choices about when to marry, when to engage in sex, and whether to
use birth control are instead based on deeply rooted cultural norms (including religious
prescriptions) about what is expected, i.e., a logic of appropriateness.

A third approach to fertility focuses on institutions. Public policies may prohibit or
encourage the use of contraception and abortion. Policies may explicitly discourage large families in
countries where the government wishes to depress the birth rate (e.g., the one-child policy in China)
or encourage larger families in countries where de-population is regarded as a problem (e.g., child
allotments, offered in many advanced industrial countries).

A fourth approach focuses on psychological factors affecting the decision to have
intercourse, utilize contraception, nurture a fetus to term, and seek appropriate medical assistance. A
fifth approach considers fertility from within an adaptive framework. In some settings, groups are
thought to engage in fertility competitions, in which only the most fertile groups survive. A sixth
approach might enlist a coordination framework. Arguably, high fertility is a product of situations
where the costs of a child are dispersed across a community while the benefits are centered on the
child’s biological parents. A communal setting of this sort establishes a collective action problem in
which individual incentives are not aligned with community interests.

Other explanatory frameworks can easily be imagined. The point is, for any given outcome
there are usually many possible explanations. Some may be truer than others; in this respect,
explanatory frameworks are rivals. But all may be true, in varying degrees, or at different levels. If so,
the various causal factors at work are likely to be interrelated in a complex fashion. Institutions may
affect norms, and norms may also affect institutions, as seems likely in the previous example.

**FORMAL MODELS**

Sometimes, a theory is stated in formal terms, with an explicit set of assumptions (axioms) upon
which a deductive structure is assembled. **Formal** theories are usually expressed in mathematical
language, where a single equation or set of equations (a model) embody the argument. Sometimes,
the model can be solved, i.e., it offers a single solution or prediction. Sometimes, the assumptions
must be narrowed in order to attain a solution. Insofar as it is possible to compute solutions for a
formal model using various assumptions (e.g., imposing specific parameter values on the variables), we refer to the exercise as a computational model and to the various computations as simulations.

Whether an explanation can be represented in formal, mathematical terms depends upon its simplicity. This is not just a matter of the number of moving parts but also the extent to which each moving part can be understood as a “quantity.” Some concepts, like institutionalization, are not very amenable to formal treatments.

Formal models often rest on preferences because they are relatively tractable (at least more so than norms and psychology). A preference refers to what an individual (or group) prefers across a range of options. Usually, it is assumed that preferences are complete, which is to say that for any two options the actor can decide whether option A is preferred to option B, option B is preferred to option A, or both are regarded as equal. In addition, it is usually assumed that preferences are transitive: if A is preferred to B, and B is preferred to C, then A is preferred to C. This allows preferences to be ordered, producing a utility function. When an individual’s behavior follows this logic, s/he is said to maximize his/her utility. (Additional assumptions are sometimes added; but comparability and transitivity are almost always included in a preference-based explanation.) Although this framework may seem complex – and indeed some formulations are extremely technical and involve a great many assumptions – the core idea is quite simple: people’s behavior is a product of their preferences.

Let us see how this framework might apply to vote choice. Consider an election conducted with single-member districts (a single candidate is selected from each electoral district) and winner-take-all (aka first-past-the-post) rules, where two parties (Democrats and Republicans) compete. We shall assume that voters’ preferences may be understood according to their issue-preferences (about particular policies) and that these policies occupy a unidimensional issue-space from left (liberal) to right (conservative). For illustrative purposes, we also assume that voters are evenly distributed along this spectrum (although this assumption can be relaxed to get the same result). And we shall assume that ideological positioning by the parties is unconstrained by activist members, by their history, by the problem of credible commitment, by changes to voter turnout, or by other factors. Under these circumstances, where should the parties (and their candidates) position themselves?

Early work on voting employed a simple spatial model, illustrated in Figure 5.3, in order to understand why candidates in a general election might move toward the middle – understood as the median voter (that individual whose views lie in the center of the ideological spectrum, with an equal number of persons to his/her left and right). Consider panel (a), which offers a plausible description of the positions of the parties according to their aggregated personal preferences – with the Democratic party on the left of the spectrum and the Republican party on the right. This setting is not in equilibrium, however, because either party can enhance its position by moving to the center, as the Democratic party does in panel (b). In this scenario, the Democrats win by a landslide simply by moving along the spectrum and thereby capturing more votes in the center. The assumption is that they lose no votes in doing so because voters to the left of the Democratic party have no viable option that lies closer to their ideal-point. Of course, the Republican party may also play this repositioning game. In panel (c), the Republicans move almost all the way to the Democrats, occupying a larger portion of the spectrum and assuring themselves of a landslide victory. This setting is also not in equilibrium because another move by the Democrats – leap-frogging the Republicans – would assure them of victory. Under the terms of our (admittedly simple) model, the only stable equilibrium is one where both parties hug the center, as illustrated in panel (d).

The intuition, then, is that if voters attempt to maximize their preferences (understood as policy preferences), they will vote for the candidate whose policy positions most closely mirror their own. In a two-party competition, parties (and candidates) maximize their vote-share by moving toward the center, generating a centripetal dynamic.
Figure 5.3: A Spatial Model of Vote Choice and Party Positioning

Panel (a)

Panel (b)

Panel (c)

Panel (d)
Here, issue-preferences structure candidate behavior, which in turn structure voting behavior. If most voters behave in this way we can use this simple folk wisdom to explain the results of an election.

Likewise, if we wish to explain why certain individuals or parties join together in a coalition, we might begin by examining their preferences, with the idea that those with similar utility functions are most likely to form enduring coalitions. The same logic may be employed to explain consumer choice, the structure of markets, the choice of marriage partners, social networks, and many other phenomena of interest to social science.

Of course, inferring what a person’s true preferences are is not always an easy matter. One may ask directly (via a survey or interview). But people are not always forthcoming. Sometimes, they tell you what they think you want to hear or what they feel is most acceptable. They will rarely reveal preferences that go against the law or against social mores (see section on Surveys in Chapter 13).

One may also observe an individual’s behavior. Typically, preferences predict behavior. If we observe how someone votes we are presumably observing their preferences at work. This approach to measurement is referred to as revealed preferences. Sometimes, however, people act strategically. Consider an election contest where three candidates, A, B, and C, compete for a single seat. Let us suppose that Bob’s preference-ordering is A>B>C. However, he knows (from opinion polls) that A’s probability of winning is miniscule. Under these circumstances, Bob might opt for B even though B is not his first preference. On the basis of Bob’s behavior, and knowing something about the situation, one can infer that Bob prefers B to C. But it is impossible to tell whether he prefers A to B.

Where asking and observation are impossible or for some reason flawed, one may try to infer preferences from other aspects of a person’s situation. This approach may be reasonable if our focus is on the preferences of many people rather than on a single individual. In this case, it does not matter if our guesses about preferences are correct or incorrect for “Bob” so long as they are generally correct for members of Bob’s group.

If one is trying to understand a series of decisions (taken by an individual or a group) in a particular sequence, a decision-tree model may be applied. This draws on a branch of math and economics known as decision theory. If actions by an individual or group depend upon the actions of others, the resulting model may be described as game theory. Here, each player is understood as playing a game insofar as their actions are strategic, with an eye to the probable behavior of others.

**CONCLUSIONS**

In this chapter, we have set forth a number of explanatory frameworks, summarized in Table 5.1. Although this typology is by no means comprehensive, it encompasses many of the arguments you are likely to run across in social science studies. The final section of the chapter discusses how frameworks interact with one another in the service of causal explanation. It also discusses – in a very general way – how and why explanations might be formalized in a mathematical model. This may prove helpful as you peruse the literature or construct your own arguments about a topic.

Naturally, every model of human behavior requires simplifying assumptions. And mathematical models are generally more reductionist than models elucidated in prose. On this ground, they may be criticized (for leaving out aspects that may be important for understanding an outcome) as well as praised (for their parsimony). This is the irreducible dilemma faced by any attempt to theorize about complex phenomena connected to human behavior.
The utility of any causal explanation – formal or non-formal, mathematical or non-mathematical – is contingent on the extent to which its assumptions are realistic and its predictions are testable. This brings us to the topic of the next chapter.
KEY TERMS

- Motivational frameworks
- Interests (that which benefits an actor)
- Norms (values, beliefs, ideology, culture, identity)
- Psychology (cognition, emotion, personality, genetics)
- Structural frameworks
- Material factors (financial resources, modernization, geography, technology)
- Human capital/Demography (health, education, skills, migration, life-cycle events, demographic transition)
- Institutions (formal rules and organizations)
- Interactive frameworks
- Adaptation (feedback, competition, selection, evolution, efficiency, functionalism)
- Coordination (collective action, externalities, credible commitment, transaction costs)
- Diffusion (demonstration effect, exposure, ideas, information, knowledge, exposure)
- Networks (informal associational structures, social capital)
- Path dependence (contingent moments [critical junctures] leading to fixed trajectories)
- Critical juncture
- Simulation
- Formal models
- Preferences
- Utility function
- Revealed preferences
- Decision tree
- Game theory

INQUIRIES

1. Take a published article of your choosing (perhaps drawn randomly from a journal in your subfield of interest). Or, alternatively, consider your own work. Now try to classify the main argument into one or more of the categories listed in Table 5.1. Does it fit neatly? Does it fit at all? If not, why not? How might the typology be reformulated to include your argument?

2. Are there important social science explanations (theories) that can't be classified within the typology of Table 5.1? Think about the field or subfield that you know best.

3. Consider an outcome of interest in your field or subfield. Which theoretical frameworks are generally invoked to explain that outcome? Is this theoretical choice understandable (justifiable)? How might other theoretical frameworks be applied? See if you can generate plausible explanations that apply a variety of frameworks, as listed in Table 5.1 (as we did for fertility in the final section of the chapter).
4. Consider democratization, an outcome we have used as an example throughout this book. How might various theoretical frameworks (listed in Table 5.1) be applied to explain this outcome? Which theoretical frameworks are most/least satisfactory?

5. Consider the act of voting (casting a ballot in an election). How might various theoretical frameworks (listed in Table 5.1) be applied to explain this outcome? Which theoretical frameworks are most/least satisfactory?

6. What are the generic strengths and weaknesses of each explanatory framework listed in Table 5.1? Are there some that seem more “scientific” than others? (On what grounds?) Which ones appeal to you? (On what grounds?)

7. Social scientists often say that the influence of norms is harder to study (causally) than the influence of other causes. This may account for why norm-based explanations are less common in social science than in the humanities and in popular accounts. Why might this be so? (Note: This question may be easier to answer after you have read Chapters 6-9.)

8. Path dependence explanations are sometimes criticized for being idiosyncratic, insofar as they apply to the specific circumstances of a specific case. Is this true? Is it impossible to generalize a path-dependent explanation?

9. Interest-based explanations are discussed first in this chapter because they provide what might be considered the first-resort explanation for social behavior. That is, when we try to explain why a person voted for one candidate rather than another, or why they chose one product over another, the first explanation we are likely to consider is that it is in the person’s self-interest to do so. If we find a compelling interest-based explanation we are likely to feel that we have satisfactorily explained the phenomenon. If not (that is, if self-interest explanations do not work, or are indeterminate), we may be more inclined to look to other explanatory frameworks. Is this a good way to proceed?

**RESOURCES**

*Note:* The long list of books and articles that follows serves a heuristic purpose, namely to illustrate the range of each explanatory framework discussed in the text. Here, readers can see examples – both classical and contemporary – of the each framework as applied to different disciplines. We do not suppose that you, or anyone else, will read all of these works, so do not feel intimidated by its length. However, if you are especially interested in one of these frameworks this reading list will offer a useful point of departure.

**CAUSAL FRAMEWORKS, GENERALLY CONSIDERED**


INTERESTS (THAT WHICH BENEFITS AN ACTOR)


NORMS (VALUES, BELIEFS, IDEOLOGY, CULTURE, IDENTITY)


**INFORMATION (IDEAS, KNOWLEDGE, LEARNING)**

• See also sources listed under **Diffusion**

**PSYCHOLOGY (COGNITION, EMOTION, PERSONALITY, GENETICS)**

MATERIAL FACTORS (RESOURCES, MODERNIZATION, GEOGRAPHY, TECHNOLOGY)


INSTITUTIONS (FORMAL RULES AND ORGANIZATIONS)


See also sources listed under Coordination.

HUMAN CAPITAL/DEMOGRAPHY (HEALTH, EDUCATION, SKILLS, MIGRATION, LIFE-CYCLE EVENTS)


DIFFUSION (DEMONSTRATION EFFECTS, EXPOSURE)


• Spolaore, Enrico; Romain Wacziarg. 2009. “The Diffusion of Development.” Quarterly
Journal of Economics 124:2, 469-529.


**Adaptation (Feedback, Competition, Selection, Evolution, Efficiency, Functionalism)**


**Coordination (Collective Action, Externalities, Transaction Costs)**


• See also sources listed under *Institutions*.

**NETWORKS (ASSOCIATIONAL STRUCTURES, SOCIAL CAPITAL)**


**PATH-DEPENDENCE (CONTINGENT MOMENTS LEADING TO FIXED TRAJECTORIES)**


**SIMULATIONS**


**FORMAL MODELS**


**Examples:**


6. CAUSAL HYPOTHESES AND ANALYSES

In the previous chapter we discussed general causal frameworks, the building blocks of a causal explanation. In this chapter, we focus on specific hypotheses, where one factor is thought to generate change in another factor.

We begin by clarifying the concept of causality. In the next section, we discuss the criteria of a good (well-constructed) causal hypothesis. The rest of the chapter is devoted to causal analysis. First, we outline the criteria that all causal research designs seek to achieve. Next, we discuss the problem of reaching causal inference.

This chapter is fairly complex. A number of new terms are introduced, some of which may be unfamiliar to the reader and some of which are used in slightly different ways in different disciplines. Although the vocabulary may seem bewildering at first, try to familiarize yourself with these concepts — which you are likely to encounter in your reading and in your future work. The topics covered here are critical for understanding how evidence is used to infer causality in social-science settings. Whether you are primarily a consumer or a producer of social science the following chapters bear a close read and a good think.

CAUSALITY

A causal hypothesis involves at least two elements: a cause and an outcome. A cause may be referred to variously as a condition, covariate, ecogenous variable, explanatory variable, explanans, independent variable, input, intervention, parent, predictor, right-side variable, treatment, or simply “X.” An outcome may be referred to as a dependent variable, descendant, effect, endogenous variable, explanandum, left-side variable, output, response, or “Y.”

Whatever the terminology, to say that a factor, X, is a cause of an outcome, Y, is to say that a change in X generates a change in Y relative to what Y would otherwise be (the counterfactual condition), given certain background conditions (ceteris paribus assumptions) and scope-conditions (the population of the inference).

Now, let’s unpack things a bit. As an example, we shall focus on the causal role of a worker training program. A reasonable hypothesis is that participation in the program (X) will enhance an unemployed person’s subsequent earnings (Y). If the relationship is causal, her earnings should be higher than they would be if she had never participated in the program. Let us represent the treatment, X, as a binary (dichotomous) variable, which takes one of two values. X=1 means that the individual has attended the program (the treatment condition). X=0 means that the individual did not attend the program (the control condition).

We shall refer to participants in the program as members of the treatment group since they receive the treatment (the worker-training program). Unemployed persons who did not participate in the training program – if they are similar to those who did – may be understood as a control group.

A causal effect (aka treatment effect) is the change in Y (ΔY) or “delta Y”) brought about by a change in X (ΔX). When X changes, Y must also change, at least probabilistically, if the relationship is causal.
When a treatment is binary, as above, $\Delta X$ refers to a change from 0 to 1, i.e., from the control condition to the treatment condition (attending the worker-training program). It might also refer to a change from 1 to 0, though in this case it is difficult to imagine un-learning what one has been exposed to in a worker-training program.

When an outcome is continuous, $\Delta X$ affects the value of $Y$ along an interval scale. Thus, we hypothesized that exposure to a worker training program increases one’s earnings. This hypothesis may be written as a statement of inequality…

$Y_{X=1} > Y_{X=0}$

Read: “Earnings are higher for someone in the treatment group ($X=1$) than for someone in the control group ($X=0$).” Or, alternatively, “Earnings are higher for someone after they have been treated ($X=1$) than before ($X=0$).”

When an outcome is also binary, $\Delta X$ affects the probability ($P$) of $Y$ achieving an outcome. Thus, one might hypothesize that exposure to a worker training program increases the probability of employment ($Y=1$), as opposed to unemployment ($Y=0$). This may be written as

$P(Y=1|X=1) > P(Y=1|X=0)$

Read: “The probability of employment is greater conditional on having attended a worker-training program than conditional on not having attended such a program.”

Estimating a causal effect involves comparing a factual (that which actually happened) with a counterfactual (that which might have happened). Continuing with our example, let us say that we wish to determine what effect (if any) a job-training program has on the earnings of unemployed people who have completed the program. We know what the program participants – the treatment group – earned before they entered the program and after they finished the program. But we do not know what they would have earned had they never attended the program. Unfortunately, this is precisely what we need to know if we are to estimate a causal effect. That is, we want to know whether members of the treatment group have higher earnings than they otherwise would have (had they never received the treatment).

Because causal effects depend upon counterfactuals that are not directly measurable they seem to require something on the order of a time-machine. Traveling back in time, one could re-play events so as to observe the relevant counterfactuals. Thus, one could cancel the worker-training program and observe the earnings of those who would have participated in it, leaving everything else in the world as it is (ceteris paribus). In this fashion, one could obtain a measure of the true causal effect.

Lacking time-machines, we must resort to cruder expedients. Specifically, we must infer the unobservable counterfactuals – what-would-have-beens – from available evidence. For example, one might obtain an estimate of the causal effect by subtracting the earnings of members of the treatment group before completing the worker-training program ($t$) from their earnings after participating in the program ($t+1$), expressed as…

$Y_{t+1} - Y_t$

In this setting, one must be willing to believe that background conditions remained constant from $t$ to $t+1$ – a difficult assumption, in many circumstances. One might also subtract the earnings of those in the treatment group from those in a suitable control group in order to get an estimate of the causal effect, expressed as…
Here, one must be willing to believe that members of the treatment and control groups are similar in all respects (relevant to their earnings) – also a difficult assumption, in most circumstances.

Causal inference is a tricky business, and you need to know the many tricks of the trade if you are to be successful at it. In subsequent chapters we will introduce a variety of research designs whose purpose is to arrive at a valid – and, if possible, reasonably precise – estimate of $X$’s impact on $Y$.

It bears emphasis that whenever one asserts that $X$ causes $Y$ one is asserting that the actual probability of an event is increased by $X$, not simply a theory’s predictive capacity. This is what distinguishes causality from description or prediction. To be causal, the factor in question must generate, create, or produce an effect. A correlation between $X$ and $Y$ does not necessarily indicate that a causal relationship exists; it may be spurious.

Of course, it is not always possible to specify precisely why $X$ generates a change in $Y$. Yet, in identifying $X$ as a cause of $Y$ one is presuming the existence of some causal mechanism – understood here as the pathway or process or chain of intermediary variables by which $X$ affects $Y$, illustrated as $M$ in Figure 6.1. In our example, the mechanism might be the factual information about a particular occupation that is imparted in the worker training program. Or it could be the job search skills that one learns as part of a program (such as how to construct a resume or how to present oneself in an interview). Or it could be general workplace comportment (how to relate to one’s boss and co-workers). Lots of potential mechanisms might be identified, and any combination of them may be at work.

Causal relationships occur against a background of other factors that also influence the outcome. In our example, many additional factors may affect earnings including experience, effort, and various demographic characteristics of the worker (race, sex, age, and so forth). These background conditions are generally assumed to hold constant. When one asserts that a particular causal factor, e.g., worker-training, increases earnings one is not asserting that other factors do not matter. One is asserting that participation in worker-training programs increase earnings, all other things being equal. This is known as the ceteris paribus assumption, and is implicit in all causal arguments.

Causal relationships are always bounded in some fashion; they do not exist everywhere and always. In this case, one would not expect worker training programs to have much causal impact in a country where jobs (and promotions) are allocated by clientelistic networks. Likewise, one would not expect a worker training program to affect earnings in a collapsing economy. So we might stipulate that the scope-conditions of the argument include an intact economy and meritocratic (or at least somewhat meritocratic) appointment and promotion practices. The scope-conditions of an argument may be understood as the box that surrounds all the nodes in Figure 6.1.

One sort of background factor lies antecedent to the causal factor of interest. It affects $Y$ indirectly, through $X$. In our example, the antecedent factor is that which causes a person to participate (or not) in a worker training program, labeled $Z$ in Figure 6.1.

In any causal argument, it is important to distinguish between factors that are independent or exogeneous (causal in nature), and factors that are dependent or endogenous (outcomes). In Figure 6.1, the relationships are as follows:

- $Z$ is exogenous to $X$, $M$, and $Y$
- $X$ is exogenous to $M$ and $Y$
- $M$ is exogenous to $Y$
Likewise,
- $Y$ is endogenous to $Z$, $X$, and $M$
- $M$ is endogenous to $Z$ and $X$
- $X$ is endogenous to $Z$

As you can see, endogeneity/exogeneity is a relative matter.

Each factor in Figure 6.1 may be treated as an individual *variable* or as a *vector* (set) of variables. Although most of our examples will concern individual variables these examples can usually be generalized by treating the variable as a vector. Often, we are interested in combinations of causal factors, combinations of causal pathways, combinations of antecedent causes, and so forth (very rarely, we might be interested in combinations of outcomes).

By employing the language of “variables” our assumption is that each factor can vary (at least hypothetically) in the setting under study. We are not supposing that all variables are measurable. Nor are we supposing that all causal models are statistical in nature. Nothing in the foregoing passages necessitates a large sample or a probabilistic model. We find the language of variables to be a convenient means of simplifying and unifying our understanding of causation. It is not meant to impose a uniform approach to research design.

**Figure 6.1: A Simple Causal Graph**

\[
\begin{array}{c}
Z \rightarrow X \rightarrow M \rightarrow Y
\end{array}
\]

*Z* = Antecedent  
*X* = Causal factor of interest  
*M* = Mechanism  
*Y* = Outcome of interest
Causal Graphs

When attempting to tease apart correlational and causal patterns, and devise appropriate conditioning strategies, it is often helpful to diagram the expected relationship between two or more factors. The resulting diagram—an example of which is shown in Figure 6.1—is known as a causal graph, a language of causal inference developed by Judea Pearl and others.

The notation used in this book does not exactly replicate Pearl’s, but it is similar. Here, each node in a graph represents a variable (or factor). A one-sided arrow between nodes indicates a causal relationship from the origin to the terminus:

\[ X \rightarrow Y \] (X causes Y)

A two-sided arrow indicates a reciprocal causal relationship:

\[ X \leftrightarrow Y \] (X causes Y and Y causes X)

A line without an arrow indicates a covariational relationship that may or may not be causal:

\[ X \sim Y \] (X co-varies with Y)

Note that when a node appears in a causal graph you may assume that it is measured (and “conditioned”), unless it appears in square brackets—in which case it is un-measured (and perhaps unmeasurable). We make frequent use of these graphs in the coming chapters so it is important to have this language clear in your head.

Criteria of a Causal Hypothesis

Having defined causation minimally, and discussed various types of causal relationships, we turn to the ideal-type. What is a good causal hypothesis?

Recall from Chapter 2 that all arguments strive for precision, generality, boundedness, parsimony, coherence, commensurability, innovation, and relevance. Causal hypotheses also strive for clarity, manipulability, precedence, impact, and an identifiable mechanism. For convenience, all thirteen criteria applicable to causal theories are reproduced in Table 6.2. However, our focus in this chapter is on factors that are distinct to causal arguments.
Table 6.2: Causal Hypotheses: Criteria

<table>
<thead>
<tr>
<th><strong>ALL ARGUMENTS</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Precision</strong> (specificity)</td>
<td></td>
</tr>
<tr>
<td>• Is it precise?</td>
<td></td>
</tr>
<tr>
<td><strong>Generality</strong> (breadth, domain, population, range, scope)</td>
<td></td>
</tr>
<tr>
<td>• How broad is the scope?</td>
<td></td>
</tr>
<tr>
<td><strong>Boundedness</strong> (population, scope-conditions)</td>
<td></td>
</tr>
<tr>
<td>• How well-bounded is it?</td>
<td></td>
</tr>
<tr>
<td><strong>Parsimony</strong> (concision, economy, Occam’s razor, reduction, simplicity)</td>
<td></td>
</tr>
<tr>
<td>• How parsimonious is it? How many assumptions are required?</td>
<td></td>
</tr>
<tr>
<td><strong>Coherence</strong> (clarity, consistency; <em>antonym</em>: ambiguity)</td>
<td></td>
</tr>
<tr>
<td>• How coherent is it?</td>
<td></td>
</tr>
<tr>
<td><strong>Commensurability</strong> (consilience, harmony, logical economy, theoretical utility; <em>antonym</em>: adhocery)</td>
<td></td>
</tr>
<tr>
<td>• How well does it cumulate with other inferences? Does it advance logical economy in a field?</td>
<td></td>
</tr>
<tr>
<td><strong>Innovation</strong> (novelty)</td>
<td></td>
</tr>
<tr>
<td>• How new is it?</td>
<td></td>
</tr>
<tr>
<td><strong>Relevance</strong> (everyday importance, significance)</td>
<td></td>
</tr>
<tr>
<td>• How relevant is it to issues of concern to citizens and policymakers?</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>CAUSAL HYPOTHESES</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Clarity</strong> (<em>antonym</em>: ambiguity)</td>
<td></td>
</tr>
<tr>
<td>• Is it clear what the envisioned variation on X and Y, background conditions, and scope-conditions are?</td>
<td></td>
</tr>
<tr>
<td><strong>Manipulability</strong></td>
<td></td>
</tr>
<tr>
<td>• Is the causal factor manipulable (at least potentially)?</td>
<td></td>
</tr>
<tr>
<td><strong>Precedence</strong> (exogenous, foundational, independent, original, prime, prior, structural, unmoved mover)</td>
<td></td>
</tr>
<tr>
<td>• Is X separable from Y and prior to other causes of Y?</td>
<td></td>
</tr>
<tr>
<td><strong>Impact</strong> (effect size, magnitude, power, significance, strength)</td>
<td></td>
</tr>
<tr>
<td>• How much of the variation in Y can X explain? Is the causal effect significant (in theoretical or policy terms)?</td>
<td></td>
</tr>
<tr>
<td><strong>Mechanism</strong> (intermediary, mediator, pathway, process)</td>
<td></td>
</tr>
<tr>
<td>• How does X generate Y? What are the causal mechanisms (M)?</td>
<td></td>
</tr>
</tbody>
</table>
CLARITY

We have defined causality as a situation in which a change in X (the causal factor of theoretical interest) generates a change in Y (the outcome of interest) relative to what Y otherwise would be, given certain background conditions and scope conditions. It follows that a good causal hypothesis should provide clarity along each of these dimensions. Clarifying causal theories makes them more useful, as well as easier to test. Indeed, a theory that is ambiguous is impossible to verify or falsify; it is neither true nor false.

One issue is terminological. Writers may state that a causal factor, X, leads to an outcome, Y, is related to Y, is associated with Y, influences Y, results in Y, and so forth. Of these, only the last two are clearly causal in the sense in which we have defined the term. But all may be causal, depending upon the context. A simple suggestion for writers is to clarify whether an argument is intended to be causal or not. Intuition on the part of the reader should not be required.

A second issue is the specification of Y. To evaluate an argument we need to know the variation in Y that is the outcome of interest. Usually, this is apparent; but sometimes it remains ambiguous. There is a humorous tale of a priest who reportedly queried the notorious bank robber, Willie Sutton, about why he robbed banks. The miscreant patiently explained that this is where the money is. Evidently, the priest and the bank robber have different ideas about variation in Y. For the priest, it is robbing (Y=1) versus not robbing (Y=0). For Sutton, it is robbing banks (Y=1) versus robbing other establishments (Y=0).

A third issue is the specification of X, i.e., the change in X that is envisioned as a causal factor or treatment. If someone argues that worker-training programs cause wages to rise, the variation in X is presumably participation (X=1) versus non-participation (X=0) in a worker-training program. The latter may also be referred to as the causal counterfactual.

A fourth issue concerns the background conditions or scope-conditions of an argument. Under normal circumstances, it is not necessary to specify what these are. Sometimes, however, the background conditions of an argument are important enough, and ambiguous enough, that they really ought to be mentioned. For example, it may be important to specify that the causal role of worker-training programs in wages presumes a capitalist economy; no such causal effect can be expected in a socialist economy. It may be important to specify that the effect of worker-training programs is limited to workers who are currently unemployed, and does not extend (or may not extend) to those already employed. Further discussion of appropriate scope-conditions is found under the rubric of boundedness (Chapter 3).

MANIPULABILITY

In order to formulate a clear causal hypothesis it is helpful if the cause in question is manipulable, that is, if it can be changed by the researcher or by someone else. Formulating the argument in this fashion – as an imagined manipulation – clarifies what the relevant change in X is, and what the background conditions might be.

As an example, let us return to the worker-training program. This treatment is fairly easy to manipulate because one can determine who is admitted into a program and who is rejected. Thus, when we say that a worker-training program has a particular causal effect we know what we are talking about.

By contrast, when someone argues that democracy has an impact on peace/war it is less clear what we have in mind. Naturally, it is helpful if we can specify how we are measuring “democracy.” But this is not simply a problem of measurement. The problem is that it is difficult to envision circumstances in which a country democratizes while holding other relevant factors
constant. Bear in mind that democratization entails a momentous shift in a country’s political and social life and is apt to be accompanied by many additional factors that might impact the probability of peace and war. Consequently, it is not clear what the background (ceteris paribus) conditions of such an argument might be.

Similarly, to assert that “social capital causes economic growth” is ambiguous. We cannot directly manipulate the level of social capital in a community. Consequently, it is difficult to know what “increasing social capital” might mean in a specific setting.

Any causal argument that rests on attitudes, beliefs, or diffuse practices – including ideational arguments and cultural/ideological arguments – is also somewhat ambiguous, for we have no way of directly manipulating people’s attitudes, beliefs, or diffuse practices. Of course, we can imagine manipulable treatments that would affect people’s attitudes, beliefs, and practices. We might, for instance, ask subjects to read a passage, watch a video clip, listen to a tune, or participate in a staged intervention. These treatments might make the subject happy or sad, change their opinion on a subject, or prompt them to take some action. And they are readily interpretable as causes. However, in this scenario the causal factor is the manipulable treatment, while attitudes/beliefs/practices serve an intermediary role (as a presumed causal mechanism).

Some methodologists view manipulability as a necessary condition of any causal argument. By this interpretation, arguments about the effect of democracy, inequality, race, and other abstract factors are not really causal in nature. This seems a little extreme. Instead, we will treat manipulability as a desirable trait, among others, and one that is best approached as a matter of degrees. Even non-manipulable causes – like attitudes, beliefs, and practices – have prior causes that are manipulable. Thinking carefully about these prior causes can sometimes help us to clarify an argument.

**Precedence**

A cause must be separable from the effect it purports to explain; otherwise it is tautological (circular). This seems obvious. Yet, on closer reflection, it will be seen that separation is a matter of degrees. To begin with, Xs and Ys are always somewhat differentiated from one another. A perfect tautology (e.g., “The Civil War was caused by the Civil War”) is simply nonsense. One occasionally hears the following sort of argument: “The Civil War was caused by the attack of the South against Fort Sumter.” This is more satisfactory. Even so, it is not likely to strike readers as a particularly acute explanation. Indeed, there is very little explanation occurring here, precisely because the X is barely differentiated from the Y (the attack against Fort Sumter was of course part of the Civil War). Equally problematic is an argument that links the Civil War to a warlike relationship between North and South, one that persisted from the 1850s to the outbreak of the conflict in 1861. Again, one is at pains to distinguish between cause and effect.

Generally, the greater the separation between cause and effect – the more distance lies between X and Y – the more useful the argument will be, and the more we are likely to regard X as a cause of Y. A good causal factor enjoys temporal precedence relative to other potential causes of an outcome.

Consider, if one proposed cause of an outcome is fully explained by something else, we shall regard the first as superstructural (epiphenomenal, endogenous) and the second as structural (exogenous, foundational). The second has better claim to the status of “cause.” It embodies Aristotle’s quest for an Unmoved Mover, a factor that affects other things but is affected by nothing.

---

Of course, every causal factor is affected by something. There are no unmoved movers. Yet, some factors are entirely (or almost entirely) explained by one or two other factors. Here, we are dubious about calling it a cause for it is mostly endogenous. By contrast, the causes of other causes are random or so numerous that they would be difficult to verify.

Consider Figure 6.1. If $X$ is largely explained by $Z$ (if most of the variation in $X$ is due to variation in $Z$), then $Z$ is probably more correctly regarded as “the” cause of $Y$. $X$ is subsumed by $Z$. Once one knows the status of $Z$ one can predict the status of $X$, $M$, and $Y$. $X$ and $M$ are not informative – except with respect to causal mechanisms, as discussed below.

If, on the other hand, $Z$ explains only a small portion of $X$ – which is a product of many factors, some of which may be purely stochastic – then $X$ may properly be regarded as the cause of $Y$. It is not subsumed by $Z$.

The greater its precedence, the higher its standing among the various causes of some outcome. Indeed, debates about causal questions often rest on which causal factor is properly judged to come first. Which $X$ explains all the other $X$’s?

Consider the various factors that have been proposed as explanations of long-term economic development, i.e., for explaining why some nations are rich and others poor. A short list of such causal factors would include geography, colonialism, domestic political institutions, technology, human capital, culture, population, and demographic transitions. Note that arguments among partisans of these different schools are not simply about whether a single factor – say, demography – has a large impact on long-term economic development. They are also, perhaps more importantly, about relationships among the various causal factors, namely which are independent and which are dependent. In this argument, geography has an important advantage: it is not easily explained. Indeed, geography approximates Aristotle’s unmoved mover. Of course, there are geological explanations for why land masses were formed in certain ways, why rivers appear, why some are navigable and others are not, and so forth. However, these explanations would be quite complex and would involve a considerable amount of contingency. Geographic explanations would be difficult to explain away. By contrast, cultural explanations seem quite vulnerable, as they are often endogenous to other factors. Those who wish to restore the status of cultural explanation must show that a set of values and practices that impacted economic development is not superstructural, that it has causal independence in the long sweep of history.

**Impact**

Causal arguments strive to explain variation in an outcome. The more variation the causal factor explains – the greater the **impact** of $X$ on $Y$ – the more significant that argument is likely to be. This may also be articulated to as a question of effect size, magnitude, power, or strength.

Necessary-and-sufficient causal arguments are compelling because they explain **all** the variation in $Y$, while remaining admirably concise. It is no wonder that they continue to be regarded as the ideal-type causal argument. By contrast, wherever there are exceptions to a causal argument, or where some factor other than $X$ accounts for variation in $Y$, we can see that the argument is weakened: it no longer suffices to account for $Y$.

There are a number of ways in which the question of relative impact can be gauged. In a regression format, where the relationship between $X$ and $Y$ is assumed to be probabilistic, impact is measured by the coefficient (slope) for $X$ or by a model-fit statistic such as $R^2$ for $X$, a vector of independent variables.

Of course, estimates of causal impact from an empirical model depend upon the specifics of that sample and model, and may or may not correspond to real-world impact. If the model is not realistic in this respect, then a separate evaluation of impact – perhaps in a more speculative mode –
may be required. It is often helpful to consider the impact of $X$ on $Y$ in practical terms, e.g., as a matter of public policy. Could a significant change in $Y$ be achieved by manipulating $X$? At what cost and with what opportunity costs?

The impact of $X$ on $Y$ may also be gauged by comparing its impact to other factors. If the impact of these other factors is well understood, this may provide a useful metric of significance (i.e., relative impact).

Whatever the metric of evaluation, the impact of $X$ on $Y$ is a key measure of success. One of the criteria of a good causal argument is that it explains a lot about the phenomenon of interest. It is not trivial.

**MECHANISM**

We have said that causes generate – alter, change, condition, create, effect – outcomes. It follows that there must be a causal mechanism, or mechanisms, at work. The mechanism is “the agency or means by which an effect is produced or a purpose is accomplished.” It is the causal pathway (aka process, mediator, or intermediate variable) by which a causal factor of theoretical interest is thought to affect an outcome, illustrated by $M$ in Figure 6.1. This might be a single factor, a large number of factors acting independently, a causal chain of discrete events, a continuous process (e.g., a billiard ball rolling across a table and hitting another ball) – in short, anything that is affected by $X$ and that, in turn, affects $Y$.

Sometimes, the working of a causal mechanism is obvious and can be intuited from what we know about the world. For example, if product sales increase when the price is lowered we may infer that consumer decisions were influenced by price. Often, however, the mechanism operating within a causal relationship is obscure. For example, insofar as one might regard economic development as a causal factor ($X$) in democratization ($Y$) – rich countries being more likely to democratize or to consolidate their democracy – it is difficult to say which components of economic development might be causing this result. $X$ might impact $Y$ through income, infrastructure, urbanization, or education, and the mechanism might be interest-based or norm-based (or some combination of the two).

An $X/Y$ hypothesis without a clear causal mechanism is an argument in search of an explanation. It may be true, but it will be not be very meaningful, will be difficult to generalize upon, and may also be difficult to prove in a convincing fashion. We want to know why $X$ causes $Y$, not simply that $X$ causes $Y$.

Thus, it is incumbent upon the writer to clarify the causal mechanism(s) at work in a causal argument if it cannot be intuited from context. This may be accomplished in prose, in diagrams, and/or in mathematical models, and is implicit in the very act of theorizing, as discussed in the previous chapter.

**CAUSAL ANALYSIS**

Having established the nature of causal hypotheses, we turn to the problem of proof or demonstration. Given a specific hypothesis about the relationship of $X$ to $Y$, how might one assess whether the relationship is actually causal (rather than merely correlative)? If it is causal, how might we estimate the causal effect?

---

In Chapter 4, we observed that all research designs strive for validity, precision, sample representativeness, and sample size. In this chapter, we focus on two additional criteria that are specific to causal analysis: covariation and comparability. For convenience, all six criteria are reproduced in Table 6.3.

### Table 6.3: Causal Research Designs: Criteria

<table>
<thead>
<tr>
<th>ALL RESEARCH DESIGNS</th>
<th>(Chapter 4)</th>
<th>CAUSAL RESEARCH DESIGNS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Is the estimate valid (true)?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Is the estimate precise (reliable)?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample representativeness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Is the sample representative of the intended population?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample size</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Is the sample large enough to overcome stochastic threats to inference?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Covariation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(association, correlation)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Does X co-vary with Y?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comparability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(equivalence, unit homogeneity)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Is each observation causally comparable to other observations in the sample?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Covariation

Empirical evidence of a causal relationship consists of patterns discovered in the data. Specifically, if a causal relationship between X and Y exists we expect to find covariation (association, correlation) between the two variables (probabilistically). If, for example, the causal factor of interest is a worker-training program (X), and the program has the hypothesized causal effect on wages (Y), we expect that the earnings of participants (\(Y_{X=1}^X\)) will be higher than earnings of non-participants (\(Y_{X=0}^X\)) when measured a year after completion of the program.

Covariation between X and Y is thus a necessary condition of causality. That is to say, if we learn nothing about Y’s probable value from looking at the value of X then X cannot be considered a cause of Y. Likewise, it is from this covariational pattern that we estimate a causal effect.

Imagine a hypothetical sample of 100 observations, each of which records a value for X, the causal factor of interest, and Y, the outcome of interest. In conditioning these variables we observe the value of Y for each value of X.

In the simplest case, both factors are binary. For all observations in which X=0, we observe the values of Y, which we designate as \(Y_{X=0}^X\). Next, for all observations in which X=1, we observe the values of Y, designated as \(Y_{X=1}^X\). Results of this test are contained in panel (a) of Figure 6.2.
this hypothetical sample, we observe a (positive) association between $X$ and $Y$. That is, knowing the value of $X$, we can predict the value of $Y$ (with some error).

Now, let us assume that $Y$ is an interval variable, varying from 0-100. Again, we compare $Y_{X=1}$ and $Y_{X=0}$. Results of this test are contained in panel (b) of Figure 6.2. Again, we see a covariational relationship. Knowing the value of $X$, we can predict the value of $Y$ (with some error).

Finally, let us assume that both $X$ and $Y$ are measured along an interval scale stretching from 0-100. For each value of $X$, we record the values for $Y$. This produces a scatterplot, as shown in panel (c) of Figure 6.2. Once again, we observe a (positive) covariational relationship.

All of these hypothetical data distributions point to a positive (and statistically significant) relationship between $X$ and $Y$. If the relationship were reversed – if the value of $Y$ decreased with each increase in $X$ – then a negative relationship would be in evidence. These are monotonic relationships (as explained in Chapter 2). All sorts of non-monotonic relationships are also possible. For example, the values of $X$ and $Y$ might be positive as $X$ increases from 0 to 50, where they peak; from 50 to 100 the relationship might be negative, as shown in panel (d).

This hump-backed (inverted-U) relationship may be quite common in the social world. The “Kuznets curve” is a wellknown (but contested) relationship between economic development and inequality named after Simon Kuznets, who hypothesized that income inequality would increase during the early period of industrialization and then begin to decrease as a country became more developed.\(^{78}\)

A final possibility is that there is no discernible pattern between $X$ and $Y$, or at least none that can be distinguished from what stochastic (random) factors might produce. Such a situation is illustrated in panel (e) of Figure 6.2. In this situation, the null hypothesis (of no relationship) cannot be rejected.

---

\(^{78}\) Kuznets (1955).
Figure 6.2: Covariational Patterns
COMPARABILITY

Finding a covariational pattern between $X$ and $Y$ does not assure that the relationship is causal. It might be merely associational. “Correlation does not equal causation,” as the saying goes. If we find an association among unemployed persons between attending a worker-training program ($X$) and higher earnings ($Y$) it does not automatically follow that $X$ causes $Y$. 

In order to infer a causal relationship between $X$ and $Y$ from the covariation of $X$ and $Y$ across a set of observations we must be able to convince ourselves that other things are equal (the *ceteris paribus* assumption). That is, we need to be assured that the variation we observe in $Y$ is a product of $X$ and not of some other cause(s). If *causal comparability* is present across the sample, we can say that the estimate derived from the analysis is valid. So comparability is the specific feature of a research design that allows one to make claims about *validity*, an issue introduced (in a general fashion) in Chapter 4. If comparability does not exist then the pattern of $X/Y$ covariation is spurious, i.e., characterized by systematic error or *bias*.

A more technical way of phrasing the assumption of comparability is that the *expected value* of $Y$ for a given value of $X$ should be the same across the studied observations throughout the period of analysis:

$$E(Y|X)$$

If this is the case, we can say that a group of observations is causally comparable (equivalent) with respect to a given hypothesis.\(^7\)

If, for example, a set of unemployed persons is causally comparable with respect to a worker-training program then we shall expect them to experience roughly the same change in earnings when subjected to the same worker-training program. This must remain true throughout the period of analysis, i.e., from the first observation to the last (if several observations of the outcome are taken through time).

Our (minimal) understanding of causal comparability requires only that units be comparable to one another *on average*. Error (non-comparability) across units is satisfactory so long as its distribution is centered on the true mean (i.e., so long as the error is random).\(^8\) That is, some persons may be positively affected by a worker-training program, while others weakly or even negatively affected. Yet, on average (across the sample), or in expectation, the effect of $X$ on $Y$ is assumed to be the same.

---

79 Strictly speaking, a single observation cannot be causally comparable to another because a single observation does not register variation between $X$ and $Y$. Causal comparability is the attribute of a set of observations, sometimes understood as a case or unit.

80 A maximal understanding of causal comparability (sometimes expressed as *unit homogeneity*) is that units should evidence identical responses of $Y$ to a given value of $X$ across units. The latter ideal is rarely, if ever, realized in the world of social science, and perhaps not even in the world of natural science. However, the minimal definition seems too minimal. After all, non-comparabilities are always somewhat problematic – at the very least, they introduce problems of noise (random error). They also may hide heterogenous causal effects within the chosen sample (where different units respond differently to the same treatment). Thus, we shall regard this goal as a matter of degrees. Greater causal comparability in a sample of observations is desirable, even if perfect comparability is rarely (if ever) accomplished.
CONCLUSIONS

We began this chapter by defining the concept of causality at some length, with the assistance of causal graphs. In the next section, we introduced criteria intended to define a *good* (well-constructed) causal hypothesis — that which researchers strive to achieve (or ought to strive to achieve) in their work. This includes **precision**, **generality**, **boundedness**, **parsimony**, **coherence**, **commensurability**, **innovation**, **relevance**, **clarity**, **manipulability**, **precedence**, **impact**, and **mechanism** (see Table 6.2).

The chapter proceeded to discuss the challenge of causal *analysis*, i.e., the formulation of an effective research design in order to test a hypothesis. Here, we dwelt on the importance of **covariation** and **comparability**.

In subsequent chapters, we put these tools to work in the discussion of specific research designs. Experimental designs will be addressed in Chapter 7, nonexperimental designs in Chapter 8, and case study designs in Chapter 9. Although these chapters address different topics, they are closely inter-related and build upon one another in a cumulative fashion. Indeed, all of the concepts introduced in this chapter will reappear in later chapters, so you will have a chance to review and to see how they apply in various research settings.
KEY TERMS

- Backdoor
- Cause
- Outcome
- Treatment group
- Control group
- Causal effect
- Counterfactual
- Scope-conditions
- Antecedent
- Endogenous
- Exogenous
- Monotonic causality
- Non-monotonic causality
- Causal equifinality
- Interaction effects
- Conjunctural causality
- Multiplier effect
- Sequential cause
- Proximate
- Distal
- Causal laws
- Probabilistic causes
- Causal heterogeneity
- Causal graph
- Clarity
- Manipulability
- Separation
- Precedence
- Impact
- Causal mechanism
- Causal comparability
- Expected value
- Covariation
- Confounder
- Common cause confounder
- Compound-treatment confounder
- Circular confounder
- Mechanistic confounder
• Collider confounder
• Causal chain
• Causal pathway
• Conditioning
• Binary
• Reversible causality
• Irreversible causality
• Linear, non-linear causality
• Stochastic

INQUIRIES

1. Label each of the following terms as (a) a synonym (or near-synonym) for causal factor, (b) a synonym (or near-synonym) for an outcome, or (c) neither.
   • Covariate
   • Response
   • Y
   • Explanatory variable
   • Explanans
   • Independent variable
   • Input
   • Intervention
   • Exogenous variable
   • Parent
   • Endogenous variable
   • X
   • Dependent variable
   • Condition
   • Descendant, effect
   • Explanandum
   • Right-side variable
   • Left-side variable
   • Treatment
   • Predictor
   • Output

2. “Correlation does not equal causation.” What does this mean?

3. What is a causal counterfactual?

4. What is a causal mechanism?
5. What are the background conditions (ceteris paribus assumptions) associated with a causal argument?

6. Why is causal exogeneity/endogeneity a matter of degrees (in nonexperimental settings)?

7. What does it mean to label a factor as a “variable”?

8. Construct a causal graph of this scenario: “When the IMF, in conjunction with other lenders, forced African countries to adopt austerity measures in the 1980s, most states cut their payroll and their personnel. This resulted in lower state capacity and contributed to unrest and civil war.”

9. Define each of these types of causal relationship:
   - Monotonicity/Non-monotonicity
   - Linearity/Non-linearity
   - Reversibility/ Irreversibility
   - Path-dependency (aka critical juncture)
   - Discrete/Constant
   - Independence
   - Equipotentiality
   - Interaction effects
   - Conjunctures
   - Multiplier
   - Sequence
   - Proximal/ Distal
   - Causal chain
   - Causal laws
   - Set-theoretic causes
   - Probabilistic causes
   - Causal heterogeneity

10. How does an author achieve clarity in a causal argument? What are some examples of ambiguous causal arguments (that should be avoided)?

11. What does manipulability refer to? Why is it important? Is it always achievable?

12. What does causal exogeneity refer to? What makes one factor more exogenous than another?

13. A cause that has greater impact on an outcome may be more useful than a cause with minor impact on an outcome. How might this be judged?

14. Causal mechanisms are sometimes obvious and sometimes difficult to discern or to test. Does this mean that the study of mechanisms cannot be empirical in the same sense of the study of causal effects?
15. Think about the causal frameworks introduced in Chapter 5 and summarized in Table 5.1. Do these frameworks have characteristic strengths and weaknesses in terms of the general criteria that causal arguments strive to achieve (summarized in Table 6.2)?

16. What does it mean to say that a set of units are \textit{causally comparable} with respect to a hypothesis? Consider the example of worker training, in which some unemployed persons participate in the program and others do not. Under what circumstances would one say that the two groups are causally comparable?

17. What is \textit{covariational} (correlational, associational) evidence? (What is \textit{not} covariational evidence for a causal hypothesis?) If covariation is different from causation, why is the former necessary for the latter?

\textbf{RESOURCES}

7. EXPERIMENTAL DESIGNS

Recall (from Chapter 6) that estimating a causal effect involves comparing a factual (that which actually happened) to a counterfactual (that which might have happened). Returning to our perennial example, let us say that we wish to estimate the effect of a job-training program on the earnings of unemployed people after they have completed the program. Those who participate in the program are members of the treatment group. Non-participants are members of the control group.

Let us say that we know the earnings of participants and non-participants after completing the program. These are the factuals. What we do not know is what their earnings would have been if their roles had been reversed. These are the counterfactuals.

Although the only sure way to compare factuals with counterfactuals is to employ a time-machine, a well-designed experiment comes close to the mythical time-machine insofar as the control group exemplifies the (unobserved) counterfactual for the treatment group, and the treatment group exemplifies the (unobserved) counterfactual for the control group. Under certain conditions, this is a plausible scenario.

For present purposes, the defining criterion of an experiment is that the treatment is randomly assigned (“randomized”) across subjects. We do not care who controls the experiment – the researcher conducting the study or someone else. (Sometimes, an experiment that unfolds naturally, without intervention by a researcher, is referred to as a natural experiment, as discussed below.)

We begin this chapter with a review of the problem of confounding as it applies to experimental research. We proceed to introduce various approaches to experimental research. The final section offers a series of examples of experiments conducted on a variety of diverse topics. Together, these sections should give the reader a sense of the opportunities for, and limitations of, experimental research in the social sciences.

EXPERIMENTS WITH AND WITHOUT CONFOUNDING

In its simplest version, a single treatment (e.g., a worker-training program) is randomly assigned to members drawn from a known population (e.g., unemployed people of a given age). That is, some are chosen to participate in the program (the treatment group) and others are chosen to become part of the control group. The causal effect of the program is measured by comparing annual earnings of program participants with annual earnings of non-participants one year after he completion of the program.

Since the assignment of the treatment is random, we can assume that all background factors that might affect earnings are equal (on average) across the treatment and control groups. The only difference between the two groups is that one has been subjected to the treatment and the other has not, as illustrated in Figure 7.1. Causal comparability is achieved, which is to say the expected value of the outcome is the same for each group, contingent upon its receiving the treatment or not (see Chapter 6).

Specifically, we compare the factual with the counterfactual by comparing two factuals: the earnings of those in the treatment group with the earnings of those in the control group. If the
program has the desired effect, those in the treatment group should have higher earnings, on average, than those in the control group. If there is no difference between the earnings of these two groups, or if the difference does not surpass that which might have been obtained through stochastic (random) factors, then we can conclude that there is unlikely to be a (positive) causal effect. And if the control group has higher earnings than the treatment group we may conclude that the program has adverse effects.

This simple data generating process may also be diagramed in a causal graph. In Figure 7.2 (which replicates Figure 6.1), \( Z \) is the random assignment process, \( X \) is the assignment itself (0=control, 1=treatment), \( Y \) is the outcome (subsequent earnings), and \( B \) represents other factors that might impact the outcome (e.g., motivation, job skills, interpersonal skills, luck). Note that \( B \) is orthogonal to (uncorrelated with) \( X \) because the treatment has been randomly assigned: persons in the treatment group are assumed to have, on average, the same level of motivation, job skills, interpersonal skills, and luck as those in the control group.

Let us assume that the training program has a positive impact on earnings. If so, we should observe a pattern of data that looks something like the \( X/Y \) graph in Figure 7.3. The mean value of \( Y \) when \( X=0 \) (\( \bar{Y}_{X=0} \)) is $22,248. The mean value of \( Y \) when \( X=1 \) (\( \bar{Y}_{X=1} \)) is $32,915. We can therefore conclude that the causal effect of a hypothetical change from \( X=0 \) to \( X=1 \) is given by the difference in means: \( \bar{Y}_{X=1} - \bar{Y}_{X=0} \), or $11,667. This may be regarded as an average treatment effect (ATE), since it averages results in the treatment and control groups. Implicit in this analysis is the idea that if one re-assigned those in the treatment group to the control group, and those in the control group to the treatment group, and then ran the experiment again, the measured causal effect would be the same.

This is the simplest – and best – research design imaginable. We need only condition on two variables – \( X \) and \( Y \) – observing their covariation. (The problem of discerning when differences between treatment and control groups are greater than might be expected to occur as a result of chance – i.e., “statistical significance” – is postponed until Part IV.) Of course, we could decide to condition on orthogonal factors (\( B \)) as well in order to increase the precision of our estimates. However, doing so would not affect our estimate of the causal effect, precisely because these additional factors are uncorrelated with \( X \).

**Figure 7.1: Experimental Data without Confounders: An Illustration**

![Figure 7.1: Experimental Data without Confounders: An Illustration](image)

\( E(Y|X) \) is the same for treatment and control groups
Figure 7.2: Experimental Data without Confounders: A Causal Graph

X = Causal factor
Y = Outcome
Z = Antecedent
B = Orthogonal covariates
[] = Not conditioned
$t$ test of statistical significance = -6.790.
**NON-COMPLIANCE**

By assigning a treatment randomly, experiments avoid one sort of confounding. But they are still open to other sorts of confounders. Understanding these potential confounders is important not only for understanding of experimental research but also for our understanding of non-experimental research – since many of these confounders also appear in circumstances where a treatment cannot be randomly assigned.

One potential problem is **non-compliance**. Suppose that we have assigned unemployed persons to our treatment group but they refuse to attend the worker-training program, or they drop out before the program is complete (attrition). There is little that we can do to solve problems of non-compliance in many research settings. Yet, the effect is to render a treatment group in which some members have not been treated, i.e., they have not been exposed to the treatment or have been exposed to less of it than others, as illustrated in Figure 7.4.

**Figure 7.4: Experimental Data with Non-Compliance: An Illustration**

<table>
<thead>
<tr>
<th>Control group</th>
<th>Treatment group</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X=0$</td>
<td>$X=1$</td>
</tr>
</tbody>
</table>

Some units assigned to the treatment group are not exposed to the treatment. Non-compliers indicated by $\otimes$
Several approaches may be taken to this problem. One may regard the estimate of causal effects as an estimate reflecting intention-to-treat (ITT) rather than an average treatment effect (ATE). This is a reasonable approach in settings where non-compliance is unavoidable and one wishes to take this into account in the final analysis. For example, a policymaker may wish to know what the causal effect of a worker-training program is likely to be on a population of unemployed people, given that some of those chosen to participate in the program are likely to drop out prior to completion. If attrition is inevitable, the ITT estimate may be more relevant than a (hypothetical) ATE estimate.

One may also choose to exclude non-compliers from the analysis. Thus, we might compare compliers in the treatment group with the (entire) control group. However, this raises a problem of comparison. Note that non-compliers are not likely to be similar to compliers; they may be less motivated to find work, for example. If we compare the compliers with the control group we shall be comparing apples and oranges, i.e., groups for which the expected value of $Y$, given $X$, is not the same.

One option is to try to identify those characteristics that differentiate compliers from non-compliers and then use these as controls in a regression or matching analysis. The purpose here is to identify subjects within the control group who are similar to compliers so that comparability is restored. However, it is often difficult to identify and measure those characteristics – especially if they are attitudinal (e.g., motivation). Another approach is to regard the random assignment to treatment and controls as an instrument in a two-stage analysis, as described in the next chapter (see Instrumental variables).

A final option is to forget the control group entirely, focusing instead on a pre- and post-comparison of compliers in the treatment group. If one has a pre-treatment measure of earnings, this approach is feasible. However, it must be borne in mind that since the compliers are different from the non-compliers in key respects we cannot regard the estimate as representing an average treatment effect (ATE). Rather, it should be understood as an average treatment effect on the treated (ATT). This means that the causal effect applies to units that are similar to those that actually received the treatment, but not to units drawn randomly from the population (in this case, the population of unemployed persons).

**Contamination**

A second potential problem with experimental designs occurs when the treatment and control conditions are not effectively isolated from each other, and thus are allowed to “contaminate” each other, as illustrated in Figure 7.5.

Contamination is rarely a problem in laboratory experiments precisely because experimenters can control all aspects of the situation. However, it is quite common in field experiments, where subjects in the treatment and control groups may intermingle or otherwise communicate with each other. (This distinction is discussed later in the chapter.)

Typically, contamination is caused when members of the control group manage to receive the treatment, or some aspects of the treatment. In our ongoing example, this might occur if participants who attend the worker-training program communicate with those in the control group, thereby spreading their learning among those who are not supposed to be exposed to the treatment.

There is no easy fix for contamination unless one can successfully model the process by which contamination occurs – an uncertain prospect, in most settings. A better approach, if it is possible, is to re-design the experiment so that each group is effectively isolated from each other.
Figure 7.5: Experimental Data with Contamination: An Illustration

$E(Y|X)$ is not the same for treatment and control groups
**COMPOUND TREATMENTS**

Sometimes, an experimental treatment combines two elements, only one of which is of theoretical interest. Because the treatment group is subject to both \( X \) (the factor of theoretical interest) and \( C \) (the background factor), one cannot discern the independent effect of \( X \) on \( Y \), which may be confounded by \( C \), as illustrated in Figure 7.6.

The most common scenario involves researcher (or experimenter) effects. This refers to a setting in which the subject is aware that she is part of a study and this knowledge alters her behavior. Since this is not the treatment of theoretical interest, it serves to confound the analysis.

For example, participants in a job-training program may be motivated to find a job simply because they know they are being studied. If so, this feature may generate differences in the outcome across treatment and control groups which have nothing to do with the content of the program and therefore offer a misleading picture of program impact.

A famous example occurred in a series of studies focused on workers at a General Electric plant in Hawthorne, Illinois. Here, it appears that workers in the treatment group applied themselves with greater diligence than those in the control group regardless of what the treatment consisted of. It was eventually determined that workers were responding not to the intended treatment but rather to the situation of being intensively monitored. \(^81\) Experimenter effects due to the condition of being studied have been known ever since as Hawthorne effects.

Another form of researcher effect occurs whenever subjects are influenced by the ascriptive features or personality of a researcher. For example, respondents may react differently when interviewed by white and black surveyors, thus introducing an important confounding factor into any study of racial attitudes.

---

The best approach to compound-treatment confounders is to re-design the experiment to alleviate the confounder. If the problem is caused by researcher effects one may “blind” the subjects of an experiment so that they are unaware of their status as members of treatment or control groups. If one suspects that the researcher, rather than the subject, is instilling bias into the analysis then the researcher should be “blinded.” And if both are suspected of introducing bias in the analysis, a double-blind protocol is desirable.

Of course, it is not always possible to achieve this research design feature due to practical or ethical constraints. Many social experiments are not truly blind (where subjects do not know whether they are in the treatment or control group), much less double-blind (where neither subjects nor researchers know who is in the treatment and control groups). And even when subjects are ignorant of their status they may respond to aspects of the treatment that are not of theoretical interest, thus introducing confounders into the analysis. Consequently, experimenter effects are difficult to avoid, and – worse – are not always apparent to the researcher.

An alternative is to subject the control group to the same “experimenter effects” as the treatment group, thus canceling out this feature of the design and rendering the two groups comparable. However, this may be difficult to implement.

Before concluding this section it is important to bear in mind that compound-treatment confounders are not limited to researcher effects. Other sorts of compound-treatment confounders will require different approaches.

**Varieties of Experiments**

We are accustomed to speak of experiments as if they all followed the same research design. This is true, but only to a point. While all experiments feature a randomized treatment (according to our
adopted definition), other features vary, sometimes quite a lot. In this section, we review experimental designs, randomization mechanisms, and research settings.

**TYPOLOGY OF DESIGNS**

There are many ways to establish an experimental protocol. However, most follow one of seven templates, summarized in Table 7.1. First, we need to introduce some notation that will be used in this table and in subsequent tables.

Observations \((O)\) through time \((t)\) are signaled as \(O_t\). \((O_t)\) refers to the third observation. By an observation we mean a point in time when someone (e.g., the researcher) measures the value of \(X\), \(Y\), and relevant covariates (if any) for all units in the study. Thus, in a study of the impact of worker-training programs \((X)\) on earnings \((Y)\) the researcher might measure \(X\) and \(Y\) for each person in the sample before \((O_{t1})\) and after \((O_{t2})\) the intervention. This generates a table of data, where each row represents a new observation, as discussed in Chapter 17. As previously, \(X^{ct}\) refers to the treatment condition (e.g., program participation), while \(X^{ct}\) refers to the control condition (e.g., program non-participation). Likewise, \(Y^{ct}\) refers to the value of the outcome (e.g., earnings) for members of the treatment group, while \(Y^{ct}\) refers to the value of the outcome for members of the control group.

A **group** refers to a set of units (e.g., individuals) that are exposed to the same treatment/control conditions. This might be a simple control condition, a simple treatment condition, or a series of treatment/control conditions administered through time. Sometimes, there is no pure control group: all groups receive the treatment at different times or receive different treatments. Whatever the case, each group in a research design is represented by a Roman numeral \((I, II, \ldots)\) in Table 7.1.

The size of a group – the number of subjects or units who exposed to a particular treatment or set of treatments (or to the control condition) – is indeterminate. However, it is important to bear in mind that the randomization procedure is premised on the existence of a sufficient number of observations to overcome stochastic threats to inference. The greater the number of units, the greater our assurance that results obtained from a randomized design are not the product of chance. Indeed, the virtues credited to the technique of randomization become plausible only as the number of units increases. And this, in turn, means that the experimental method is implicitly a large-\(N\) method of analysis. (See discussion of sample size in Chapter 4.)

For each experimental protocol, an appropriate statistical test must be adopted for evaluating the impact of an intervention(s). This might be as simple as a **t-test** measuring the difference of means between two groups. It might involve a multivariate model incorporating additional factors to reduce background noise. Or it might be more complicated.\(^{82}\) Typically, more complex models are required whenever confounders are suspected. Discussion of statistical models is postponed until Part IV; in this section we are concerned with issues of research design.

The simplest and most common experimental protocol is the **post-test only** design (#1 in Table 7.1). This involves a single treatment that is randomly assigned across the sample, creating two groups: the treatment group and the control group. Properly administered, only one feature – the presence or absence of treatment – differentiates the two groups. All relevant background features should be similar, on average, across the groups. A single post-test measures the outcome of concern at some point after the intervention.

Other experimental designs may be regarded as variants of the post-test only design.

Sometimes, observations are taken before and after the intervention, providing **pre- and post-tests** (#2 in Table 7.1). The potential effect of the treatment is calculated by comparing

\(^{82}\) Rubin (1991) reviews four approaches.
difference scores, i.e., differences on the measured outcome between the first set of observations (the pre-test) and the second set of observations (the post-test) for each group.

Sometimes, it is suspected that the effect of a treatment varies over time (after the initiation of the treatment), requiring **multiple post-tests** (#3 in Table 7.1) in order to track these variations. Indeed, wherever long-term causal effects are of interest multiple post-tests staged at regular intervals are advisable so that the endurance or attenuation of a treatment effect can be gauged.

Sometimes, it is desirable to stage an intervention across several sequences so that the same treatment is administered to all groups at different times. In varying the timing of an intervention, a **roll-out** design (#4 in Table 7.1) overcomes potential confounders that are coincident with the first intervention. Moreover, one is able to offer the treatment to all groups, which may be important for political or ethical reasons.

A **crossover** design (#5 in Table 7.1) may be employed if one is interested in testing sequencing effects. Does it matter whether $X_1$ is introduced before $X_2$, or vice-versa? Sometimes, it does, and a crossover design allows one to test the possibility by constructing groups that are subjected to different sequences among the same treatments.

A **factorial** design tests the interactive effects of several categorical treatment variables (#6 in Table 7.1). In the simplest version, two binary causal factors are combined into four possible treatments. If, for example, treatments consist of $X_1$=participation in a worker-training program and $X_2$=small classes, these can be combined into four groups by randomizing across two dimensions, as shown in Table 7.1.

This concludes our itinerary of experimental protocols. Note that well-established names exist for only a few of these. For others, a variety of names (usually rather long and cumbersome) may be employed. Despite the lack of a standard terminology, the construction of these designs is remarkably consistent across diverse research settings. Most experimental designs may be understood as variants of the classics.
### Table 7.1: A Typology of Experimental Designs

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>I.</th>
<th>II.</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Post-test only</td>
<td>X=1</td>
<td>O₂ get to 1</td>
<td>O₂ get 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>X=0</td>
<td>O₂ get to 0</td>
<td>O₂ get 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Pre-test post-test</td>
<td>O₁ get to 1</td>
<td>X=1</td>
<td>O₂</td>
<td>O₂ get 0</td>
<td>O₂ get 0</td>
<td>O₂ get 0</td>
</tr>
<tr>
<td></td>
<td>O₁ get to 0</td>
<td>X=0</td>
<td>O₂</td>
<td>O₂ get 0</td>
<td>O₂ get 0</td>
<td>O₂ get 0</td>
</tr>
<tr>
<td>3. Multiple post-tests</td>
<td>O₁ get to 1</td>
<td>X=1</td>
<td>O₂</td>
<td>X=0</td>
<td>O₂ get 0</td>
<td>O₂ get 0</td>
</tr>
<tr>
<td></td>
<td>O₁ get to 0</td>
<td>X=0</td>
<td>O₂</td>
<td>X=0</td>
<td>O₂ get 0</td>
<td>O₂ get 0</td>
</tr>
<tr>
<td>4. Roll-out</td>
<td>O₁ get to 1</td>
<td>X=1</td>
<td>O₂</td>
<td>X=0</td>
<td>O₂ get 0</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>O₁ get to 0</td>
<td>X=0</td>
<td>O₂</td>
<td>X=0</td>
<td>O₂ get 0</td>
<td>X</td>
</tr>
<tr>
<td>5. Crossover</td>
<td>O₁ get to 1</td>
<td>X₁=1</td>
<td>X₂=0</td>
<td>O₂</td>
<td>X₁=0, X₂=1</td>
<td>O₂</td>
</tr>
<tr>
<td></td>
<td>O₁ get to 0</td>
<td>X₁=0, X₂=1</td>
<td>O₂</td>
<td>X₁=1</td>
<td>X₂=0</td>
<td>O₂</td>
</tr>
<tr>
<td>6. Factorial</td>
<td>O₁</td>
<td>X₁=0, X₂=0</td>
<td>O₂</td>
<td>O₂</td>
<td>O₂</td>
<td>O₂</td>
</tr>
<tr>
<td></td>
<td>O₂</td>
<td>X₁=0, X₂=1</td>
<td>O₂</td>
<td>O₂</td>
<td>O₂</td>
<td>O₂</td>
</tr>
<tr>
<td></td>
<td>O₁</td>
<td>X₁=1, X₂=0</td>
<td>O₂</td>
<td>O₂</td>
<td>O₂</td>
<td>O₂</td>
</tr>
<tr>
<td></td>
<td>O₂</td>
<td>X₁=1, X₂=1</td>
<td>O₂</td>
<td>O₂</td>
<td>O₂</td>
<td>O₂</td>
</tr>
</tbody>
</table>

**RANDOMIZATION MECHANISMS**

Purposeful **randomization** of the treatment is the defining element of an experiment. By this, we mean that all members of the chosen sample have an equal opportunity of receiving the treatment (or the various treatment conditions, if there is more than one). This may be accomplished by choosing names from an urn, with a random number generator, or through some other method. In order to be fully realized, the goal of randomization must be intentional, which usually means that it is controlled by the researcher him/herself.

Sometimes, randomization is carried out within identified strata rather than across the entire sample. In this case, known as **blocking** (or sometimes as **matching**), units chosen for analysis are first stratified according to some background feature(s) relevant to the outcome of interest. Thus, in an experiment on worker-training programs might stratify a sample of unemployed persons by race, sex, age, and parental educational attainment. Within each stratum, the treatment (e.g., participation in the program) is randomized. If the strata consist of only two units each (e.g., two students), then one is working with blocked-pairs (i.e., blocks of two). This is a common technique to limit background noise wherever background heterogeneity is great and sample size is limited.

**RESEARCH SETTINGS**

Experiments may be carried out in a wide variety of settings. If the setting is constructed and controlled by the researcher it may be referred to as a **laboratory experiment**. If the setting is natural (i.e., more realistic) it may be referred to as a **field experiment**. However, there is no hard
and fast distinction between these two types of settings, which are, after all, matters of degree. (What is a “laboratory”?)

To be sure, laboratory settings are likely to involve more researcher control, and this may in turn result in fewer threats to inference (e.g., by contamination). However, the simple fact of being conducted in a laboratory is no assurance that threats to inference have been removed. Likewise, the fact of being conducted “in the field” is no reason to believe that experimental protocols have been violated. Ultimately, it is up to the researcher – and the consumers of a study – to judge.

Some experiments involve nothing more than a survey instrument. The only difference between a survey experiment and a regular survey is that the former has a randomized treatment. That is, some respondents have been selected at random to receive a different version of the survey, generating a treatment version and a control version whose responses can be compared. The cheap and easy availability of on-line surveys (see Chapter 13) have contributed to the growing popularity of survey experiments in recent years.

Typically, the researcher controls the assignment of the treatment – the randomization procedure – as well as other design features of an experiment. Occasionally, a randomized treatment occurs naturally, i.e., outside the control of the researcher. This may be referred to as a natural experiment. For example, a government-run lottery assigns winners randomly from among those with lottery tickets (if the lottery is honest). Those who win may be considered part of a treatment group, while those who lose may be considered part of a control group. Researchers have nothing to do with it. So long as randomization has been achieved (with a fairly high degree of certainty) we shall regard it as an experiment. (If, on the other hand, random assignment is not achieved, or is highly questionable, we shall classify the resulting study as observational, as discussed in the next chapter.)

**Examples**

It remains to be seen how applicable the randomization approach might be to the research agenda of the social sciences. Indeed, the hard methodological question is not whether experiments work (there is little doubt that they do), but what they work for. As an entrée into this issue, we review four examples drawn from diverse fields: (1) employment discrimination, (2) corruption control, (3) historic election campaigns, and (4) gender and leadership. Each will be briefly discussed to give a flavor of the sorts of subjects that experimenters have tackled in recent years, as well as potential methodological difficulties.

**Employment discrimination**

Employment discrimination is a policy question of obvious and enduring importance. Of course, few employers will admit to racially-based hiring practices.83 This must be inferred.

One genre of experiment involves sending white and black (or male and female) applicants – matched in all respects and coached to answer questions in a similar fashion – into the same job interviews.84 If white candidates experience higher success rates, this fact may be accorded to race-based discrimination. However, this procedure – known generically as an audit study – is subject to researcher bias. In particular, those conducting the experiments may act in ways that affect the

---

83 But see Kirschenman and Neckerman (1991).
84 E.g., Kenney and Wissoker (1994), Neumark et al. (1996). For an overview of this genre of field experiment see Pager (2007).
employer’s decision but has nothing to do with race per se. Note that since the pseudo-job candidates are aware of the objectives of the experiment they may (consciously or unconsciously) allow this to affect their behavior in the job interview. Black confederates may appear less eager to receive the job than white confederates because they believe this is how many black job candidates might behave or because they believe that this is how they will be perceived by the white employer. If so, then the reported results of the experiment are the product of experimenter effects rather than the treatment of interest (skin color), generating a compound-treatment confounder.85

In order to alleviate this potential bias, Bertrand and Mullainathan remove the experimenters entirely from the conduct of the experiment. Instead, they mail out hundreds of resumes in response to jobs advertised in the Boston and Chicago areas.86 These applications differ in one key respect: some of the names on the resumes were distinctively African-American and the others identifiably white. The researchers find that applications with “white” names like Emily and Greg were more likely to be contacted by employers for a follow-up interview than applications with recognizably “black” names like Lakisha and Jamal.

This clever field experiment provides some of the strongest evidence to date of employment discrimination. However, it too may be subject to a compound-treatment confounder. Note that the names chosen by Bertrand and Mullainathan as typically black are also unusual. This is because African-Americans parents adopt a much wider array of names for their children than white parents. Nonetheless, it could be that the lack of employment success realized by “Lakisha” and “Jamal” – relative to “Emily” and “Greg” – is due to the fact that employers find these names strange. Persons with exotic names may be tagged by employers (probably unconsciously) as less trustworthy, less likely to take direction and work well with others, and so forth. And this, rather than their race per se, may account for differential hiring decisions.

**Corruption control**

The causes of corruption are a central preoccupation among citizens and among scholars across the world. Yet, the question remains agonizingly diffuse. Studies generally build on cross-sectional analyses of countries (or states within a country), whose varying levels of corruption are measured through surveys of the public or country experts (e.g., the Transparency International and World Bank indices of corruption). If an institution is found to correlate with a higher level of corruption (taking all other relevant and measurable factors into account), then it may be interpreted as a causal relationship.87 This genre of work is evidently open to the same, familiar objection: perhaps it is a merely correlational relationship, accountable to some unmeasured common cause.

In order to bring experimental evidence to bear on the question of corruption, Ben Olken observes levels of corruption in road projects spread across over six hundred Indonesian villages. Corruption is measured by a variety of direct methods – most interestingly, by sampling the cores of selected roads in order to determine whether materials used were standard or substandard (chapter five).88 Two theories are tested. The first concerns the effect of an impending government audit, a

85 For other criticisms of the audit technique see Heckman and Siegelman (1993) and discussion in Pager (2007).
86 Bertrand and Mullainathan (2004).
87 E.g., Gerring and Thacker (2004).
88 Olken (2007: 203) explains: “I assembled a team of engineers and surveyors who, after the projects were completed, dug core samples in each road to estimate the quantity of materials used, surveyed local suppliers to estimate prices, and interviewed villagers to determine the wages paid on the project. From these data, I construct an independent estimate of the amount each project actually cost to build and then compare this estimate with what the village reported it spent on the project on a line-item by line-item basis. The difference between what the village claimed the road cost to build and what the engineers estimated it actually cost to build is the key measure of missing expenditures” – i.e., the measure of corruption.
The top-down approach to corruption control. The second concerns grassroots participation in monitoring the road project, a down-up approach to corruption control. Each of these treatments is randomized across the 600 villages. Olken finds that the threat of an audit had a much greater effect in reducing corruption than institution of village-level monitoring (though both had some effect).

**HISTORIC ELECTION CAMPAIGNS**

Traditionally, the use of experimental methods has been understood as a prospective, rather than retrospective, exercise. Because the research is designed by the experimenter, it may help us shed light on general phenomena that pertain to the past, but it cannot shed light on particular events in the past. While generally true, there are exceptions to this rule.

The role of the infamous “Willie Horton” ad in the 1988 US presidential campaign has exercised scholars and pundits since the day it aired. In the political advertisement, sponsored by a group loyal to George Bush’s campaign, a black man, Willie Horton, is shown along with a voiceover explaining that Horton was released from prison on furlough by Massachusetts Governor Michael Dukakis, after which he proceeded to rape a woman and viciously beat her fiancé. Was this a “race-baiting” ad, or was it really more about crime? What sort of effect might it have had on the general public? And what, more generally, is the effect of attack ads that seek to capitalize on fear?

Tali Mendelberg’s ingenious approach to these questions is to expose a sample of white college students to a laboratory experiment in which the treatment is exposure to the Horton ad, which is embedded within a news story. (Students chosen for the experiment had little or no knowledge of the ad, and are thus unaware of its notoriety.) The control consisted of a similar news story, without the Horton excerpt. Mendelberg finds that the ad enhanced the salience of race, rather than crime, and interacted with existing prejudices so as to affect subjects’ views on a range of issues in a more conservative direction.

Of course, this study is unable to determine (or even estimate) how long this effect lasted, how many voters it reached, and how many votes (if any) it changed. These very specific historical outcomes are beyond the scope of experimental methods to explore. Even so, Mendelberg’s influential study prompts us to consider ways in which experimental protocols might be enlisted to shed light on past events – a relatively new purview for the venerable experiment.

**GENDER AND LEADERSHIP**

Does the sex of a politician affect his/her policy decisions? The question has been much studied and much debated. However, there is little strong evidence to show whether gender has effects on the quality of political decisions, primarily because observational data are replete with potential confounders. If we simply compare the behavior of male and female legislators we run the risk of confusing gender effects with other factors that happen to co-vary with the gender of a legislator, e.g., party identification, the character of the election, or the character of the district. And if the empirical comparisons are cross-country, the number of potential confounders is even greater. A recent study by Dollar, Fisman, and Gatti reports a negative correlation between representation of women in parliaments and corruption. The causal question, raised pointedly by Chattapadhyay and

---

91 Additional issues pertaining to internal and external validity are discussed in a subsequent study by Huber and Lapinski (2006). See also the colloquy between Huber/Lapinski and Mendelberg in Perspectives on Politics 6:1 (March 2008) and Hutchings and Jardina (2009).
Duflo, is whether this means that women are less corrupt (as Dollar et al. claim), or that “countries that are less corrupt are also more likely to elect women to parliament.” This gets to the heart of our topic.

To shed light on this issue, Chattapadhyay and Duflo take advantage of a federal law passed in India in 1993 requiring that one-third of all village council heads (an elective position) be reserved for women. Because the assignment of women to positions of authority is randomized (though not by the researchers), it is possible to interpret policy choices made by village councils under male and female leadership as an indication of the causal effect of gender. Prior to this, of course, it is necessary to determine what the varying preferences of men and women in West Bengal and Rajasthan, the two states under study, might consist of. This is accomplished by examining the types of formal requests brought to the village council by male and female citizens.

In West Bengal, women complain more often than men about drinking water and roads, and there are more investments in drinking water and roads in [village councils] reserved for women. In Rajasthan, women complain more often than men about drinking water but less often about roads, and there are more investments in water and less investment in roads in [village councils] reserved for women. The authors find that these preferences are indeed reflected in the sort of public goods provided by governments in villages where a female council head is in charge – understood relative to villages in the control group, which are generally governed by men. In short, the sex of political leaders counts. Goods valued more highly by women are more likely to be distributed in villages where women hold important leadership positions.

The design features of this natural randomization are close to ideal. There are many units to study, the process of randomization seems to have been strictly adhered to, and there are viable outcome indicators by which one might judge the impact of the treatment. In short, the set-up in this experiment appears to be about as sound as one might have achieved even with an experiment that was controlled by the researcher.

Of course, the precise causal mechanisms at work in this setting are somewhat open to interpretation. Is the difference in policy outcomes between quota and non-quota villages a product of female leaders’ desire to represent the interests of women in their constituencies or a product of the personal attributes and life-histories of the chosen female leaders? (Would well-off and politically empowered female leaders behave differently?) The authors try to address this question by running statistical analyses that control for various characteristics of female leaders, finding no effect. Even so, one may suspect that there is insufficient variation in these properties to provide a proper test. The general point is that once one moves from questions about the main causal effect (which is randomized) to questions of causal mechanisms (which, in this case and most others, is not randomized) the precision and confidence of the analysis suffers accordingly. (This also provides a good example of a study in which experimental and nonexperimental styles of analysis coincide.)

A final point of clarification is in order. Chattapadhyay and Duflo are careful to present their research as a test of gender quotas, not a test of gender per se. Note that it is the gender quota that is being randomized not the gender of specific political leaders or the gender of constituents within the studied communities (of course). What one may reasonably conclude from this experiment is that gender quotas influence public policy outcomes wherever gender preferences diverge (men and women in a community want different things) and one group is politically disadvantaged. We do not learn what features of gender are driving divergent preferences, either on the part of elites or of masses. What about “gender” affects public policy?

95 Chattapadhyay and Duflo (2004: 1411).
CONCLUSIONS

We have defined the experiment as a research design in which the treatment is randomly assigned across units. This avoids common-cause and circular confounders. But it does not solve threats to inference that arise after the treatment has been assigned – specifically, the problems posed by non-compliance, contamination, and compound treatments, as discussed.

Anticipating and identifying these confounders is, first and foremost, a matter of careful attention to the details of the research design and vigilant observation of the research as it progresses in real time. In this respect, constructing a sound experiment involves an ethnographic set of skills and resources. One must know one’s subjects and one must be in a position to speculate what the reaction to a treatment might be in order to devise an adequate experimental test. Experiments cannot be constructed by rote adherence to an experimental handbook. Local knowledge, including qualitative evidence, is essential.96

In the event, some pre- or post-treatment threats to inference may be unavoidable. This does not mean that all is lost. Sometimes, adequate corrections can be applied ex post. However, these work-arounds usually involve the introduction of assumptions (e.g., about the nature of noncompliers) that are difficult to test empirically, and which therefore compromise the confidence with which we may regard an experimental finding. Insofar as an experiment depends upon “statistical” corrections, it begins to look more like observational research (the topic of the next chapter).

Before concluding, it is important to say a few words about external validity – the validity of experiments for a broader population of units (see Chapter 4). This is often regarded as the Achilles heel of experimental research. Note that units chosen for experimental analysis are rarely sampled randomly from a known universe. Instead, subjects are taken opportunistically. As a result, it is often difficult to generalize the results of an experiment – whether conducted in a laboratory or in the field. This common feature of experiments is not coincidental. Because experimental work involves an intentional manipulation – subjects are “treated” – it is necessary to obtain consent. As such, many prospective participants are likely to demur, leaving the researcher with a less than random sample.

This does not doom the experimental method to triviality. Because of their generally high internal validity, experimental case studies have the potential to cumulate – if the protocols employed in these studies are standardized, and hence replicable across diverse settings. Multiple experiments on the same subject in different settings may help, ultimately, to bring a larger picture into focus if the key features of those individual experiments are commensurable with one another. In this fashion, we have learned from repeated field experiments what effects a variety of different treatments have on voter turnout (at least, in the United States).97 The same replication may, over time, contribute to the cumulation of knowledge on other subjects of interest to social science. Of course, this will not happen quickly, or cheaply, and will require significant adjustment of scholarly incentives – which do not generally reward the faithful replication of extant studies.

Even so, some subjects are likely to escape experimental protocols. Try as one might, it will be difficult to test the impact of economic development on democratization in a randomized trial. It is not even clear how economic development should be operationalized, even if one could distribute it randomly throughout the universe of nation-states.

---

96 Paluck (2010).
97 For a compilation of experimental studies on turnout see the Get Out The Vote (GOTV) web site maintained by Yale University’s Institute for Social and Policy Studies (http://research.yale.edu/GOTV/).
KEY TERMS

- Assignment
- Treatment group
- Control group
- Expected value
- Average treatment effect (ATE)
- Non-compliance
- Intention-to-treat (ITT)
- Average treatment effect on the treated (ATT)
- Researcher (or experimenter) effects
- Hawthorne effects
- Experimental designs (post-test only, pre-and post-test, multiple post-tests, roll-out, crossover, factorial)
- Contamination
- Randomization
- Reputation effects
- Selection bias
- Laboratory experiment
- Field experiment
- Survey experiment
- Natural experiment

INQUIRIES

1. An experiment is the next-best-thing to a time-machine (for purposes of reaching causal inference). Explain the logic of this assertion.

2. What does a randomized treatment mean? What are the different techniques used to randomize a treatment?

3. Suppose we conduct an experimental test of a new program to improve student attendance at the secondary level. The program is administered at the individual level; that is, students are given rewards or punishments to encourage their attendance in classes. Students are randomly assigned to a treatment and control group and their attendance is measured for several months after the inauguration of the program. We then compare mean attendance rates for the two groups (treatment and control). Next, we use information about these students – their age, GPA, prior attendance record, family income, and so forth – to conduct a multivariate analysis. That is, we examine the treatment effect controlling for these other factors. Should we expect to find much difference in these two analyses (the simple bivariate analysis and the multivariate analysis)? Why, or why not?
4. What is *non-compliance*? Why is it such a persistent problem in social experiments? What are some approaches to solving (or mitigating) this problem?

5. What is *contamination*? How can it be avoided (at least in principle)?

6. What is a *compound treatment*? Why is it a problem? How might it be avoided?

7. There are many ways to design an experiment. Here are diagrams showing some of the most common approaches. What are the distinctive goals/purposes of each of these designs?

<table>
<thead>
<tr>
<th>Design</th>
<th>Groups</th>
<th>Observations through time</th>
<th>Control condition</th>
<th>Treatment condition</th>
<th>Interventions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Post-test</td>
<td>I.</td>
<td>$X=1$</td>
<td>$O_{t1}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>only</td>
<td>II.</td>
<td>$X=0$</td>
<td>$O_{t1}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Pre-test</td>
<td>I.</td>
<td></td>
<td>$X=1$ $O_{t2}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>post-test</td>
<td>II.</td>
<td></td>
<td>$X=0$ $O_{t2}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Multiple</td>
<td>I.</td>
<td>$X=1$</td>
<td>$O_{t1}$</td>
<td>$O_{t2}$</td>
<td>$O_{t3}$</td>
</tr>
<tr>
<td>post-tests</td>
<td>II.</td>
<td>$X=0$</td>
<td>$O_{t1}$</td>
<td>$O_{t2}$</td>
<td>$O_{t3}$</td>
</tr>
<tr>
<td>4. Roll-out</td>
<td>I.</td>
<td>$X=1$ $O_{t1}$</td>
<td>$O_{t2}$</td>
<td>$X=0$ $O_{t4}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>II.</td>
<td>$X=0$ $O_{t1}$</td>
<td>$O_{t2}$</td>
<td>$X=1$ $O_{t4}$</td>
<td></td>
</tr>
<tr>
<td>5. Crossover</td>
<td>I.</td>
<td>$X_1=1, X_2=0$ $O_{t1}$</td>
<td>$O_{t2}$</td>
<td>$X_1=0, X_2=1$</td>
<td>$O_{t4}$</td>
</tr>
<tr>
<td></td>
<td>II.</td>
<td>$X_1=0, X_2=1$ $O_{t1}$</td>
<td>$O_{t2}$</td>
<td>$X_1=1, X_2=0$</td>
<td>$O_{t4}$</td>
</tr>
<tr>
<td>6. Factorial</td>
<td>I.</td>
<td></td>
<td>$O_{t1}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>II.</td>
<td></td>
<td>$O_{t1}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>III.</td>
<td></td>
<td>$O_{t1}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>IV.</td>
<td></td>
<td>$O_{t1}$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

8. What is the difference between a *laboratory experiment* and a *field experiment*?

9. For many, experiments are the hallmark of natural science and a legitimate goal for social science. But others feel that experimental methods would effectively exclude too much of the work that is traditionally conducted in the social sciences. How much of social science is amenable to an experimental approach? How much is not?

10. Consider the question of democracy and development. Does regime-type affect development? Could aspects of this question be explored experimentally?
11. Consider the question of social capital and governance. Does improved social capital lead to improvements in the quality of governance? Could aspects of this question be explored experimentally?

12. Consider the question of worker-training programs. Does attendance in a worker-training program enhance employment and earnings among unemployed persons? Are there any aspects of this question that cannot be addressed in an experimental setting?

13. If you are working on a research question, consider how it might be addressed with an experiment.

RESOURCES

PUBLISHED WORK


**WEBSITES**

8. Large-N Observational Designs

There are many potential problems with experimental designs, as reviewed in the previous chapter. However, it is generally the case that experiments have a much stronger claim to internal validity than non-experimental research. Insofar as the latter can be justified it is either because experiments are impossible or because they lack external validity. As it happens, this is quite common. And this, in turn, accounts for the ongoing predominance of observational designs in the work of social science.

In this chapter, we review several approaches to the analysis of large samples without randomization. (The subsequent chapter addresses the analysis of small samples without randomization, i.e., case studies.) Five archetypal research designs will be distinguished: cross-section (CS), time-series (TS), time-series cross-section (TSCS), regression discontinuity (RD), and instrumental variable (IV), as summarized in Table 8.1.98

These terms are defined by the observed variation in X, the causal factor of interest, or the instrument by which the treatment is assigned. In a cross-section design, observed variation in X is across units; this is what provides the leverage for causal inference. In a time-series design, observed variation in X is through time, typically with a single unit or case. In a time-series cross-section design, observed variation in X is both across units and through time. In a regression discontinuity design, observed variation is in X just above and below an arbitrary cutoff. In an instrumental variables design, observed variation is in an instrument, Z, that affects assignment to treatment, X.

It is important to bear in mind that most of these terms also carry a narrower, more technical definition as a type of statistical model. For example, a cross-section in the statistical sense is an analysis in which a set of units is observed at a single point in time. A time-series in the statistical sense is an analysis in which a unit is observed over time. And a time-series cross-section is an analysis in which a set of units is observed laterally (across units) and longitudinally (through time). Here, our emphasis is on research design properties that might be conducive to causal inference. Note, for example, that a time-series (in the statistical sense) is not very useful unless there is observed variation in X contained within the time-series, so we define a time-series design in this fashion rather than as an iterated set of observations through time. As previously in this book, our emphasis is on the design properties of causal inference rather than on statistical methods of analysis (a topic postponed until Part IV).

That said, there is greater discussion of methods of statistical analysis in this chapter than in previous chapters. This is because correcting for the deficiencies of observational designs requires more complex methods of data analysis. While a well-constructed experiment can be analyzed simply by comparing the mean values of Y across treatment (X=1) and control (X=0) groups, reaching causal inference with observational data is not so simple. Bluntly stated: if you can’t manipulate the treatment (ex ante) you need to do a bit of (ex post) data manipulation in order to arrive at results that can plausibly be regarded as evidence of a causal relationship. This means that we must introduce techniques of data analysis such as regression and matching—techniques that are not

---

98 Another approach to causal inference is through the mechanism (M) that connects X with Y. However, this “front-door” approach remains unusual. For further discussion see Glynn, Gerring (2013), Morgan, Winship (2007), Pearl (2009).
dealt with formally until Part IV. Readers who find this method of exposition confusing may wish to peruse the relevant chapters in Part IV prior to reading this chapter.

Table 8.1: Large-N Observational Research Designs

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross-section (CS)</td>
<td>Variation in $X$ analyzed across units but not through time.</td>
</tr>
<tr>
<td>Time-series (TS)</td>
<td>Variation in $X$ analyzed through time but not across units (all units are exposed to the same treatment conditions, or there is only one unit).</td>
</tr>
<tr>
<td>Time-series Cross-section (TSCS)</td>
<td>Variation in $X$ analyzed across units and through time (e.g., panel, difference-in-difference, and pooled cross-section)</td>
</tr>
<tr>
<td>Regression discontinuity (RD)</td>
<td>Variation in $X$ analyzed just above and below an arbitrary cutoff.</td>
</tr>
<tr>
<td>Instrumental variables (IV)</td>
<td>Variation in $Z$ affects assignment to treatment ($X$) but not the outcome ($Y$).</td>
</tr>
</tbody>
</table>

CROSS-SECTIONAL DESIGNS

Consider the situation when a worker-training program is administered on a voluntary basis: all unemployed persons who wish to attend do so, and those who do not form the control group. A post-test taken one year after the completion of the program measures worker earnings for those who attended the program and those who did not. The design is illustrated with $X$’s and $O$’s in Table 8.2.

Table 8.2: Cross-sectional Design

<table>
<thead>
<tr>
<th>Groups</th>
<th>$X=1$</th>
<th>$O_{it}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>$X=0$</td>
<td>$O_{it}$</td>
</tr>
</tbody>
</table>

- $O_{it}$: Observations
- $X=1$: Treatment condition
- $X=0$: Control condition
Superficially, this design is identical to the post-test only experimental design, illustrated in Table 7.1. However, under conditions of voluntary choice we are not likely to achieve comparability across the treatment and control groups, for those with greater motivation are likely to sign up, and their greater motivation – rather than the content of the program itself – may account for their superior earnings relative to the control group. As such, the background factors are not the same across the two groups, a feature illustrated in Figure 8.1. Accordingly, the expected value of $Y$, given $X$, is not the same across the groups.

**Figure 8.1: Cross-sectional Data: A Typical Scenario**

\[
\begin{array}{c|c}
\text{Control group} & \text{Treatment group} \\
\hline
X=0 & X=1 \\
(Weakly motivated) & (Strongly motivated)
\end{array}
\]

$E(Y|X)$ is not the same for treatment and control groups
The problematic confounder in this case is a common-cause confounder, so-called because it impacts both the causal factor ($X$) and the outcome ($Y$). In this instance, a highly motivated unemployed person is more likely to select into the treatment and is more likely to have higher earnings relative to those with lower motivation. The associated causal graph is illustrated in Figure 8.2.

There are potential solutions to this sort of problem. The simplest and most common is to condition on $C$, the confounder. What this means is that rather than simply looking at the relationship between $X$ and $Y$ we must now examine a multivariate relationship – $X/Y$, holding $C$ constant.

This is easiest to explain if we assume that $C$, like $X$, is binary. Let us say that motivation is a binary quality; one either has it or one does not; and let us imagine that this feature of a personality can be measured. This allows us to present the combination of possible values in a 2x2 matrix, as shown in Figure 8.3.

With this matrix, we pursue two sorts of comparisons. First, we compare values for the outcome, $Y$, when $X=0$ and $X=1$, holding $C$ constant at 0 (unmotivated people). By subtracting $Y_{X=0,C=0}$ from $Y_{X=1,C=0}$ we obtain the causal effect for all cases where $C=0$. Next, we do the same for the next row. Subtracting $Y_{X=0,C=1}$ from $Y_{X=1,C=1}$ we obtain the causal effect for all cases where $C=1$ (motivated people).
This may be achieved in the pencil-and-paper method we have just reviewed, or with estimators (e.g., regression or matching) that are capable of accommodating a large number of confounders. Note that although the mechanics of punching buttons in a software program are easy, there are quite a number of assumptions wrapped up in this sort of analysis. Each must hold true, or the result will be spurious.

Returning to our exemplar, let us examine a few of the potential problems. First, we must assume that we have adequately accounted for the suspected common-cause confounder, motivation. Unfortunately, this is a difficult matter to measure. One cannot simply ask people whether they are highly motivated and expect to receive an honest answer. This is not simply because people may lie but also because people are not always aware of how motivated they are or how persistent they will be in pursuit of their goals. Of course, it would help if we measured motivation in a more sensitive fashion. Rather than a binary scale we could use an interval scale, stretching, let us say, from 0 to 100. One must then assume a particular function – say, linear or log-linear – to represent the relationship between motivation and earnings, and one must adopt a regression framework to conduct the analysis, involving additional assumptions (see Part IV).

Second, one must assume that one has accounted for all the additional confounders that may be present. Motivation is just one possible confounder. It could be that those who sign up for the worker-training program are more skilled – or less skilled – than those who do not. It could be that males, or whites, or Protestants, are over-represented relative to the unemployed people who do not join the program. Each of these features may affect the outcome, measured as earnings, and thus may serve as a confounder. How is one to know whether one has included all of these factors as controls in the analysis? (What if some of them are impossible to measure?) This is referred to as omitted variable bias.

Third, one must assume that all those factors included as controls in the analysis actually are confounders. If they are not confounders they may introduce confounding to an otherwise valid analysis. This is apt to be the case if factors are measured in a post-test, i.e., after the treatment (the program) has been administered. Suppose, for example, that we measure motivation at the end of a worker-training program rather than at the beginning. In this instance, we might imagine that the motivation levels of unemployed persons participating in the program would be influenced by their participation in the program. If it is a well-run program they may feel more motivated to find work than they did at the outset. Indeed, this may serve as an important causal mechanism if the program has its desired impact. Conditioning on this factor in a multivariate analysis will block the impact of
$X$ on $Y$, resulting in an attenuated estimate and perhaps occluding the causal effect entirely, as illustrated in Figure 8.4.

**Figure 8.4: Causal Graph with Mechanismic (post-treatment) Confounder**

```
X -> M -> Y
```

$X =$ Causal factor  
$Y =$ Outcome  
$M =$ Mechanismic confounder

Another potential problem posed by unwittingly conditioning on non-confounders is the **collider**. In its simplest version, a confounder is generated when one conditions on a factor that is affected by both the causal factor of interest ($X$) and the outcome ($Y$), as diagramed in Figure 8.5. Consider the following scenario. In a bid to condition on all possible common-cause confounders in the analysis of a worker-training program one decides to include union membership as a covariate in the causal model. However, if union membership ($C$) is affected by participation in the worker-training program ($X$) and by earnings ($Y$), a spurious relationship between $X$ and $Y$ is generated. That is, among union members ($C=1$) one will observe covariation between participation in the program and earnings that is not causal. Likewise, among non-union members ($C=0$) one will observe covariation between participation in the program and earnings that is not causal.
Finally, if entrance into the program is voluntary one can imagine that the outcome—earnings—might influence participation. For example, if the program is not free then those with higher earnings (prior to their current spell of unemployment) may be in a better position to participate. Even if it is free, they may be more likely to participate because they can better afford to forego current earnings for enhanced future earnings. Those with little or no savings may be obliged to accept the first job they find, presumably a low-paying job, rather than to attend a job-training program that promises higher rewards at some point in the future. This sort of circularity between cause and effect, illustrated in Figure 8.6, will also bias the analysis. In this case, it will result in a stronger correlation between X and Y than is warranted by the independent effect of X on Y. As with common-cause confounders, one may be able to restore comparability between treatment and control groups by blocking the circular path. In this case, that would require conditioning on pre-treatment income or wealth.
Figure 8.6: Causal Graph with Circular Confounding

![Causal Graph with Circular Confounding]

$X =$ Causal factor  
$Y =$ Outcome  
$C =$ Confounder  
[] = Not conditioned

**EXAMPLE**

Ethnic groups are everywhere, but only in some instances do they become fodder for politics, i.e., lines of cleavage between party groupings. Here lies a classic question regarding the construction of political identities. Daniel Posner surmises that the political salience of ethnic boundaries has a lot to do with the size of the ethnic groups relative to the size of the polity. Specifically, “If [a] cultural cleavage defines groups that are large enough to constitute viable coalitions in the competition for political power, then politicians will mobilize these groups and the cleavage that divides them will become politically salient.”

In order to pursue this hypothesis, Posner takes advantage of the arbitrary nature of political borders in Africa, where national boundaries are largely the product of intra-European colonial struggles rather than indigenous nation-building. This means that, unlike political boundaries in Europe, borders in Africa may be regarded as random elements of the political universe. The assignment problem is presumably (or at least plausibly) solved. In particular, Posner focuses on the border between Zambia and Malawi, which has separated members of two tribes, the Chewa and Tumbuka, since 1891, when these territories were held by the British (as Northeastern and Northwestern Rhodesia). As a product of this line-drawing exercise (conducted purely for administrative purposes, Posner says), Chewas and Tumbukas became very small minorities within the Zambian polity (7% and 4% respectively of the national population) and large minorities within the much smaller – Malawian polity (28% and 12% of the national population). Posner argues that this difference in relative size explains the construction of ethnic group relations in the two countries. In Zambia, Chewas and Tumbukas are allies, while in Malawi they are adversaries. This is borne out by surveys that Posner administers to villagers within each ethnic group on both sides of the border, and is also the received wisdom among scholars and experts.

Of course, a good deal of time elapses between the treatment (whose causal effect presumably begins with the initial partition of the territory in 1891, and accelerates after the independence of the two countries in 1964) and the post-test (in the early twenty-first century).

---

Typically, institutional factors exert a small but steady causal influence over many years, so this is a reasonable way to test the theory of theoretical interest. And yet, whenever a great deal of time elapses between a treatment and an outcome of interest it is difficult to reach firm conclusions about causality. And when pre-tests are lacking, as they are (by definition) in all cross-sectional designs, inferential difficulties are compounded. In these respects, cross-sectional (post-test only) designs are a lot weaker than panel designs.

In Posner’s study, even if the assignment problem is solved there are still a large number of potential confounders that threaten to creep into the research design after (or coincident with) the establishment of national borders. Specifically, any factor correlated with the treatment — “country” — is a potential confounder. It might well be, for example, that ethnicity is treated differently in Zambia and Malawi for reasons other than the sheer size of the ethnic groups. Posner looks closely at several of these alternative accounts including the actions of the colonial power, missionaries, ethnic entrepreneurs, and diverse national trajectories. This portion of the study draws on auxiliary evidence composed of causal-process observations (Chapter 10).

Posner does a good job of addressing the historical evidence. Even so, such confounders are difficult to dispense with, and stochastic threats to inference (factors that cannot be readily identified or theorized) are equally problematic. Under the circumstances, it might help to compare the politicization of ethnicity across small and large groups within each country, to study an ethnic group that is found in a large number of countries (e.g., Han Chinese), or to observe changes in the politicization of ethnicity as an immigrant community grows in size over time within a single country (a longitudinal design). There are many ways to skin this cat. Nonetheless, Posner offers an ingenious and plausible test of a difficult causal question.

**TIME-SERIES DESIGNS**

Having reviewed some of the complications faced by cross-sectional designs with nonexperimental data, we can conclude the obvious: it is difficult to restore the virtues of randomization with a non-randomized treatment. Despite our best efforts, confounders may persist, or may be unwittingly generated by our attempt to overcome them. And, in the end, we have no way of knowing for sure whether we have achieved causal comparability between treatment and control groups, i.e., whether the expected value of $Y$, given $X$, is the same for both groups.

Another option dispenses with the control group entirely, focusing instead on the group receiving the treatment. This will be referred to as a **time-series design**. Instead of comparing the treatment and control groups one observes a group through time — before and after treatment — as a clue to causal relations. One may think of this as a treatment and control conditions observed through time rather than across groups. The pre-treatment condition exemplifies the “control” group and the post-treatment condition exemplifies the “treatment” group.

Typically, this involves a single unit (or case). However, occasionally a group of units are exposed to the same treatment condition(s) at the same time. Here, a number of units may be observed simultaneously. However, because $X$ takes on the same values for all of the units there is no “control” group (in the usual sense).

Table 8.3 distinguishes three sorts of time-series research design. The simplest involves a single treatment with pre- and post-tests. For example, one might observe a group of unemployed persons who join a worker-training program, measuring their earnings before they join the program and again one year after they complete the program.
A second involves a single treatment accompanied by multiple pre- and post-tests. This is commonly referred to as an **interrupted time-series**. For example, one might observe the earnings of a group of unemployed people at monthly intervals for several years prior to, and after, they take part in a worker-training program.

A final variety involves the multiple iteration of a single treatment. This is known as a **repeated observations** (or **repeated measures**) design. For example, one might observe a group of unemployed persons who attend one or several worker-training programs over several decades, measuring their income every year.

Methodologically, these three designs are similar, though in some settings the interrupted time-series and repeated observations designs offer advantages over the simple pre- and post-test design.

### Table 8.3: Time-series Research Designs

<table>
<thead>
<tr>
<th>Design</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. One group pre/post test</td>
<td>[ O_{t1} \quad X=1 \quad O_{t2} ]</td>
</tr>
<tr>
<td>2. Interrupted time-series</td>
<td>[ O_{t1} \quad \ldots \quad O_{t2} \quad O_{t3} \quad O_{t4} \quad X=1 \quad O_{t5} \quad O_{t6} \quad O_{t7} \quad \ldots ]</td>
</tr>
<tr>
<td>3. Repeated observations</td>
<td>[ O_{t1} \quad \ldots \quad O_{t2} \quad X=1 \quad O_{t3} \quad X=1 \quad O_{t4} \quad X=1 \quad \ldots ]</td>
</tr>
</tbody>
</table>

- **I**: A group
- **O**: Observations through time
- **X=1**: Treatment condition

### Figure 8.7: Time-series Data: A Typical Scenario

\[ E(Y|X) \] is not the same for the group, pre- and post-treatment
In this context, the question of comparability concerns the status of the group prior to, and after, being exposed to the treatment. That is, the expected value of $Y$, given $X$, must be the same pre- and post-treatment, i.e., at $T_1$ and at $T_2$ ($T$=time). Any violation of this criterion will introduce bias into the analysis.

Unfortunately, there are often confounders lurking in a time-series design. Consider the worker-training program. Some unemployed persons may undergo a marked increase in their level of motivation. This might arise from a conversation with a friend or family member, a religious experience, the birth of a child, or some other life-changing event. This increased motivation may incline them to (a) join a worker-training program and (b) apply themselves with renewed vigor to the arduous process of finding a job. As a result, we may find a spurious time-trend. Those who participate in worker-training programs may experience higher subsequent earnings not because they benefited from the program but rather because they are more motivated than they were in the pre-treatment period. The expected value of $Y$, given $X$, is thus not the same at $T_1$ as it is at $T_2$. The common-cause confounder is again motivation. However, in the time-series research design it is a change in motivation, rather than a static feature of individuals, that creates the problem.

As with any other design, any feature that correlates with the treatment and also affects the outcome is a potential confounder (if not conditioned). This is a common state of affairs in time-series designs simply because treatments are often temporally associated with other things that also impact an outcome. If one wishes to find a job one is likely to take multiple approaches to this problem. Consequently, it is difficult to parse out the impact of a worker-training program from all the other actions a highly motivated unemployed person might take. Likewise, if one is measuring a policy initiative at a state or national level it is likely that the policy initiative will be accompanied by lots of other policy initiatives, undertaken at more or less the same time. Consequently, it will be difficult to distinguish the causal effect of one from the causal effects of all the others.

Threats to inference in a time-series design include any temporal feature that affects the time-trend. For example, suppose that earnings are increasing across the board in a growing economy – a constant trend. At some point, a time-series design is implemented to test the impact of a worker-training program. If one simply compares earnings before and after the treatment within this group one is likely to find improvement. But this improvement may be due to the existing time-trend rather than the program itself.

One time-trend is so common, and so vexing, that it has a special name: regression to the mean. Commonly, we take action on a difficult matter only when the urgency is great. Individuals, like governments, wait for emergencies in order to institute reform. For an unemployed person, this might be the point when her bank account reaches zero. For a government, it might be point at which societal unemployment reaches a particularly high point (in light of that country’s historical experience). At such moments, heroic action is taken – the worker decides to sign up for a grueling job-training program, a government decides to institute an expensive worker-training program. Many things in life observe a cyclical pattern, and unemployment is probably one of these. So it will not be surprising if, subsequent to joining a worker-training program, the unemployed person finds work, or if, subsequent to instituting a worker-training program, the unemployment rate goes down. However, it would be a mistake to attribute these changes to the worker-training program. They are, instead, examples of regression to the mean, i.e., a return to a normal (average) state of affairs.

To be sure, there is some hope of handling the problem of pre-existing trends (but not omitted variables coterminous with the treatment). This can be accomplished with a careful examination of a trend-line, followed by corrective action. This requires a good deal of temporal data; a simple pre- and post-test will not suffice. Where the data are rich, an extensive set of operations has been developed for “de-trending” time-series data so that $X$’s true effect on $Y$ can be
correctly estimated. It should be recognized that each of these operations involves significant, and difficult to test, assumptions about the data-generating process.\textsuperscript{100} Time-series econometrics, even in the most sophisticated hands, is plagued with ambiguity. If the trend is complicated – involving let us say, a long-term non-linear trend, a short-term cyclical trend, and lots of stochastic variation – one will be at pains to estimate the true causal effect of $X$ on $Y$.

At first glance, the repeated measures approach to time-series analysis appears to solve these problems. To be sure, if the unit returns to equilibrium after each intervention, then each intervention may be understood as an independent test of a given proposition. A single case observed longitudinally thus serves the function of a number of treatment and control cases, observed latitudinally. In effect, one tests and re-tests a single unit.

However, in many other situations common to social science there are enduring testing effects. Typically, the effect of an intervention is to change the unit experiencing the intervention. If so, the \textit{tabula} is no longer \textit{rasa}. Even if the unit remains the same, other contextual elements may vary from $T_1$ to $T_2$, rendering the second test non-equivalent to the first. This is why repeated measures designs often offer a poor substitute for a spatial control group.

\textbf{EXAMPLE}

Time-series designs may be quite strong, especially if the factor of theoretical interest is subjected to multiple independent tests. An example of this procedure can be found in a study of employment discrimination conducted by Claudia Goldin and Cecilia Rouse.\textsuperscript{101} We have already shown the potential of randomized experiments for analyzing the effects of employment discrimination in low-skill jobs. High-skill jobs offer a special obstacle to causal assessment because there are fewer positions, they are less standardized (and hence less comparable to one another), and the selection process is based on skills that are difficult to manipulate artificially, e.g., through audit or resume experiments. And yet, suspicion persists that a “glass ceiling” prevents the movement of women and minorities to the top of highly skilled occupations.\textsuperscript{102}

An opportunity for testing this hypothesis arose recently when a number of orchestras instituted blind audition procedures. Before entering into the specifics of the study it is worth considering that a classical orchestra is perhaps the ideal prototype of a skill-based occupation. All that matters, or should matter, is how one plays an instrument. Moreover, there are shared standards about what constitutes good playing in the field of classical music. (It is conceivable that aesthetics are race- or gender-based, but this is not the general impression.) Thus, from a certain perspective, the producers of “classical” music fall into a sector of high-skill occupations that are \textit{least} likely to exhibit discriminatory practices.

Goldin and Rouse exploit the change from non-blind to blind auditions in order to determine whether this shift in hiring practices has any effect on the propensity of women to attain positions in professional orchestras – where they were, and are, grossly under-represented relative to their presence in the general populace. The study gains leverage on the problem by looking closely at variation before and after the initiation of treatment, a point in time that varies from orchestra to orchestra. Specifically, they compare the probability of a female orchestral candidate passing various stages in the interview process (from the first audition to the final audition and job offer) prior to institution of blind auditing procedures and after the institution of blind auditing procedures. Data are collected for several decades prior to, and after, the change in hiring protocol. Thus, the analysis thus compares the success of female candidates in years prior to the change with their success in

\textsuperscript{100} Hamilton (1994).
\textsuperscript{101} Goldin and Rouse (2000).
\textsuperscript{102} England et al. (1988).
years after the change. Since the experiences of multiple orchestras are analyzed separately, this study may be understood as an interrupted time-series design, iterated for each orchestra under study.

The authors find that the existence of a screen separating the artist from the orchestral decision-makers (and thus concealing the gender of the player) increased the probability that a woman would be hired severalfold. This seems to prove the thesis that women face obstacles to upward mobility that are due to their gender only, not to job-relevant characteristics. Indeed, it is difficult to identify any possible confounder in this research design. Of course, the analysis does not illuminate precisely why this form of gender discrimination persists. But it does show the power of time-series designs for estimating causal effects, at least in some circumstances.

**TIME-SERIES CROSS-SECTION (TSCS) DESIGNS**

Having discussed the strengths and weaknesses of cross-sectional and time-series designs, we turn to a family of research designs that combines both sorts of comparisons—across units and through time. This will be referred to as a *time-series cross-section* (TSCS) design. Here, several observations are taken from each unit and there is variation in $X$ through time (at least in some units) and across units. The TSCS design combines temporal and spatial comparisons.

Typically, a TSCS design involves repeated measurements from a set of units, observed through time. Thus, one might observe individuals, or nation-states, every year for 20 years, establishing a panel with twenty observations through time. Sometimes, there is missing data, i.e., some units are not observed for every time-period; this generates an unbalanced panel.

Sometimes, observations through time are not of the same units but rather of randomly chosen units from a larger population. For example, if one is constructing a TSCS analysis from survey data composed of a sample of 2000 individuals drawn randomly from the US population every year over the course of 20 years, each annual sample (or panel) includes a different set of respondents. We shall refer to this sort of TSCS as a *pooled cross-section*.

TSCS includes a diverse family of research designs, as illustrated in the “miscellaneous” row of Table 8.4.

In order to focus our discussion we shall dwell on a simple variety of TSCS design known as the difference-in-difference (DD) design, illustrated in the second row of Table 8.4. Note that Group I receives the treatment while Group II exemplifies the control condition. Outcomes are measured prior to and after the intervention. Estimates of the causal effect derive from a comparison of the change in the outcome for the treated group with the change in outcome for the control group: $\Delta Y$ (Group I) - $\Delta Y$ (Group II). Hence, a difference-in-difference.\(^{103}\)

Suppose, for example, that we are comparing earnings for unemployed persons who (voluntarily) join a worker-training program with those who do not. We observe their earnings prior to starting the program (presumably, zero, unless we are measuring it several years prior) and again one year after the completion of the program. We then compare the change in earnings for the treatment and control group in order to estimate the causal effect (if any) if the program on earnings.

In some respects, this design looks very similar to a cross-sectional design. Yet, we are now looking at change over time between the pre- and post-tests rather than a simple post-test. To what extent does this make causal inference more secure?

\(^{103}\) Those who are interested in the DD estimator will find a short discussion in Gerring (2012b: 280-283).
In the cross-sectional design the assumption of causal comparability requires that all background factors that might impact the outcome be equal, on average, across the treatment and control groups. In the DD design, we require that all changing background factors be equal, on average, across the treatment and control groups. In most contexts, this is an easier assumption to satisfy. It is especially convincing if the time-period separating pre- and post-tests is relatively short and background factors change slowly. Under these circumstances, it seems plausible to suppose that causal comparability has been achieved.

Granted, any background factor that varies with \( X \) is still a potential confounder. So, if some subjects are propelled by a motivational change to (a) join the worker-training program and (b) get a job, this will generate a spurious result. But in other respects, the DD design is more robust than the corresponding cross-sectional or time-series designs. Threats from history (extant time-trends or regression to the mean) are not problematic so long as they affect both treatment and control groups equally. Circular confounding is less problematic because it is unlikely to affect the change in \( Y \) across treatment and control groups. Mechanistic confounders are still possible. However, because the number of potential confounders is much smaller in the DD design than in the corresponding cross-sectional or time-series designs, one is less likely to mistakenly condition on a post-treatment variable. In short, there is much to recommend the DD design (and its analogs in TSCS analysis) relative to cross-sectional and time-series designs.

One critical caveat must be inserted. Because the treatment is not randomized, one would not normally expect the treatment and control groups to be equal in all background characteristics that are relevant for the outcome. Specifically, we might not expect members of the control group to respond in the same fashion as members of the treatment group. Where self-selection is at work, those who choose to be exposed to a treatment are more likely to respond positively to that treatment. In this case, those who choose to attend a worker-training program may be in a better position to make use of that knowledge to find a job. Consequently, their gains from attendance may be greater than a typical member of the control group. It follows that a causal effect calculated with a TSCS design is often better understood as an average treatment effect for the treated (ATT) rather than an average treatment effect across the treatment and control groups (ATE).

**Example**
The effect on employment of minimum-wage laws is a primary topic in labor economics. Despite multiple studies and a great deal of theorizing, the empirical issue remains elusive. As with other social science questions, a key methodological obstacle is the non-random aspect of the treatment. States (or countries) that set high minimum wages are also likely to be different in other respects from states (or countries) that set low (or no) minimum wages. These heterogeneous factors, relating to other labor market regulations, fiscal policy, or the character of societies and labor organizations, serve as potential confounders. The institution of labor market reforms may also be a response to features of macroeconomic performance, introducing threats from circularity.

In a widely cited paper, David Card and Alan Krueger approach this problem by focusing on an episode of policy change: the rise in minimum wages in New Jersey in 1992. Their data-collection strategy focuses on a single sector – fast-food restaurants – that is likely to be sensitive to changes in the minimum wage. Several hundred restaurants in New Jersey, and a neighboring state, Pennsylvania, are surveyed to determine whether levels of employment, wages, and prices underwent any change before and after this statutory change came into effect. Pennsylvania restaurants thus serve as the spatial control group. (Comparisons are also drawn between stores in New Jersey which paid more than, and less than, the newly instituted minimum wage. Since the former were unaffected by the rise in minimum wages, this group forms a second control group.)
Card and Krueger’s empirical approach is a difference-in-difference model in which the change in fast-food employment in New Jersey (the treatment group) is compared with the change in fast-food employment in Pennsylvania. The analysis shows that changes in the two states through this period were quite similar, suggesting that a legislated rise in the minimum wage in New Jersey did not enhance unemployment in that state.

It is an impressive study, though – like all studies – not without potential difficulties. Questions might be raised, for example, about the representativeness of the chosen sector (is the total economy-wide effect of a minimum wage law reflected in the behavior of a single industry?). One also wonders about the tightness of the timeline (would the economic effects of an increase in minimum wages become manifest in the short space of eight months – the time elapsed between pre- and post-tests?). One wonders whether economic conditions in the two states were sufficiently similar to constitute a good paired comparison and whether the remaining differences were adequately modeled in the statistical analysis. It may also be questioned whether the research design incorporates enough power to constitute a fair test of the null hypothesis. (Has the positive hypothesis – that minimum wages affect labor market behavior – been given a fair chance to succeed?) At least one commentator has questioned whether the rise in minimum wages actually represents the factor of theoretical interest, or whether it should be regarded as an instrument for that underlying (un-measured) factor – wages actually paid to workers.104

Some of these issues might have been overcome by slight alterations in the research design; others are inherent by virtue of the fact that the treatment cannot be directly manipulated.105 The problem of non-random assignment haunts all DD designs (just as it does all other non-experimental designs). One can never be entirely sure, for example, that businesses in Pennsylvania (the control group) would have responded to a rise in minimum wages in the same manner as New Jersey. If not, then the generalizability of the finding is cast in doubt. One’s doubts about causality are amplified wherever the treatment is non-randomly assigned because one worries that there might be something about the assignment of the treatment – some unmeasured factor – that differentiates the treatment group from the control group, and accounts for their responses. More specifically, one worries that the rate of change in the outcome might differ across the treatment and control groups. If this is the case, then the chosen comparison-case (in this case, Pennsylvania) is not doing the job of an experimental control.

REGRESSION-DISCONTINUITY (RD) DESIGNS

A regression-discontinuity (RD) design, in its simplest and most typical form, looks like an experimental post-test only or cross-sectional design. One group receives the treatment and the other the control condition, and a single post-test measures the outcome of interest, as diagramed in Table 8.5. However, the details of RD design, i.e., how these groups come to be defined, are quite distinct.

Consider a worker-training program with a means-test. Only those who have been out of work for at least a year are eligible, and all such applicants are admitted. Let us imagine that this means-test is established after applications have been received (perhaps because of an unexpected budget cut). Thus, the program receives many applications from those who have been out of work for less than a year, and are thereby excluded from participation. This has the effect of establishing a clean cutoff between program participants and non-participants. While we can expect many background differences between the short-term unemployed and the long-term unemployed, we expect relatively minor differences between those out of work for 11-12 months and those out of work for 13-14 months. These groups should be similar in all respects that might impact the outcome of theoretical concern, earnings. Since one group receives the treatment and the other does not, we have a situation that resembles a true experiment with randomized treatment.

Methods of analysis for RD designs vary. One may approach the observations lying just above and below the cutoff as equivalent on all background characteristics; in this case, a simple difference-of-means test will suffice to measure the causal effect. Alternatively, one may include all members of the sample, downweighting observations lying further from the cutoff point. In our example, this would be accomplished by conditioning on length of unemployment spell. But the common
characteristic of the RD design is its exploitation of an arbitrary cutoff point separating treated and un-treated units.

Several potential weaknesses of the design should be noted. First, if the cutoff is not rigidly observed – if, for example, applicants who have been out of work for 11-12 months are admitted when members of the program staff judge that the applicant is especially needy – the comparability of observations lying on either side of the cutoff is likely to be compromised. Second, if the subjects in an RD design are aware of the cutoff rule they may be able to circumvent it, e.g., by lying on their applications. This will have the same effect, vitiating the comparison of treated/un-treated units. Third, there must be sufficient units lying just below and just above the cutoff point – or, alternatively, one must have great confidence in the covariates (e.g., length of unemployment spell) utilized to control for dissimilarities. Finally, since the analysis centers on units lying near the cutoff point it may not be possible to generalize findings to units lying far from that point, limiting the generalizability of a study. These points notwithstanding, the RD design is still very attractive. Of all the observational designs reviewed here it probably lies closest in spirit and in fact to a randomized design (i.e., a true experiment).

**Examples**

As an example, let us consider Richard Berk and David Rauma’s study of the California penal system. In 1978, California extended unemployment insurance to recently released prisoners, in the hopes of easing their transition to civilian life and reducing rates of recidivism. Former inmates were eligible only if they had worked a requisite number of hours while in prison, thus setting up a cutoff point that provides the basis for an RD design. Subjects become part of the sample if they actually apply for benefits, which means that the analysis compares those who apply and are eligible for benefits with those who apply but are ineligible (presumably because they were unaware of their ineligibility). The data model assumes the following simple form:

\[
Y: \text{Failure (re-incarceration)} = X: \text{Benefits (the binary treatment variable)} + Z: \text{Hours worked (the assignment criterion)} + C: \text{Control variables (background characteristics that might affect recidivism)}
\]

On the basis of this procedure, Berk and Rauma conclude that members of the treatment group experienced thirteen percent lower re-incarceration rates than those in the control group (i.e., those ineligible for the program), suggesting that providing post-prison assistance reduces recidivism by an appreciable degree.

A second example of the RD design is drawn from a recent study of labor unions. It is sometimes alleged that the formation of labor unions causes businesses to fail by imposing extra costs that cannot be recouped through enhanced sales, increased productivity, or changes in a firm’s pricing structure. In order to test this proposition, John DiNardo and David Lee examine the fate of over 27,000 American firms over the course of a decade and a half. Noting that unionization occurs as a partial product of a secret ballot election (the results of which are publicly available), the authors use this cutoff to conduct an RD design in which firms where a union vote narrowly wins are compared with firms where a union vote narrowly loses. They find that a successful union election scarcely affects firms’ subsequent survival rate; moreover, little evidence of a causal effect on employment levels, output, and productivity is discovered. (They also consider the possibility that the threat of a successful union drive may alter the wage structure, and hence a firm’s chances of survival, prior to the vote for recognition.)

---

Some uncertainty necessarily remains about the randomness of the above-cutoff/below-cutoff comparison, as the treatment is not truly randomized. In particular, one must be wary of circumstances where participants in a sample are aware of the consequences of a threshold, and able to self-select. For example, in the RD design conducted by DiNardo and Lee, where the effect of unionization on firm survival, employment, output, productivity, and wages is tested, we must count the possibility that the failure/success of unionization drives are not random with respect to the outcomes of interest. Consider that workers in a union representation election may be conscious of the potential effect of their vote on the financial health of the firm. Management often argues that a union will put the firm at a competitive disadvantage and lead, ultimately, to a loss of jobs. Under the circumstances, workers may be more inclined to support unionization if they are convinced of a firm’s strength, and less inclined if they feel that the firm is in a vulnerable position. If enough workers vote strategically on this basis, and if their hunches have some basis in fact (presumably they are familiar with their firm’s market position), then the results of this RD design speak only to local average treatment effects (LATE). That is, we may be willing to believe that firms that were/are unionized are no more likely to go bankrupt than firms that remain un-unionized, but we should not infer from this that unionization – if assigned randomly across the universe of firms – would have no causal effect on the probability of firm failure. (DiNardo and Lee are careful not to over-generalize from the limited data at-hand.)

### Table 8.5: Regression-Discontinuity (RD) Design

<table>
<thead>
<tr>
<th>Groups</th>
<th>(X=1)</th>
<th>(O_i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>II.</td>
<td>(X=0)</td>
<td>(O_{ii})</td>
</tr>
</tbody>
</table>

- \(O_{\text{NN}}\): Observations through time
- \(X=1\): Treatment condition
- \(X=0\): Control condition

### Instrumental-Variable (IV) Designs

We have seen that the assignment of a causal treatment is often subject to confounding whenever the assignment principle is not random. A partial solution to this problem can sometimes be found if at least one factor influencing assignment to treatment is not subject to confounding. This “instrument” offers the opportunity of a two-stage analysis, setting up an instrumental variable (IV) design.

Consider a worker-training program in which applicants self-select. A cross-sectional, time-series, or TSCS design is subject to potential confounding if the suspected confounders (e.g., motivation) cannot be measured and thereby conditioned. A regression-discontinuity design is not feasible because there is no cutoff for eligibility. However, among applicants, it is noticed that those living within easy travel distance to a worker-training center are more likely to matriculate and more likely to complete the program than those living at a further remove. Apparently, travel-time
increases the opportunity costs of attending. Fortuitously, it is not a factor that applicants are likely to be aware of when applying, as they do not know the precise locations of multiple centers where worker-training programs are held within a metropolitan area, or the travel-time required to reach them. Travel-time for each applicant can be measured easily using Google’s mapping algorithm, providing an instrument that predicts program participation.

Of key importance is that the chosen instrument does not affect the outcome directly and is not correlated with other factors (not controlled in the analysis) that affect the outcome. Travel-time must impact earnings only through the factor of theoretical interest, i.e., participation in the worker-training program. If, let us say, people living further away from worker-training centers are also further away from potential employers then the chosen instrument is subject to confounding and will render a biased estimate of the true causal effect. (In this case, it will bias the estimate downward, as those attending worker-training programs face a higher commute time to potential employers.) Let us stipulate that this requirement, sometimes referred to as an exclusion restriction, is satisfied.

In this situation, one can use the instrument $(Z)$, to establish a predicted value for the causal factor of interest $(X)$ that is free of confounding $(C)$. One then examines the covariation between $\tilde{X}$ and $Y$, controlling for any additional confounders that can be identified and measured. A causal diagram of the relevant assumptions is embedded in Figure 8.8.

**Example**

A recent influential application of instrumental variables addresses the classic question of long-run economic development. Why are some countries so much richer today than others? Acemoglu, Johnson, and Robinson (hereafter AJR) suggest that a principal factor affecting secular-historical growth rates is the quality of institutions, i.e., the strength of property rights. The methodological obstacle is that we have at our disposal no measure of institutional quality that is assigned in a fashion that is random with respect to economic development. Wealth and good institutions tend to go together. In order to surmount this difficulty, AJR construct the following causal story. Over the past several centuries, European colonial powers established strong property-rights protections in some parts of the world (e.g., North America) and not in others (e.g., most of Africa and Latin America). Schematically, they protected property rights in areas where large numbers of Europeans decided to settle and instituted “extractive” regimes in areas where Europeans were outnumbered by indigenous populations. This, in turn, was a factor of geographic circumstances such as the prevalence of tropical disease, which determined the likelihood of European survival in Africa, Asia, and the New World. Europeans settled, and thrived, where they had high survival rates. Estimates of varying mortality rates for European settlers in the course of the nineteenth century thus provide a suitable instrument for patterns of colonial settlement and, ultimately, for the quality of institutions that AJR presume the colonists are responsible for. This allows for a two-stage analysis, which may be simplified as follows:

\[
X = Z + B + U_1
\]

\[
Y = \tilde{X} + B + U_2
\]

where $X$=property rights measured in the late twentieth century (expropriation risk), $Z$=the instrument (European settler mortality), $B$=covariates (other causes of $Y$), $Y$=per capita GDP, $\tilde{X}$=the fitted values from Equation 1, and $U$=error terms for the two equations. (Intercepts are omitted.)

As with other corrections for non-randomized treatments, the IV technique is not without its difficulties. Indeed, the three assumptions outlined above seem rarely to be fully satisfied in empirical work. The chosen instrument, \( Z \), may be weakly correlated with the theoretical variable of interest, \( X \); \( Z \) may have an effect on the outcome, \( Y \), other than through \( X \); or there may be a common cause operating on both \( Z \) and \( Y \) (an unconditioned confounder). As with most modeling assumptions, these potential violations are difficult to test,\(^{108}\) and perhaps best viewed as theoretical priors. For example, in the study explored above, critics have suggested that a common cause—geography—attracts both settler mortality and current levels of economic development in ways not mediated by property rights.\(^{109}\) If this story about the data generating process is true then the chosen instrument is not valid. Even so, the two-stage analysis is probably more convincing than any conceivable one-stage analysis for this particular problem; in this respect, and to this extent, the IV approach is useful.

\[ \text{Figure 8.8: Instrumental Variable (IV) Design} \]

\[ \begin{align*}
Z & \rightarrow X \\
(C) & \\
X & \rightarrow Y \\
\text{Causal factor} & \\
\text{Outcome} & \\
\text{Instrument} & \\
\text{Confounder} & \\
\text{Not conditioned} &
\end{align*} \]

---


A few limitations of this design should be noted. First, the requirements of the analysis – diagramed in Figure 8.8 – are difficult to verify in most settings. Often, they are highly suspect on theoretical grounds. That is, one may wonder whether $Z$ affects $Y$ other than through $X$ or whether other (unconditioned) confounders are associated with the relationship between $Z$ and $X$ or $Z$ and $Y$.

Second, the IV analysis estimates a causal effect that relates to those units that are encouraged to receive the treatment due to the chosen instrument(s). This is referred to as a local average treatment effect (LATE), as distinguished from an average treatment effect. If the instrument ($Z$) is weakly correlated with the causal factor of theoretical interest ($X$), then the estimated LATE may have little practical or theoretical import. Moreover, it is rarely possible to identify the units encouraged into treatment by the instrument. Consequently, it may be difficult to say what sort of cases an IV result with a weakly correlated instrument would apply to.

It is easy to quarrel with many IV analyses found in the social sciences today. However, when the assumptions undergirding an IV analysis (as diagramed in Figure 8.8) are plausible, the IV approach to causal inference is often more compelling than other observational research designs that might be adopted.

**Conclusions**

In this rather complex chapter we have introduced five research designs for use in situations where a treatment is not randomized (or is imperfectly randomized) and a large number of observations are available for statistical analysis. A cross-sectional analysis relies on comparisons across units at a single point in time. A time-series analysis relies on comparisons through time. A TSCS analysis enlists both latitudinal and longitudinal comparisons. A regression-discontinuity design compares units on either side of an arbitrary cutoff, which distinguishes treatment and control groups. An instrumental-variable analysis rests on a factor that influences assignment to treatment but has no direct effect on the outcome.

Each of these designs attempts to re-capture the virtues of an experimental design. While this quest is noble, its achievement is a matter of doubt, as our discussion has shown. Even so, these designs often provide the best available method of analysis. Sometimes, an observational design is the only method that is practicable, or the only method whose results are likely to be generalizable to a larger population.
KEY TERMS

- Cross-section (CS)
- Time-series (TS)
- Time-series Cross-section (TSCS)
- Regression discontinuity (RD)
- Instrumental variables (IV)
- Common-cause confounder
- Omitted variable bias
- Collider
- Interrupted time-series
- Repeated observations
- Regression to the mean
- Difference-in-difference (DD) design
- Local average treatment effect (LATE)
- Estimator
- Exclusion restriction
- Spurious

INQUIRIES

1. What is a cross-sectional research design? What sorts of confounders does one need to be concerned with (typically)?

2. What is a time-series research design? What sorts of confounders does one need to be concerned with (typically)?

3. What is a TSCS research design? What sorts of confounders does one need to be concerned with (typically)?

4. What is a regression-discontinuity (RD) research design? What sorts of confounders does one need to be concerned with (typically)?

5. What is an instrumental-variable (IV) research design? What sorts of confounders does one need to be concerned with (typically)?

6. Of the above-mentioned research designs, which are most/least common?

7. Which research designs are generally strongest with respect to internal validity?

8. Why do advantages in internal validity not always carry over to advantages in external validity?
RESOURCES

REGRESSION-DISCONTINUITY (RD) DESIGNS


INSTRUMENTAL VARIABLE (IV) DESIGNS


OBSERVATIONAL DESIGNS, GENERALLY CONSIDERED

9. **Case Study Designs**

In the previous chapter we discussed observational designs (where the treatment is not randomly assigned) employed for use with large samples. In this chapter, we discuss observational designs whose purpose is to analyze a very small sample, typically one or several cases.

A **case** connotes a spatially delimited phenomenon (a unit) observed at a single point in time or over some period of time. It may be a political unit with a defined area of semi-sovereignty (e.g., empire, nation-state, region, municipality), organization (e.g., firm, non-governmental organization, political party, school), social group (as defined, e.g., by ethnicity, race, age, class, gender, or sexuality), event (e.g., foreign policy crisis, revolution, democratic transition, decision-point), or individual (e.g., a biography, case history). However defined, a case must comprise the type of phenomenon that an argument attempts to describe or explain. In a study about nation-states cases are comprised of nation-states (observed over time). In a study that attempts to explain the behavior of individuals, cases are comprised of individuals. And so forth.

A **case study** research design is an intensive study of a single case or a small number of cases that promises to shed light on a larger population of cases. The individual case(s) is viewed as a case of something broader, just as large sample analysis is also generally viewed as exemplary of a broader phenomenon. Both forms of analysis infer from a sample to a population, even though the sample sizes are very different.

Case study research may incorporate several cases. The latter is a defining characteristic of **comparative historical analysis**, associated with the work of Barrington Moore, Theda Skocpol, David Collier, and James Mahoney. However, as the sample of cases expands it becomes less and less feasible to investigate each case intensively. The case study format is thus implicitly a small-sample format.

To be sure, large samples of observations may be drawn at a lower level of analysis, i.e., from within the case(s) of interest. Commonly, a case study focused on a nation-state will include **within-case** analysis focused on individuals, perhaps drawn from a survey. These observations may be analyzed in any fashion, using any of the research designs introduced in the previous chapters.

Consider a case study of the democratic peace theory that rests on an intensive study of two (non-democratic) states that went to war. The cases of theoretical concern are nation-states. Yet, the researcher is likely to mobilize all sorts of within-case observations, some of which may be large in number (necessitating quantitative analysis) and others of which may be small in number (necessitating qualitative analysis).

It follows that case study research must enlist qualitative evidence at the level of the case because there are too few cases to allow for a statistical model. This is what makes it distinctive, methodologically speaking, and this is why one commonly equates case study research with qualitative analysis. However, **within-case** evidence may enlist small- and/or large-N samples. Consequently, case studies often combine qualitative and quantitative styles of research.

We begin this chapter by discussing case-selection strategies, each of which is associated with a specific research design. Next, we briefly touch upon rules of thumb for **within-case** analysis. A concluding section discusses some of the tradeoffs inherent in case study research.
In keeping with our focus in this section of the book we shall assume that the goal of the case study is to shed light on causal relationships. Readers should bear in mind that many case studies are descriptive in nature and thus omitted from this discussion.110

**CASE SELECTION**

Case study research begins with the selection of cases that will be intensively studied. The reader will recall (from Chapter 4) that the ideal selection procedure for large-sample research is to sample randomly from the population of interest (or to include the entire population). This generates a sample that is very likely to be representative, and therefore good fodder for theories that are likely to be true across that broader population.

In case study research, cases chosen for intensive analysis are usually chosen in a non-random fashion. *Purposive* case-selection enhances the likelihood that the chosen case(s) will provide leverage on the question of theoretical interest, even though it may also compromise the representativeness of the chosen sample.

Before entering the details of how this might be accomplished it is important to emphasize that case study analysis is never entirely restricted to the cases chosen for intensive analysis. Other cases always form a basis of comparison. These comparisons might be explored informally, through *shadow cases*. Or they might be explored explicitly by integrating data from a large number of additional cases in a cross-case model – a multi-method (aka mixed-method) approach to research. Thus, while case selection is important we want to emphasize that generalizations based on case study research rest on a larger empirical base than it might appear.

In Table 9.1 we set out a variety of case selection strategies.111 These are understood according to their function, which might be *exploratory* (to identify a hypothesis, $H_x$) or *confirmatory* (to assess $H_x$). For each general aim or specific function there are several viable strategies of case selection, indicated by bullet points in Table 9.1.

Column 2 specifies the number of cases in the case study. It will be seen that case studies enlist a minimum of one or two cases, with no clearly defined ceiling. Column 3 clarifies which dimensions of the case are relevant for case-selection. This may include the descriptive features of interest ($D$), the causal factor ($X$), background factors that may serve as confounders ($C$), and/or the outcome ($Y$). Column 4 specifies the criteria used to select a case(s) from a universe of possible cases.

In the following sections, we outline each case selection technique, offer one or two examples, and suggest ways in which cases might be chosen from a large population of potential cases. Before beginning, two general points should be underlined.

First, case selection criteria can usually be understood cross-sectionally or longitudinally (through time). For example, the test of “deviance” might be a case’s status at a particular point in time, or its change in status over an observed period of time. Usually, the latter is more informative for purposes of in-depth case analysis. A case that happens to be deviant may have undergone a change at some point in the past, or perhaps its deviance is a product of changes in other cases. By contrast, if we can ascertain the point of inflection – the time at which a case becomes deviant – we can focus subsequent study of the case on this time-period. This is usually more informative. Thus, wherever

---

110 See Gerring (2016a).
111 A longer list of strategies is presented in Gerring (2016a: ch 2).
possible, researchers should choose cases based on their values through time rather than their value at a particular point in time.

Second, virtually all case-selection strategies may be implemented in an informal “qualitative” fashion or in a formal, “quantitative” fashion. The qualitative approach involves reviewing potential cases known to the researcher according to the criteria listed in Table 9.1. The quantitative approach involves translating the criteria in Table 9.1 into an algorithm that can be applied to a large sample of potential cases. Case-selection algorithms are an excellent, and probably underutilized, tool for selecting a pool of potential cases. However, the algorithmic choice should probably not be followed slavishly in the final selection of cases, which usually depend upon details of the research context that are hard to represent in an algorithm.

Table 9.1: Case Selection Strategies

<table>
<thead>
<tr>
<th>Goals/Strategies</th>
<th>α</th>
<th>Factors</th>
<th>Criteria for cases</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Exploratory</strong> (to identify H₀)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>● Outcome</td>
<td>1+</td>
<td>Y</td>
<td>Maximize variation in Y</td>
</tr>
<tr>
<td>● Deviant</td>
<td>1+</td>
<td>C Y</td>
<td>Poorly explained by C</td>
</tr>
<tr>
<td>● Most-similar</td>
<td>2+</td>
<td>C Y</td>
<td>Similar on C, different on Y</td>
</tr>
<tr>
<td><strong>Confirmatory</strong> (to assess H₀)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>● Longitudinal</td>
<td>1+</td>
<td>X C [Y</td>
<td>X changes, C constant or biased against H₀</td>
</tr>
<tr>
<td>● Most-similar</td>
<td>2+</td>
<td>X C [Y</td>
<td>Similar on C, different on X</td>
</tr>
</tbody>
</table>

H₀ = causal hypothesis of interest. P(H₀) = the probability of H₀. X = causal factor(s) of theoretical interest. X → Y = apparent or estimated causal effect. Y = outcome of interest. C = vector of background factors that may affect X and/or Y.

**EXPLORATORY**

Frequently, the goal of a case study is to identify a possible cause, or set of causal factors, that may help explain an outcome of interest. Case selection thus hinges on the status of Y or Y & C – but not X (which is unknown at the start of the study). In this context, the case study plays an exploratory role. Two exploratory approaches to case selection are pervasive: outcome and deviant.

● **Outcome.** An outcome case maximizes variation on the outcome of interest. This may be accomplished by choosing a case that exhibits extreme values on Y (or ∆Y), which is then compared – formally or informally (through “shadow” cases) – with the norm. Thus, in a study of democratization one might examine a country that transitions quickly and completely. Likewise, studies of welfare state development often focus on the world’s largest welfare states, located in Northern Europe. Studies of war often focus on one of the two world wars. Studies of genocide often focus on the Holocaust. It is almost second-nature to identify phenomena by their most extreme manifestation.

When an outcome is conceptualized in a binary fashion, and one value (generally understood as the “positive” value) is especially rare, there are good reasons for choosing such cases for intensive analysis. Case studies of democratization generally focus on regime transitions – continuity being the more common condition. Case studies of war generally focus on wars – peace being the more common condition. Case studies of revolution generally focus on revolution – non-revolution being the more common condition. Note, again, that the rare value is – explicitly or implicitly –
compared with the normal value, either by comparing a case to itself through time or by comparing a case to other (less intensively studied) cases.

- **Deviant.** In many settings, what one knows about the background causes of an outcome can help identify useful cases. A deviant (anomalous) case deviates from an expected causal pattern. The purpose of a deviant case study is to explain the deviant case—and, in so doing, to explain other cases, providing a generalizable hypothesis about the phenomenon under study.

  For example, Cuba is the only country in the Americas that shows no signs of democratization at the present time, constituting an anomaly in the region. A study of this case might reveal something new about the causes of democratization.

  Another example is provided by Werner Sombart’s (1906/1976) classic study, *Why is There No Socialism in the United States*, often regarded as the first study of American exceptionalism. The US seemed deviant to Sombart at the turn of the twentieth century because, despite the advance of capitalism, it failed to exhibit the expected rise of a sizeable socialist party. In explaining this non-event, Sombart called attention to the consciousness of American workers, which seemed favorably disposed to capitalism and to the American system of government.

- **Most-similar.** Sometimes, one can identify cases that exhibit strong similarities on background conditions (C) but divergent outcomes (Y). The generic label for this research design is most-similar, aka Method of Difference.

  To apply this approach to the study of democratization one might compare neighboring countries that exhibit different regime outcomes. For example, in the 1990s Mali, one of the world’s poorest countries, introduced multi-party elections. Although threatened by an insurgency in the North of the country and a recent military coup, Mali’s democratization has so far survived. Other countries in the region such as Burkina Faso, Guinea, Mauritania, and Niger have taken no steps to open up political competition, despite sharing many geographic, cultural, and historical features with Mali. This sets up a most-similar comparison in which the question of interest is posed by this striking juxtaposition of different outcomes among similar circumstances.

**Confirmatory**

Case studies may also perform a confirmatory function—helping to confirm, disconfirm, or refine a hypothesis (garnered from the literature on a subject or from the researcher’s own ruminations) and identifying the generative agent at work in that relationship. Here, case-selection involves the causal factor of interest (X), background factors (C), and—sometimes—the outcome (Y). Common strategies may be classified as longitudinal or most-similar.

- **Longitudinal.** A longitudinal case study mimics a one-group experiment, where X changes in an as-if random fashion while C remains constant and Y is observed before and after the intervention.

  As an example, let us consider the “resource curse” hypothesis—that autocratic countries with natural resource wealth are less likely to democratize than similarly situated countries without resource wealth (Ross 2014). A longitudinal case would be a country where mineral wealth (such as oil) was discovered, and subsequently developed. The intervention may be understood as as-if random, at least in the contemporary era, insofar as mineral exploration is undertaken pretty much everywhere with similar technology. Likewise, investment resources are generally available thanks to the international markets that allocate capital and the international corporations that conduct research and development. Resources, once discovered, are apt to be exploited. So the assignment of the treatment is conditioned only by the international price of the commodity in question (e.g., oil), the state of technology (for research and development), and chance. A case study of this subject
would presumably focus not only on the outcome of interest (democratization) but also on the causal mechanisms thought to be at work in the relationship between resource wealth and regime transition.\footnote{A large-\(n\) crossnational version of this research design is conducted by Tsui (2011).}

- **Most similar.** A *most similar* design is similar to the longitudinal design with the notable addition of a control case – which (ideally) experiences no change on either \(X\) or \(C\).\footnote{Also referred to as the “comparative method” strategy (Lijphart 1975), discussed in Glynn & Ichino (2015).} For example, to explore the resource curse hypothesis in a most similar setting one might begin with a longitudinal case, adding a second case that exhibits similar features on background conditions but where there is no discovery of natural resource wealth.

**Omnibus Criteria**

In addition to criteria that allow us to distinguish among case-selection procedures, several additional criteria apply broadly to the selection of cases. Omnibus criteria include (1) **intrinsic importance**, (2) **data availability**, (3) **logistics**, (4) **case independence**, and (5) **sample representativeness**. We introduce these criteria last, though not out of lack of respect. Indeed, these features often outweigh the others. While the diversity of case selection procedures is emphasized in Table 9.1, case selection also has some generic features.

**Intrinsic Importance**

The selection of cases is often influenced by the (perceived) **intrinsic importance** of a case. Some cases – such as world wars, genocides, key inventions, revolutions – matter more than others because they have an obvious world-historical significance. Others matter because they are important for a specific group of readers. We presume that every social group or organization is interested in its own history, and this may justify the choice of cases. These sorts of case studies are sometimes described as **idiographic** (Eckstein 1975; Levy 2008; Lijphart 1971). However, most case studies have an idiographic feature or element, which is to say that the researcher is interested in saying something about the chosen case(s) as well as about some larger subject. It is only the latter that qualifies a study as a case study (following our definition), so we are not concerned with studies that are not generalizable in any fashion. Or, to put the point more finely, we are concerned only with the elements of a study that are generalizable.

**Data Availability**

Any case chosen for in-depth analysis must afford enough **data** to address the question of interest. If sources are unreliable, scarce, or for one reason or another inaccessible, the case is of little value. Recall that the purpose of a case study is to extend our knowledge beyond what it is possible to explore in a large sample. One important way that a case extends our knowledge is by providing information on things that we cannot measure, or cannot easily assemble, across a large number of units. This includes **within-case** evidence.

**Logistics**

The availability of data is, in turn, often a product of the researcher’s personal attributes – his or her language facilities, connections, and previous acquaintance with a region, time-period, or topic. We
assume that these logistical features are taken into account – implicitly if not explicitly – in any case-selection process. Sometimes, it is referred to as convenience sampling, though researchers may be loath to admit that they have chosen one case over another simply because it is easier to study the former than the latter. Nonetheless, if a researcher has special access to Site A, but not Site B, we should be grateful if she chooses A over B (so long as other criteria are not sacrificed). And we must acknowledge the fact that many cases find their authors, through some serendipitous process that could scarcely be predicted or replicated, rather than the reverse. Darwin did not select the Galapagos Islands from a universe of potential cases.114

**CASE INDEPENDENCE**

Chosen cases should, ideally, be independent of each other and of other cases in the population. If cases affect each other they are not providing independent evidence of the proposition. This may be referred to as Galton’s problem, interference, or a violation of SUTVA. (The exception would be a situation where interaction across cases happens to be the subject of investigation, as it would in a study of diffusion.)

**SAMPLE REPRESENTATIVENESS**

Although case studies are focused on one or several cases they also aim (by definition) to represent features of a larger population. Some cases are chosen primarily because of their presumed representativeness (typical cases) – though one should not imagine that typicality assures representativeness. For most case studies, representativeness is just one of several attributes of concern. Note that deviant cases perform their function – identifying new causes of \( Y \) – only if they are representative of a larger population. (If the result of a deviant case study is to develop an idiosyncratic explanation – pertaining only to the chosen case – it is not very useful.) Likewise, influential cases are likely to be dismissed – as lying outside the population of the hypothesis – if they are too idiosyncratic. (A necessary condition may be disproven by a single case only if that case is judged to be part of the population of interest, for example. Otherwise, scholars will rightly deal with the apparent exception by adopting tighter scope-conditions.)

**WITHIN-CASE ANALYSIS**

Once cases are chosen, they must be analyzed for clues into causal inference. Within-case evidence might be gathered longitudinally (the case observed over time) or from a lower level of analysis (“inside” the case). Evidence enlisted for this analysis might draw on any sort of source, e.g., primary sources, secondary sources, interviews, ethnography, or standardized surveys, as discussed in Chapter 13. And techniques for analyzing that evidence might be primarily qualitative or it might combine qualitative and quantitative techniques. In short, reaching causal inference with case study evidence is not so different from reaching causal inference with large-N samples of observational data, as discussed in Chapter 8. Do not imagine that you are entering a different universe.

However, one element is somewhat distinctive. Case study research must rely, at least in part, on qualitative evidence drawn from a single case or a very small number of cases. (Studies based on large samples may also incorporate qualitative evidence, but they need not do so; and when they do so the use of qualitative evidence is apt to be different from its use in case study research – where the data is focused, by definition. on one or several cases.)

Since this is the most distinctive feature of case study research it is important to spend some
time discussing how one might draw inferences based on this sort of data. Our suggested “rules of
thumb” cover several interrelated topics: the identification of a hypothesis or theory, testing
theories, analyzing temporal relations, and examining background assumptions.

IDENTIFYING A HYPOTHESIS/THEORY

Insofar a case study is generalizable it must center on causes that might conceivably apply elsewhere.
Idiosyncratic causal factors should not be major protagonists in the narrative.

The problem is that it is not always easy to discern which factors might be generalizable, and
the same cause may be differently framed. For example, a key event in the Spanish transition from
autocracy to democracy in the 1970s was the “Moncloa Pact” (so named because it was signed at the
prime minister’s residence by that name), which brought key actors – politicians, political parties,
and trade unions from differing ideological perspectives – into formal agreement on how the
economy might be handled during the transition. As such, it is a particular event with a particular
context. But it might also be framed as an example of a more general phenomenon – a political pact,
which some political scientists view as instrumental to the process of democratization. O’Donnell &
Schmitter (1986: 37) define a pact as “an explicit, but not always publicly explicated or justified,
agreement among a select set of actors which seeks to define…rules governing the exercise of power
on the basis of mutual guarantees for the ‘vital interests’ of those entering into it.”

In order to be considered as general causes a factor must be stated in a manner that has
plausible application to other cases in a larger population. At the same time, in order to make sense
of what actually happened in the case – including confounders, mechanisms, and alternate outcomes
– the researcher must also keep track of the specifics. Common nouns generalize, while proper
nouns particularize. Both play important roles in a case study. But their respective roles should not
be confused. Make clear – to yourself, and to readers – which factors are potentially generalizable
and which are not.

TESTING THEORIES

Once a hypothesis has been identified (an original hypothesis or one drawn from someone else’s
work), it is important to canvas widely for rival explanations. This canvas should include extant work
on the particular case under examination, general theoretical frameworks that might be brought to
bear on the subject, as well as your own intuition. Try to anticipate rival explanations that critical
readers of your work might construct.

For example, if the theory of interest focuses on the role of political pacts in democratization
then you must consider alternative narratives that might account for the trajectory of the case(s) you
have chosen. Alternate theories of democratization include modernization (more developed societies
are more likely to democratize), diffusion (countries in democratic neighborhoods are more likely to
democratize), polarization (ideologically polarized societies are less likely to democratize), and so
forth.

Where rival causes can be (convincingly) eliminated as factors in a particular case, this
elimination serves two important functions: (1) it eases concerns about potential confounders, and
(2) by the logic of elimination, it enhances the likelihood that X is the cause of Y. For example, if
Linz & Stepan can prove that democratization in Spain was not caused by diffusion this cannot
serve as a confounder and their favored explanation – resting on social pacts – becomes more
plausible.

To test a theory – your favored theory and rival theories – it is vital to examine all relevant
hypotheses suggested by the theory. For example, O’Donnell & Schmitter (1986: 38) write that pacts
are typically negotiated among a small number of participants representing established (and often highly oligarchical) groups or institutions; they tend to reduce competitiveness as well as conflict; they seek to limit accountability to wider publics; they attempt to control the agenda of policy concerns, and they deliberately distort the principle of citizen equality… At the core of a pact lies a negotiated compromise under which actors agree to forgo or underutilize their capacity to harm each other by extending guarantees not to threaten each others’ corporate autonomies or vital interests. This typically involves clauses stipulating abstention from violence, a prohibition on appeals to outsiders…

Each one of these assertions constitutes a hypothesis that may be affirmed or disconfirmed by looking closely at a case (assuming the evidence is available to do so).

Of course, it is not always easy to uncover specific, falsifiable hypotheses embedded in a theory, particularly if that theory is stated in very general terms. This may be facilitated if the theory is articulated in a formal model, but even formal models can be differently interpreted. In any case, for a theory to be falsifiable it must issue predictions about the world. And these predictions, or hypotheses, must be applicable to the case under investigation if the case is to adjudicate among rival explanations.

When testing those rival explanations treat them fairly – not as “straw men.” In order to dismiss a rival explanation it must be given a good chance to succeed. For example, do not dismiss a probabilistic theory with a single counter-example (a case or instance where the theory appears to be wrong). Think about the rival explanation as a proponent of that theory might think about it. This requires approaching your topic fresh, without all the baggage (preconceptions) you may have acquired as you developed your argument. What must be true about the case under investigation if the theory is true?

Often these hypotheses center on the causal mechanisms of the theory (M) – what lies “inside the box” between X and Y. For example, Michael Ross (2001: 327–328) suggests four mechanisms that might generate a relationship between resource wealth and autocracy:

A “rentier effect” . . . suggests that resources rich governments use low tax rates and patronage to relieve pressures for greater accountability; a “repression effect” . . . argues that resources wealth retards democratization by enabling governments to boost their funding for internal security; and a “modernization effect” . . . holds that growth based on the export of oil and minerals fails to bring about the social and cultural changes that tend to produce democratic government.

Case study analyses of the resource curse have focused on each of these putative mechanisms in an attempt to determine their veracity.

Another sort of test involves units at a lower level of analysis. For example, studies of the resource curse sometimes examine patterns across regions within a country. In this fashion, Goldberg, Wibbels, and Mvukiyehe (2008) focus on Texas and Louisiana, two American states that experienced revenue windfalls from oil resources, in order to take a closer look at the inter-relationship between resource wealth and democracy. They confirm that revenue riches lead to a decline in political competitiveness, and also discuss the possible causal pathways. In their words, “mineral rents provide cheap revenues that incumbent politicians use to purchase clientelistic support while keeping direct taxes on citizens low. It is this combination of low taxes and extensive public outlays that seems to contribute to politicians’ persistence in office” (Ibid. 506).

**Temporality**

To tease apart causal relations within a single case clues derived from the temporal ordering of events are essential. And to judge temporal relationships a chronological timeline of relevant events
is indispensable. Sometimes, authors construct chronologies in a list format. Sometimes, key events will be portrayed in a causal diagram. But always, there is an accompanying narrative explaining who did what to whom, and when. Juan Linz’s (1978) analysis of the breakdown of democracy in Spain during the 1930s involves a meticulous assemblage of historical facts including the changing composition of the legislature (disaggregated by ideological faction), the duration of cabinets, and the role of key actors in each event.

A chronology should begin before the causal factor of interest and extend all the way up to the outcome of interest, and perhaps beyond. Note that a qualitative analysis of X’s relationship to Y is in some respects like a quantitative time-series analysis. The longer the temporal relationship can be observed, the greater our opportunities for gauging X’s impact on Y, and identifying potential confounders. It is, however, unlike a time-series analysis insofar as one is unlikely to be able to observe X and Y throughout the whole period; or, it is irrelevant to observe them over a long period because they are not changing. In this setting, which often characterizes qualitative analysis, the relevant “covariational” features are the initial change of X and the eventual change in the outcome. What is left to observe are the factors that may have contributed to ΔX and ΔY. So, look for things that preceded X and things that lie in between X and Y.

The latter are the essential features of what has come to be known as process tracing, represented by M in this book. A good chronology includes all relevant features of M. It is complete. Unfortunately, it is not always apparent how to interpret completeness or continuity, i.e., which features are suitable for inclusion in a chronology and which features may be considered redundant. But one may assume that completeness exists when the connection between the events included in a chronology is tight – such that it is easy to see how a phenomenon evolved from step 1 to step 2, and step 2 to step 3, and difficult to see how any confounder could have disrupted that path.

The model of dominoes has served as a metaphor for the ideal. If one wishes to fully explain how the first domino is causally connected to the last domino one would want to construct a chronology of dominoes that includes each domino’s fall – the events leading up to, and causing, the outcome of interest. In this simple example, each domino serves a gateway function and there is only one pathway (flowing through all of these gates). Factors that seem to carry a “necessary” or “sufficient” quality should always be included in a chronology. When in doubt, the inclusion of ancillary details is not damaging, and certainly less damaging than the accidental exclusion of crucial details.

**Examining Background Assumptions**

A final goal of case study research – whether undertaken on its own (a stand-alone case study) or in tandem with a cross-case study (a multi-method study) – is to shed light on background assumptions. All causal models assume causal comparability. The expected value of Y must be the same for all observations in the sample, conditional on observables. For large-n samples, this is understood in a probabilistic sense: \(P(Y|X,C)\) is the same, on average, for all cases and for observations within a case (if the case is observed over time). For small-s samples, this is understood in a deterministic sense. Thus, any possible threat to this assumption for the studied cases is potentially damaging for the inference that the researcher wishes to draw.

Special attention to the assignment mechanism is warranted, as this is source of bias in many (most?) social science analyses. Where X is not intentionally randomized by the researcher – as it is not, by definition, in observational studies (a definitional criterion of the case study) – one must worry about assignment bias (or selection bias, as it is sometimes called). The blessing is that case studies are often especially insightful in providing insight into the assignment principle at work in a particular instance (Dunning 2008: ch 7).
For example, Jeremy Ferwerda & Nicholas Miller (2014) argue that devolution of power reduces resistance to foreign rule. To do so, they focus on France during World War Two, when the northern part of the country was ruled directly by German forces and the southern part was ruled indirectly by the “Vichy” regime headed by Marshall Petain. The key methodological assumption of their regression discontinuity design is that the line of demarcation was assigned in an as-if random fashion. For the authors, and for their critics (Kocher & Monteiro 2015), this assumption requires in-depth case study research – research that promises to uphold, or call into question, the author’s entire analysis.

As a second example, we may consider Romer & Romer’s (2010) analysis of the impact of tax changes on economic activity. Because tax changes are non-random, and likely to be correlated with the outcome of interest, anyone interested in this question must be concerned with bias arising from the assignment of the treatment. To deal with this threat, Romer & Romer make use of the “narrative record” provided by presidential speeches and congressional reports to elucidate the motivation of tax policy changes in the postwar era. This allows them to distinguish policy changes that might have been motivated by economic performance from those that may be considered as-if random. By focusing solely on the latter, they claim to provide an unbiased test of the theory that tax increases are contractionary.

**STRENGTHS AND WEAKNESSES**

Like all research designs, the case study method has characteristic strengths and weaknesses. These may be assessed relative to large-N cross-case analysis, the topic of Chapter 8. As such, researchers face a series of tradeoffs, summarized in Table 9.2, which should be understood as affinities rather than invariant laws. Exceptions can be found to each one. Even so, these general tendencies are often found in case study research and are reproduced in multiple disciplines and sub-disciplines over the course of many decades.

<table>
<thead>
<tr>
<th></th>
<th>Case study</th>
<th>Cross-case Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity</td>
<td>Internal</td>
<td>External</td>
</tr>
<tr>
<td>Research goal</td>
<td>Deep</td>
<td>Broad</td>
</tr>
<tr>
<td>Causal insight</td>
<td>Mechanisms</td>
<td>Effects</td>
</tr>
<tr>
<td>Population</td>
<td>Heterogeneous</td>
<td>Homogeneous</td>
</tr>
<tr>
<td>Variation in $X$ &amp; $Y$</td>
<td>Rare</td>
<td>Common</td>
</tr>
<tr>
<td>Data</td>
<td>Concentrated</td>
<td>Diffuse</td>
</tr>
<tr>
<td>Hypothesis</td>
<td>Generating</td>
<td>Testing</td>
</tr>
</tbody>
</table>

It should be stressed that each of these tradeoffs carries a ceteris paribus caveat. Case studies
are more useful for generating new hypotheses, all other things being equal. The reader must bear in mind that additional factors also rightly influence a writer’s choice of research design, and they may lean in the other direction. Ceteris is not always paribus. One should not jump to conclusions about the research design appropriate to a given setting without considering the entire set of criteria, some of which may be more important than others.

Case studies generally have stronger claims to internal validity than to external validity. It is difficult to claim representativeness across a larger population when a sample consists of only a single case or a handful of cases (see Chapter 4).

Relatedly, case studies are generally more useful when the population of interest is moderate in size (e.g., countries in South America [N=13]) and less useful when the scope of the research encompasses an enormous number of units (e.g., people in South America [N>400 million]).

Case studies are generally more effective in shedding light on causal mechanisms than in measuring causal effects. This is because the $X/Y$ relationship is generally defined in a way that can be accessed with a large sample, while the mechanism ($M$) connecting $X$ to $Y$ is often harder to observe and hence to test in a large sample.

Case studies are generally more useful when there is a great deal of heterogeneity in the population of interest. Consider that cross-case analysis is premised on cross-unit comparability (unit homogeneity). Cases must be similar to each other in whatever respects might affect the causal relationship that the writer is investigating, or such differences must be controlled for. Uncontrolled heterogeneity means that cases are “apples and oranges”; one cannot learn anything about underlying causal processes by comparing their histories. Case study researchers are often suspicious of large-sample research, which, they suspect, contain heterogeneous cases whose differences cannot easily be modeled. “Variable-oriented” research is said to involve unrealistic “homogenizing assumptions.” Under circumstances of extreme case-heterogeneity, the researcher may decide that she is better off focusing on a single case or a small number of relatively homogeneous cases. Within-case evidence, or cross-case evidence drawn from a handful of most-similar cases, may be more useful than cross-case evidence, even though the ultimate interest of the investigator remains on that broader population of cases.

Case studies are generally more useful when variation on key parameters is limited (e.g., revolution or school shootings) and less useful when variation on key parameters is ubiquitous (e.g., economic growth, election results). Rare events call for case studies, while common events militate toward cross-case analysis.

Case studies are especially useful when data is available only for one or several cases (even though the phenomenon itself may be ubiquitous). We might have good reason to focus on the history of Sweden or the United Kingdom if we wish to study long-term demographic developments, for the available demographic records in these countries is much longer than the corresponding statistics for most other countries. By contrast, if for all units the same quantity and quality of information is available, a cross-case analysis may be more informative.

Finally, and perhaps most importantly, case study research is generally more useful at an early stage of analysis, where a study serves an exploratory role – perhaps identifying a new outcome or a new causal factor of interest. Once a specific hypothesis has been identified the task is to test that hypothesis in a falsifiable fashion. For this purpose, large-N cross-case research is usually more useful.

In short, lots of things might incline one toward a case study analysis, and lots of things might incline one toward a large-N style of analysis, as summarized in Table 9.1. It all depends.

---

In this context, it is worth noting the current trend toward combining these styles of research in a single study, i.e., *multi-method* research (see Chapter 10). This may allow the researcher to take advantage of the virtues of intensive and extensive styles of research without suffering the corresponding vices. It follows that when considering the utility of the following case-selection strategies the reader should consider them as stand-alone research designs or as possible additions to a large-N research design, such as those introduced in Chapters 7-8.
KEY TERMS

- Case
- Case selection
- Exploratory
  - Outcome
  - Deviant
  - Most-similar
- Dispositive
  - Illustrative
  - Influential
- Estimating
  - Longitudinal
  - Most-similar
- Mechanistic
  - Pathway
  - Most-similar
- Case study
- Case study research
- Comparative-historical analysis
- Quantitative
- Qualitative
- Within-case

INQUIRIES

1. What is a case study?

2. What is case-study research? (What is not case-study research?)

3. Does case-study research generally enlist qualitative evidence, quantitative evidence, or both styles of evidence?

4. What are the strengths and weaknesses of case-study research?

5. What are the different strategies of case selection and their purposes?
RESOURCES

10. **DIVERSE TOOLS OF CAUSAL INFERENCE**

This section of the book has covered a lot of ground. We have presented a variety of causal frameworks (Chapter 5), defined the problem of causal inference (Chapter 6), and introduced an itinerary of research designs, which we classified as experimental (Chapter 7), large-N observational (Chapter 8), or case study (Chapter 9).

This short chapter summarizes some of the themes discussed in this long, and complex, part of the book. We begin by reviewing the problem of confounding, and the various types of confounding that impede causal inference. Next, we review the panoply of research designs discussed in previous chapters, illustrating their interrelationship in a tree diagram, and discussing whether there is a “best” research design. We conclude by discussion the multi-method approach to research.

**IDENTIFYING AND AVOIDING CONFUNDERS**

Causal inference is possible only if confounders can be avoided or controlled in the analysis. Recall that we adopted a very diffuse definition of confounding, including any factor that produces a spurious (or biased) association between $X$ and $Y$.

This idea may be rendered in a more concrete fashion using the language of causal graphs. A causal model will generally render a valid estimate if the “frontdoor” path from $X$ to $Y$ is unblocked and any existing “backdoor” paths from $X$ to $Y$ are blocked. Simply stated, a causal model should allow no covariation between $X$ and $Y$ except that which is a product of $X$’s causal impact on $Y$. This is another way of describing *causal comparability*, i.e., the expected value of $Y$ for a given value of $X$ is the same for all units throughout the period of analysis. (You will see that there are many languages by which to express the same general idea.)

Five types of confounders, introduced in previous chapters, are recapitulated in Figure 10.1. In each case, $C$ stands for the confounder, generating a backdoor path from $X$ to $Y$. Recall that in interpreting these graphs one must distinguish settings in which $C$ is conditioned from settings in which it is not (indicated by square brackets). In some settings, a confounder is created because a factor is *not* conditioned, in other cases because it *is* conditioned.

Panel (a) illustrates a *common cause confounder*, which has a causal effect on both $X$ and $Y$. Panel (b) illustrates a *compound-treatment confounder*, where $Y$ is affected by $X$ and $C$, whose effects cannot be distinguished from each other because they cannot be separately conditioned. Panel (c) illustrates a *circular confounder*, where $Y$ affects $X$ through $C$. Panel (d) illustrates a *mechanistic confounder*, where an intervening factor ($C$) between $X$ and $Y$ is conditioned. Panel (e) illustrates a *collider confounder*, produced by conditioning on a factor ($O$) that is affected by both $X$ and $Y$.
These confounders apply to any estimation strategy that relies on measuring the direct causal effect between $X$ and $Y$. The estimation of an unbiased causal effect generally depends upon unblocking the frontdoor path from $X$ to $Y$ (by refraining from conditioning on factors endogenous to $X$) and blocking the backdoor path(s) from $X$ to $Y$ (by conditioning on factors that block those paths, if they exist).

116 An alternative estimation strategy – the “frontdoor” approach – focuses on the path from $X$ to $Y$ through $M$, i.e., an indirect approach to causal inference. However, this approach remains quite rare in the social sciences and will not concern us here.
Figure 10.1: Typology of Confounders using Causal Graphs

(a) Common cause
(b) Compound treatment
(c) Circular
(d) Mechanism
(e) Collider

$X =$ Causal factor
$Y =$ Outcome
$C =$ Confounder
[] = Not conditioned
RESEARCH DESIGNS

There are many ways to investigate the relationship between $X$ and $Y$, i.e., to avoid confounders and to draw causal inferences. As a way of pulling together our knowledge on this crucial aspect of social science methodology, we shall look at a particular research question and various approaches that might be taken to it. Research design options are summarized in a taxonomic fashion in a tree diagram, as shown in Figure 10.2.

**Figure 10.2: Taxonomy of Causal Research Designs**

- **Experimental**
  - Post-test only
  - Pre-test/post-test
  - Multiple post-tests
  - Roll-out
  - Crossover
  - Factorial

- **Observational**
  - Large-N
    - Cross-sectional
    - Longitudinal
    - Panel
    - Regression-discontinuity
    - Instrumental-variable
  - Small-N
    - Exploratory
      - Outcome
      - Deviant
    - Diagnostic
      - Longitudinal
      - Most-similar
The research question is: what drives variation in earnings among working-age adults who experience long spells of unemployment? The hypothesis of interest (of relevance for research that is hypothesis-testing) is that participation in a worker-training program enhances earnings.

**Experimental designs (large-N).** *Assumption:* the researcher can assign the treatment across a large number of units.

- **Post-test only.** Randomize the intervention across groups, forming a treatment group (unemployed persons who attend the worker-training program or programs) and a control group (unemployed persons who do not), whose earnings are compared at some point after the intervention.
- **Pre-test/post-test.** Randomize the intervention across groups, forming a treatment group (unemployed persons who attend the worker-training program or programs) and a control group (unemployed persons who do not). Measure earnings before the intervention and after the intervention, comparing the change in earnings for the treatment group with the change in earnings for the control group.
- *(Other experimental designs listed in Table 7.1 might also be adopted.)*

**Observational designs (large-N).** *Assumption:* the researcher cannot assign the treatment, but X and Y can be measured across a large number of units.

- **Cross-sectional.** Compare the earnings of unemployed persons who attended a worker-training with those who did not, controlling for other factors.
- **Longitudinal.** Compare the earnings of unemployed persons prior to their participation in a worker-training program with their earnings after some period of time (when they may be employed).
- **TSCS.** Compare the change in earnings of (initially) unemployed persons who participated in a worker-training program with the change in earnings of (initially) unemployed persons who did not participate in a worker-training program, measuring earnings prior to the program start and at some point after the completion of the program.
- **Regression-discontinuity (RD).** Identify (or implement) a worker-training program that operates with an adjustable income threshold – a maximum income beyond which applicants are not accepted. Let us say that the maximum capacity for the program is 1000 and that 2000 or so apply. Applicants are prioritized according to income so that the poorest 1000 applicants are accepted and the rest put on a waiting list. (This means that the threshold cannot be known by participants in advance.) Now, compare the earnings (among those who applied to the program) of those whose income falls just below (and therefore attended the program) and above (and therefore did not attend the program) at some point after the program is completed. Alternatively, compare the change in earnings of these two groups.
- **Instrumental-variable (IV).** First, identify a feature that (a) enhances the likelihood of an unemployed person attending a worker-training program (X) and (b) has no impact on earnings and is not correlated with other things that might impact earnings. Second, analyze the probability of attending a worker-training program (X) using this instrument along with any additional exogenous factors that may affect earnings. Third, test the hypothesis of interest by regressing earnings on the predicted (fitted) value for X along with any other exogenous covariates.
Case study designs (small-N). *Assumption:* the researcher cannot assign the treatment and data is not available for a large number of units, or it is not conclusive.

- **Exploratory.** *Assumption:* one knows very little about why some people land well-paying jobs and others do not.
  - **Outcome.** Select a small sample of highly successful job seekers and try to reconstruct the causes of their success (enlisting background knowledge about those who were less successful).
  - **Deviant.** Select a small sample of highly successful job seekers who have characteristics that would seem to inhibit success, e.g., they are poorly educated, not strongly motivated, and from minority groups that often face discrimination. Try to reconstruct the causes of their success.
- **Diagnostic.** *Assumption:* one wishes to analyze a hypothesis about how worker-training promotes earnings.
- **Longitudinal.** Choose a worker-training program that initiates a change from a traditional method to the new method without any other changes to the program. Compare the attitudes and performance of administrators before and after the adoption of the new program.
- **Most-similar.** Choose several worker-training programs that are similar in background features, some of whom initiate a change from a traditional method to the new method. Compare those that adopt the new program with those that do not.

**AND THE WINNER IS…**

With these options before us, the question naturally arises: what is the *best* research design? This issue was raised initially in a very general fashion in Chapter 1. Here, we take up the issue in the context of research designs for causal inference.

Restricting ourselves to issues of internal validity the question has a simple answer. The best research design is one where the treatment has been randomized (i.e., an experiment) so long as various post-treatment confounders can be avoided or satisfactorily dealt with (ex post).

Recall that in experimental research the treatment is randomly assigned to treatment and control groups. So long as subjects are compliant, treatment and control groups are kept apart, and compound treatments are avoided, the study can generally claim to have attained causal comparability. That is, the expected value of Y conditional on X will be the same across treatment and control groups. In this fashion, an unbiased estimate of the causal effect can be obtained.

Naturally, it is not always possible to randomize a treatment of theoretical interest. And naturally, one is also concerned with the validity of a study with respect to a larger population (external validity). For these reasons, social scientists often resort to observational data. Indeed, observational studies continue to constitute the vast majority of social science work.

Among observational research designs, one rightly gives priority to those that successfully mimic the virtues of an experiment. That is, the treatment is assigned in a manner that may be regarded as random with respect to potential confounders. It is “as-if” random. This sort of setting may be referred to as a *natural experiment* or (more loosely) a *quasi-experiment*.

As an example, let us consider a study by Susan Hyde that attempts to determine whether election observers deter vote-buying and other elements of voter fraud. The 2003 presidential election in Armenia offers an ideal subject of analysis, for in this election international election
observers were allocated in a fashion that, she argues, is equivalent to a true randomized experiment. Hyde explains:

In this particular election, delegation leaders [from various international monitoring groups] gave each team of short-term observers a preassigned list of polling stations to visit during election day. These lists were made with two objectives in mind: (1) to distribute the observers throughout the entire country . . . and (2) to give each observer team a list of polling stations that did not overlap with that of other teams. These objectives were pragmatic in nature and unlikely to be correlated with any confounders. That is to say, the polling stations visited by international monitors were similar, in all important respects, to polling stations that were left un-monitored.

In order to measure the impact of these observers on the incidence of fraud, Hyde compares the incumbent vote in districts with observers to those districts without observers. The assumption is that the incumbent party is the principal instigator of vote-buying and intimidation. On the basis of this analysis, which follows a cross-sectional design, Hyde concludes that the deterrent effect of international observers is significant in both rounds of that election.

Let us return for a moment to the method of assignment. Although random-number generators were not used and Hyde herself had no control over the process, she argues that the choice of sites was orthogonal to (uncorrelated with) any possible confounders. Under the circumstances, it seems fair to regard this as a natural randomization with the desirable characteristics of a classic experiment. One element of doubt remains: because the experimenter does not control the assignment process she may never know for sure whether it is truly random, or just apparently random. This is why the distinction between a true experiment and a natural experiment is important to maintain.

Even under the best of circumstances, a natural experiment is not as convincing as a true experiment. More crucially, incidents of natural randomization such as the Armenian presidential election of 2003 are extremely rare. Alternatively, one might say that many situations appear to bear the marks of natural randomization but we simply do not know for sure and lack any means of decisively proving or disproving the point. Although one can often compare background characteristics, even if one finds that these characteristics are comparable across treatment and control groups it could be that other – unmeasurable – characteristics are not comparable. Likewise, if one finds that treatment and control groups are dis-similar on some background characteristics it could be that these characteristics are irrelevant to the $X/Y$ relationship of theoretical concern, and thus may be disregarded.

Because of this, the designation natural or quasi-experiment is likely to remain a vague moniker. All observational research – whether large-N or small-N – aims to capture the virtues of the true experiment, either by happenstance (as in Hyde’s case) or by the implementation of a procedure intended to restore random assignment to a non-randomly assigned treatment, e.g., by blocking covariates that might serve as confounders, by limiting analysis to some portion of a sample that is not subject to confounders, by focusing on longitudinal rather than cross-sectional variation, by focusing on observations that fall just above and below an arbitrary cutoff point (regression-discontinuity designs), or by isolating a feature that affect assignment to treatment but is not affected by confounding (instrumental-variable designs). If we accept that the chosen technique restores randomness to the assignment of treatment we are inclined to designate the research design as a natural experiment. If not, then not.

Consequently, the terms “natural experiment” and “quasi-experiment” must be treated with caution. They are only as convincing as the assumptions that surround a particular research setting.

and accompanying research design. Nonetheless, they focus our attention on a feature of causal inference that is absolutely critical. If the assignment process is non-random the association between $X$ and $Y$ is quite likely to be spurious and will need to be corrected.

But the goodness of a research design is not simply a matter of non-random assignment to treatment. There are many additional elements at work in the construction of a strong research design. This returns us to the main theme of the book. The best research design is the one that maximizes all the various criteria of causal inference (see Chapter 6). Best means best-possible, i.e., the best that could be achieved under the circumstances. The goodness of a research design is therefore assessable not by reference to some abstract standard but rather by reference to all possible research designs that might be devised to address the same research question. In some settings, this might be an experiment, in other settings it might a large-N observational design, and in other settings it might be a case study. It all depends.

Thus, rather than trying to arrange research designs in a neat hierarchy it is probably more helpful to consider which design offers the best fit for a particular causal hypothesis. A good instrumental-variable (IV) analysis is better than a flawed pathway-case study, just as a good pathway-case analysis is better than a flawed IV analysis. And what makes these analyses strong or weak is the extent to which the assumptions required for each research design are met in a given instance.

**MULTIPLE METHODS**

In selecting among research designs one can anticipate methodological tradeoffs. One method might be good at achieving internal validity but with questionable external validity. Another method might speak clearly to a larger population of concern but with questionable internal validity. This is just one example of the tradeoffs one may encounter in the work of social science.

In light of such tradeoffs a natural solution is to employ multiple research designs within the same study. Ideally, each design compensates for weaknesses in the others. This is known generically as *multimethod* (aka *mixed-method research* or *triangulation*).

For example, in pursuing the impact of worker-training programs one might construct an experiment to test the causal effect, coupled with close observation of the participants in order to understand the causal mechanisms at work. One might also construct a study that enlists nonexperimental data from extant worker-training programs to figure out whether the estimated causal effect from the experiment conforms to estimated causal effects from actually existing programs. Here, the experiment sits front-and-center but is complemented by other designs.

Now consider a question that is probably more suited for observational research: the impact of economic development on democratization. Here, the point of departure is perhaps a TSCS study of countries in which one searches to discern a relationship between economic development, proxied by per capita GDP, and regime-type. This might be complemented by case studies of countries that have undergone rapid economic development in order to try to understand the mechanisms that may be at work in this relationship. One might also construct experiments – or take advantage of natural experiments – in which some facet of economic development is manipulated in order to perceive its effect on political life or political attitudes. For example, one might compare regions of a country targeted for economic development with other regions not targeted for growth, observing the way in which regional politics evolves. This is a strong research design if ceteris paribus conditions are upheld, i.e., if everything else is more or less equal across the chosen regions.
Often, the most convincing approach to solving a causal problem enlists multiple methods, each of which fulfills specific goals. The sum of these different approaches may add up to a more convincing explanation than is possible when a single method is employed in isolation. Of course, conducting multi-method research is time-consuming and demands a lot of the researcher, who must cultivate diverse methodologies. Sometimes, multi-method approaches are best undertaken by multiple people, each of whom brings specialized skills to the project. Alternatively, one can think of the accumulated literature on a subject as presenting a multi-pronged approach insofar as each study brings different tools to the question.

Medical research provides a good example. Typically, research on a disease begins with the identification of a set of symptoms, which clinicians are able to classify as a particular disease-type (new or old). This spurs the collection of observational data, drawn initially from sources that are readily available. Clinicians work intensively with patients suffering from the disease, administering treatments and noting responses as best as they can (the case study method). They may also study subjects who are exposed to the disease but do not contract it (a deviant-case approach to case study research). Epidemiologists collect data from populations, attempting to identify larger patterns from a statistical analysis encompassing many data points. Biologists work at the micro-level, seeking to identify disease patterns within the cell (a form of within-case analysis, focused on causal mechanisms). Eventually, from one of these research streams, a potential treatment is devised and researchers construct an experiment to test it – first with animal subjects and then, if the results are promising, with human subjects.

In this fashion, medical research benefits from diverse methodological approaches. Each is based on distinct technical and substantive expertise, often housed in different disciplines, e.g., medicine, biology, chemistry, public health, and medical anthropology. Despite the diversity of methods, they take part in a single conversation about a topic – say, HIV/AIDS – which, over time, often leads to progress on that problem. Methodological diversity does not pose an obstacle (or at least, not an insuperable obstacle) to problem-solving.\textsuperscript{118}

Unfortunately, this sort of multimethod community of scholars does not exist in many corners of social science. Because research programs are generally uncoordinated (leading to different choices across the various dimensions of research design), and because scholars are entrenched in different research communities (publishing work in different journals and often not citing members of a rival methodological camp), work conducted in different modes does not always cumulate. Instead, it simply accumulates. However, we should still regard the multimethod research community as a goal to aspire to, and the field of medical research offers a good exemplar of what might be achieved in other areas.

\textbf{CONCLUSIONS}

Causality has been a popular topic for applied and theoretical work in social science, statistics, and philosophy in recent years. Many insights have arisen from this work, as reflected in the text and in the set of references listed at the end of each chapter in Part II. Anyone hoping to make sense of contemporary work in the social sciences will need to grapple with the wiles of causal inference.

At the same time, one must bear in mind that causal knowledge is not the only sort of knowledge that one might acquire about the social world. Other forms of knowledge – i.e., descriptive, predictive, normative, and prescriptive, as outlined in Chapter 2 – are also valuable.\textsuperscript{118}

\textsuperscript{118} The preceding paragraphs draw on suggestions from Evan Lieberman (personal communication, 2009).
These styles of analysis often accompany causal analysis; indeed, description is a pre-requisite for any causal argument. However, description is not only a precursor for causation. Sometimes, we want to know what $X$ is – rather than (or in addition to) what causes $X$ or what effects $X$ has on something else. We close this section of the book with a reminder that good social science is not reducible to causal knowledge. And good methodology is not reducible to methods of causal inference.
KEY TERMS

- Multimethod research
- Natural experiment

INQUIRIES

1. In what sense might it be argued that there is a hierarchy of research designs? What is this hierarchy (from best to worst)?

2. What is a natural experiment? Why are these sorts of research designs difficult to discern (and contentious)?

3. Multimethod research seems to offer an appealing solution to the limitations of various methods. What does it mean? Is it a good solution, and a realistic solution?

4. Is causality the only way to understand social reality? What’s the alternative? Are there limitations to a causal ordering of the world?

RESOURCES

NATURAL EXPERIMENTS


MULTIMETHOD RESEARCH

Examples:

In this section of the book we consider the process of research, as it is conducted, and the presentation of research, once it has been conducted. Chapter 11 discusses how to read works of social science (i.e., books and articles produced for an academic audience). Chapter 12 offers advice on how to find a topic for a research paper. Chapter 13 lays out various strategies for gathering data. (Since this is a large topic this is an especially long chapter.) Chapter 14 discusses the task of writing. Chapter 15 addresses the task of public speaking. Chapter 16 offers a short discussion of ethical issues associated with the research process.

As you can probably appreciate, these tasks are closely interwoven. Reading social science will help you develop a topic for your research and will help you develop a sense for how to construct your own project. Likewise, writing social science will help you appreciate the strengths, and realize the weaknesses, of other studies. Public speaking is another form of communication, closely linked to writing. Sometimes, a good way to write is to express an idea out loud (or sotto voce); sometimes, it is best to write it out first. There are lots of synergies. So these chapters – and their associated activities – are best approached together.
11. READING AND REVIEWING

This chapter is devoted to reading and reviewing the social science literature on a topic. Here, we discuss how to distinguish social science sources from other sorts of work, how to locate and survey the literature on a chosen topic, how to read strategically, how to read critically, how to figure out complex arguments, how to construct a systematic review of the literature on a topic, and how to take notes as you go along. These are closely linked topics so you will find a good deal of overlap across these sections of the chapter.

As you read social science you may be struck by the dry tone of the literature, especially in journal articles. (Books usually attempt to include some divertissement to relieve the tedium of pages and pages of prose.) Remind yourself that what is exciting about science – any science – is getting closer to the truth. The mode of exposition is secondary, and generally remains in the background. Indeed, it is important to adopt a dispassionate tone to discourage other factors from interfering with the logic of the argument. This is why authors generally avoid personal anecdotes, jokes, heavily symbolic or allegorical language, and other narrative devices.

Of course, there is a person behind the prose, and perhaps it would aid the cause of science if some personal elements – such as his/her motivations for studying a subject – were made explicit. But, for better or for worse, this is not the accepted scientific mode of communication in most social science fields. Bear with it.

In any case, new media are opening up new modes of communication, many of which are more personal in nature and less tightly structured. On virtually any given topic you can now find blogs (discussed below) or lectures and debates preserved on Youtube, as well as other multi-media presentations. The medium is changing, though the austere format of journal articles is likely to remain the workhorse of social science for the foreseeable future.

IDENTIFYING SOCIAL SCIENCE SOURCES

Reading social science presupposes that one can identify works of social science from the mass of other sources out there in print and on the worldwide web. This brings us back to an earlier discussion. What is it that distinguishes social science from other genres (such as casual conversation, journalism, or partisan rhetoric)?

In Chapter 1, we stipulated the following. First, social science involves the systematic application of reason and evidence to problems with explicit attention to method and to possible sources of error. Second, it is accompanied by realistic estimates of uncertainty with respect to whatever conclusions are drawn from the evidence. Third, it attempts to provide a comprehensive treatment of a subject within whatever scope-conditions are defined by the study. All relevant information should be included; none should be arbitrarily excluded. Finally, social science adopts a disinterested posture with respect to all goals except the truth. Its purpose is to get as close to the truth as possible, in all its complexity, rather than to provoke, entertain, elucidate moral truths, or advance polemical claims.

Studies with these characteristics generally appear in academic venues, i.e., in scholarly journals and working papers, in academic reports, and in books published by university presses. The
audience for this sort of work is comprised of academics, policymakers, and members of the public with a special interest in the subject matter. Occasionally, a work of scholarship crosses over to a broader audience. Jared Diamond’s classic *Guns, Germs, and Steel* broke through the usual boundaries separating scholarly and popular work. However, this is fairly rare, and likely to remain so. The features that define social science also limit its appeal to a broader audience.

It may seem as if we are defining the genre of social science in a circular fashion: social science writing is what social scientists write and read. This is generally the case, but not always. Social scientists have many opportunities to write for the general public outside the genre of social science. They may write editorials, blog posts, speeches, manifestos, or articles for the popular press. They may tweet. They may even compose short stories, novels, and films. Each of these genres offers a somewhat different set of opportunities and constraints. Each has distinctive rules. However, while these forms of writing may deploy aspects of social science and may lean heavily on social science research they do not fall into the genre of social science as we usually understand it. Just because work includes the phrase “social science studies say…” does not make it social-scientific. In order for it to be an example of social science the study must possess the characteristics listed above.

Consider the case of Paul Krugman, a Nobel laureate who has become a prominent liberal pundit, weighing in on all manner of political debates in the United States. His scholarly work, published in economic journals and academic presses, falls clearly into the genre of social science. His polemical work appears in newspapers and on-line, and occasionally in books. While the latter draws on academic research it is of a very different character.

Usually, readers will have no trouble discerning popular and academic genres – though occasionally they are somewhat blurred. In any case, it will be helpful to briefly review the variety of venues that are available as you research a topic. These include (a) journals, (b) books, and (c) working papers and blogs.

**JOURNALS**

Every academic discipline boasts a handful of top journals that help to define that field or subfield. You should become familiar with these journals, as they will help you to identify important work in your area of research. Sometimes, this literature is rather technical and may not be easily accessible to those without advanced training. One of the purposes of this chapter is to assist in making sense of scholarly language – though it is important to bear in mind that some of the background knowledge required to process an academic journal article is, by its nature, highly specialized and therefore impossible to survey in a general textbook such as this.

**BOOKS**

Among books, there is a well-trodden distinction between scholarly monographs – produced primarily for academic audiences – and commercial presses, who produce for a lay audience. Traditionally, academic publishers were closely associated with (and heavily subsidized by) universities. Thus, we find Cambridge University Press, Oxford University Press, Princeton University Press, Harvard University Press, and countless other presses bearing the name of a university in their title. These presses continue to publish virtually all the scholarly monographs that appear each year. By contrast, commercial presses (all the others) rarely publish books addressed to a scholarly audience unless they are written for the (lucrative) textbook market.

The imprimatur of a book – by a scholarly or commercial press – is thus an important clue.

---

to its contents. However, it is less useful than it used to be. The reason for this is that academic presses are under pressure to publish books with crossover potential (i.e., books that might reach a general readership). Likewise, academics are less motivated to publish traditional academic monographs, as the task of scholarly communication is increasingly performed by journals. In most fields, books are now regarded as an occasion to reach a broader audience (and not coincidentally, to make some money). Thus, for a variety of reasons fewer academic monographs appear every year, and the books that do get published by academic presses are more likely to paint on a large canvas with an orientation to the general reader.

WORKING PAPERS AND BLOGS
Because of the slow pace of academic publishing most studies appear initially as working papers. They may be posted on a conference web site, on the author’s homepage, on the web site of a sponsoring organization or an aggregator like the Social Science Research Network (SSRN). Unpublished papers are of varying quality. Of course, published papers are also of varying quality, but the variance is greater in papers that have not passed the threshold of peer review. Bear in mind that some authors post papers at a very early stage, while others wait until a paper is ready to publish before making it available to the world. You must be the judge.

With respect to blogs the situation is even more confused. Some blogs are written for other academics and consequently employ all the acronyms, pet phrases, and technical verbiage that distinguish a field or subfield. They presume a lot of “inside” knowledge and thus may be less accessible to outsiders than journal articles, where a formal style of presentation is adopted and where acronyms are explained and points are fully referenced. Nonetheless, academic blogs play an increasingly important role in scholarly debate, and sometimes they go down a lot easier than a turgid journal article. A short list of widely used academic blogs is appended to the end of this chapter, for illustrative purposes. They may or may not offer a good entrée into the scholarly literature on a subject you are researching. But they are certainly worth a try.

More generally, the benefit of drawing upon unpublished papers and blogs is that you are able to get closer to the frontiers of knowledge on a topic. The drawback is that those frontiers are wild and untamed. Here, you will find undigested ideas, weakly supported claims, and brainstorm ideas that may never gain acceptance in the academy. Because there is no official process of vetting (with the exception of a few working papers or conference-series papers that operate like journals), you will need to decide upon the quality of the work and whether it merits inclusion in your survey of a subject. One clue is the academic standing of the author, which you may be able to judge by his/her academic position, prior publications, or overall reputation.

GENERAL OBSERVATIONS
By way of conclusion to this discussion, let us return to the central distinction between “academic” and “popular” genres of writing. It should be clear that we are not casting aspersions on the latter. We have greater admiration for Martin Luther King’s Letter from Birmingham Jail (one of the world’s most inspired polemics against racism) than Gunnar Myrdal’s An American Dilemma (a learned monograph on the same subject). However, this chapter is not about how to write jeremiads against injustice. It is a chapter about how to read and write social science. As such, Myrdal’s book is relevant and King’s letter is not (except as a primary source). Our purpose is not to subsume or belittle other realms of discourse but rather to improve the practice and understanding of social science discourse. As a first step, it is important to understand the content and the boundaries of the

120 Myrdal (1944).
genre known as social science.

**SURVEYING THE LITERATURE**

Having defined social science – at least in a rough-and-ready way – we arrive at the problem of identifying the literature on a particular topic. This used to be an extraordinarily time-consuming task undertaken at the library, where one would laboriously page through hard-copy indexes such as the *Reader's Guide to Periodical Literature* and the card catalog, followed by a hunt through the stacks for relevant books and articles (which would then need to be photocopied). Things are more efficient nowadays. Sometimes, a trip to the library is still required in order to get advice from a librarian or to obtain a book or article that is unobtainable in electronic form. But a good deal of the work can be conducted with a computer, providing one has access to e-journals, e-books, and specialized databases. (Usually, access to these proprietary sources must be obtained through a library or university.)

Four computer-assisted approaches are generally helpful for this task: (a) recent publications, (b) search terms, (c) citation searches, or (d) other venues. As you pursue them, you may find it useful to maintain an updated list of references and/or a folder on your hard-drive where you keep PDF copies of relevant articles and reports.

**RECENT PUBLICATIONS**

The best way to start your literature review is to identify a book or article that includes a survey of the relevant literature. Of course, all studies will include citations to the literature. But some are more useful for this purpose than others. Especially useful are books – if they contain a chapter that reviews the literature and an extensive bibliography – and review articles. A good place to look for the latter are journals that specialize in literature surveys (see Resources).

A thorough literature review on your topic vastly simplifies your task. Indeed, it may be sufficient, unto itself. However, one must bear in mind that a literature review is only as current as its publication date. If it is several years old, you will be obliged to comb through the literature in subsequent years. Likewise, be aware that the person conducting the review has a point of view, like everyone else. Their review of the literature may be undertaken in order to show that some aspect of the literature on a topic is weak or biased. This may be true, or it may not. In any case, the author may neglect other aspects of the subject since they are less relevant to his/her argument. You should not assume that a literature review covers all aspects of a question. This will require further work. However, a search through the references of a literature review should provide a good jumping-off place.

**SEARCH TERMS**

A second approach to surveying the literature focuses on key words. Think about different terms that articulate various aspects of your topic. These are useful search terms which may be plugged into Google Scholar, Web of Science (which includes the Social Sciences Citation Index), or some other database that is perhaps more focused on your area of concern (see Resources). Some topics are nicely summarized in a set of key terms; these terms return items associated with that topic and only that topic. Unfortunately, many topics do not have this characteristic. When key terms are entered into a search engine they return all sorts of irrelevant material, or no relevant material. You may have to experiment a bit with different search terms before you find the right word, or the right
combination of words.

**CITATION SEARCHES**

A third approach is contingent upon identifying at least one scholarly work that bears directly on your topic. If it is a classic work, one that defines the topic in question, you are in luck. This may be entered into Google Scholar, Web of Science, or some other database and should return all relevant material published after that classic work. The reason for this is that classic works on a topic are likely to be cited by other work on that topic.

**OTHER VENUES**

A final approach involves a grab-bag of other web-based venues. You may find a Wikipedia article on your topic. While Wikipedia articles are of varying length and quality, many are carefully written and (even if they are not) they often contain a smattering of references to the relevant literature. Likewise, you may find websites maintained by scholars or hobbyists to be useful. If a subject is frequently taught, you may locate syllabi from those courses. This is a good way to identify entry-level readings, though rarely is a course syllabus detailed enough to satisfy a literature review.

**GENERAL OBSERVATIONS**

How widely should one survey? Sometimes, a topic generates work in multiple fields. It would be difficult to restrict one’s survey of the literature on social capital or democracy to a single discipline. Other topics are restricted to a single discipline or sub-discipline. Worker-training programs are solidly within the purview of labor economics, though the topic has repercussions in sociology, education, and public policy.

How far back in history should one survey? In principle, the cumulative nature of science obviates the reading of older works, since their arguments and findings should be subsumed by more recent studies. As it happens, the cumulative character of science is more apparent in the natural sciences than in the social sciences. In the social sciences ideas tend to be re-cycled with each generation, perhaps with a slight change of terminology or theoretical focus. Gaining a sense of this intellectual history is important if one is to distinguish cycling from progress. Likewise, reading the history of a subject may open new windows into that topic, allowing one to see the “obvious” in a new and perhaps illuminating way. A deep reading of the literature on a topic is not simply a means of extracting classic quotations for your epigraph (though it performs that function as well).

Once you have identified a body of work, try to identify the most relevant studies, i.e., those that are most influential (as judged by number of citations or by what other authors say about them) and those that bear most directly on your particular topic. This smaller set of articles and/or books bears close reading.

**READING STRATEGICALLY**

By virtue of your own background research (above) or by virtue of a specific class assignment, you now have an article or book to read. How shall you go about this?

Bear in mind that social science is not a sphere of activity in which close reading of texts is always advised. Occasionally, a topic is so dominated by a classic work that anyone working in that field is obliged to master the details of that classic work. But social science is not like literature or philosophy, where texts are intended to be dissected and where each turn of phrase merits attention.
Sometimes, a close reading bears fruit. However, the time invested in a close reading must be weighed against the time that might be spent on other activities — such as skimming another hundred articles on the same general topic.

You must learn to read **strategically**. Typically, readers pick and choose which aspects of a study they are interested in or how deeply they wish to engage a particular work. In order to determine how you should approach this book or article you need to have a clear idea of what you hope to get from it.

Let us briefly survey some possibilities. If you are constructing your own review of the literature it may be sufficient to read the abstract of an article or the back-jacket blurb of a book. Perhaps all you need to know is what the study is about and what its main findings are. If the study in question contradicts your own argument or your own prior beliefs about a subject, you may wish to look closely at the evidence and the author’s interpretation of that evidence. If you are reading the study for a class then you need to think about how much detail it is necessary to remember, and what sort of issues your instructor might wish you to focus on. Some classes are focused on substantive issues, others on methodological issues, and so forth. If you are asked to write a review of, or make a presentation based on, a study, then obviously you will need to understand that study in fine-grained detail. In short, there are many purposes for reading social science and each will entail a different level of engagement.

Regardless of your strategic purpose, it will be helpful to know a bit about how social science studies are put together. This is easiest with articles, which tend to follow a uniform template. (Books are more variable, though much of what we have to say is also applicable to scholarly monographs.)

Most scholarly articles are accompanied by an **abstract**. The abstract lays out the general subject, the argument, and the evidence — all in the space of a single paragraph or several short paragraphs. This is immensely useful and probably bears reading, even if you intend to read the innards. Often, an article is accompanied by a few key words; these are useful for purposes of searching, as discussed above.

The **introduction** of an article generally repeats the abstract, with additional verbiage and citations to the extant literature. It lays out the importance of the subject matter, says a bit about work that has already been conducted on it, and the author’s contribution to that literature. Finally, there will be signposting — an outline of what is to follow.

The **theory** section lays out the argument, usually combined with a more in-depth engagement with the literature. Sometimes, the theory will be presented in narrative format as well as in a formal (mathematical) model.

This is usually followed by an **empirical section(s)**, which presents the evidence. Typically, this is the lengthiest part of a social science study, and may be supplemented by an **appendix** where descriptive statistics, specialized issues, or robustness tests are dealt with in greater detail. (Nowadays, this appendix is likely to appear on-line somewhere else rather than as part of the published article.)

The **conclusion** of an article typically repeats the main findings, comments on the limitations and broader ramifications of the study, and points out directions for further research. This is the formula for a social science journal article. The utility of a common format is that you can easily skim a social science journal article, knowing where to look for what you wish to find. Naturally, there are exceptions. Some articles do not have a well-defined “theory” section; they proceed quickly to the evidence. Some articles do not have a well-defined empirical section; their emphasis is on building theory. But most articles assume the foregoing format, which is explored at greater length in Chapter 14.
Social scientists are generally pretty smart, and the work that gets published in top journals and top university presses is generally of high quality. However, it is not flawless. And sometimes, it is very flawed. In any case, you need to read critically. This does not mean that you should try to knock down every study you encounter. It means that you should learn to identify the limitations of each piece of research. Sometimes, authors are forthright and will point out these limitations in their discussion. Sometimes, these limitations or contradictions will be hidden or mentioned only obliquely. (The author may not even be aware of them.) Suffice to say, in order to be an intelligent consumer of social science you need to cultivate a critical eye.

While there is no recipe for deconstructing social science, a few standard questions should be asked of any study. With respect to the theory, one might ask:

- Is it clearly stated?
- Does it address an important problem?
- Is it original, or in what ways is it original? (Or does it tell us something that we already know, or that has already been established by prior research?) Does the author properly credit other work that contributes to the theory? What body of work is the author arguing against?
- Is it properly bounded? (Are the scope-conditions clear, and defensible?)
- Is it specific enough to be falsified?
- How should we classify the theory? Is it, for example, primarily descriptive or causal?
- If the argument is causal, what general causal framework does it invoke?
- If the argument is causal (and if the research is non-experimental), what is the envisioned counterfactual? (What is the expected variation in X and what impact does this variation have on Y?) Has the author correctly identified the causal mechanism(s) at work?
- Are there underlying assumptions that must be satisfied for the argument to hold?

These questions elaborate on points laid out in Chapter 2 (for all theories) and Chapter 6 (for causal theories).

With respect to the research design, one might ask:

- How would one characterize it, in general terms? (Is it experimental or observational? Large-N or small-N? If the former, is it properly classified as a TSCS, cross-section,…?)
- Does it fit well with the theory? Does it constitute an easy test or a hard test?
- Are there potential problems with the research design? Is the author’s proposed account of the data correct? Is there another way to account for the observed pattern of data? Is the research design prone to stochastic error or to systematic bias? Are there potential confounders that are not effectively controlled for in the analysis?
- Is the sample representative of a defined population? To what larger population are the results generalizable?
- Are there better research designs for addressing the author’s research question?
- Is the finding original? How does it fit with other work on the subject? If it is at variance with other work, which finding is most credible? Can these research streams be reconciled?

These questions bring us back to fundamental issues of research design, as introduced in Chapter 4 (for all research designs) and Chapters 6-10 (for causal analysis).
FIGURING IT OUT

As you read the social science literature you will occasionally get stuck. That is, you will be unable to figure out what an author is trying to communicate, or you have a general idea but are not sure that you are correctly interpreting the text.

Sometimes, this is because the text itself is not clear and thus open to various interpretations. (This touches on a leitmotif of this section of the book – the importance of communication skills in social science.) Sometimes, it is because the text is written for readers who possess a good deal of background knowledge about the subject or the method.

A common stumbling block is unfamiliarity with technical terms. Once upon a time it was important to have a lexicon – a specialized dictionary – at your side, with key terms in your field or subfield defined. Nowadays, you can usually solve this problem by Googling the term and finding the appropriate Wikipedia article or other on-line source. Of course, these on-line sources are of varying quality, so you must exercise judgment. Note also that key terms can mean quite different things in different contexts. Reading several definitions can usually help to sort this out.

Another common stumbling block is a mathematical formula or model that is beyond your present capabilities. Unfortunately, learning enough algebra to understand a formal theoretical model or a proof, or learning enough probability theory to understand a statistical model, is a time-consuming venture. These are matters best learned in the structured context of a course in math or statistics (something you might want to consider as you plan your future coursework).

You may take comfort in the fact that this aspect of an article or book is (generally) carefully vetted by expert reviewers. It is rare to find mathematical errors in a published work, especially if it is a top-notch journal or press. Problems are much more common in interpretation. What does the theoretical or statistical model mean? How does it bear on the theory and the evidence? What is the data generation process (the process by which the data the author is analyzing has come to be)? Does the chosen statistical model correctly describe that data? Is the research design appropriate for the task?

Of course, knowing the math will help you resolve these sorts of questions. However, you can still address those questions by reading the author’s (verbal) explication of the model. If the point at issue is a statistical model that is not well-explained by the author you may wish to consult Part IV of this book, or take a look at a stats textbook or an on-line reference source (suggestions are included in Part IV).

REVIEWING THE LITERATURE

We have now discussed how to identify a work of social science, how to identify studies of a particular topic, and how to read those studies strategically and critically. A remaining task is to review that literature in a systematic fashion.

There are essentially three versions of the literature review. The first involves combining results from multiple quantitative analyses of the same subject – published separately – in order to determine the overall causal effect of a variable, taking sampling error and other sorts of error into account. This is known as a meta-analysis, and is beyond the purview of the book.

A second sort of review is prefatory to an original piece of research (usually conducted with primary sources). Its purpose is to situate that research within a broader literature, showing how the author’s work provides an original contribution to (some aspect of) that literature. This will be
discussed in Chapter 14.

The third sort of review is a general review of the literature on a topic, where the author’s sole purpose in the paper is to survey that literature. This is equivalent to the sort of article one might find in any of the “Annual Review” journals (listed at the end of this chapter). You might wish to peruse one of these as an example of how to conduct such a survey of the literature.

Literature reviews of this sort are commonly assigned in seminar courses, as they can be conducted without access to primary sources and provide an opportunity for students to integrate a large amount of material, perhaps spanning the entirety of the course. They may serve a number of purposes. For example, they may serve to illustrate how scholars have conceptualized a topic, what research designs they have employed, and what data sources they have used. They may serve to identify recent trends in research or the trajectory of a subject over a longer period of time (its intellectual history). They may serve to identify inconsistencies and controversies within the literature, perhaps offering a reconciliation of divergent positions. They may serve to identify omissions in scholarly work on a topic, areas that are relatively untouched by serious research, or areas that are fraught with methodological problems. They may identify hidden biases in a subject. They may identify hypotheses that bear further research, thus pointing the way to future work on a subject.¹²¹

Note that a literature review is also an original contribution to that literature, for it synthesizes that literature in a unique way, pointing out strengths and weaknesses or new approaches that may not have been apparent. Occasionally, a cogent literature review becomes a standard reference, reorienting a field in a new direction.

Think carefully about your overall argument before you begin, and modify it as you go along, as necessary. A plausible place to begin is by defining the scope of the review. Like research with primary data, the literature review must have clearly established scope-conditions. It must be bounded in a logical fashion. This is equivalent to establishing the population of studies that the review is intended to encompass. For example, you must decide the temporal scope of the review. Many reviews limit their purview to recent work, e.g., work published over the past two decades. Others are more ambitious, intending to uncover the intellectual history of a topic. You must also decide what sort of literature you wish to include: Published and/or unpublished material? Articles and/or books? Scholarly and/or popular sources? All social science fields, or just one or two? Other scope-conditions relate to the topic itself, which may be defined broadly (e.g., “democracy”) or more narrowly (e.g., “democratization in Latin America”).

Having defined the purview of the review, you must now integrate the literature that falls within that purview. This means painting a broad picture, citing as many studies as possible while discussing the details of only a few. These few should be chosen carefully. Usually, they exemplify a general feature that one finds in the literature; they are examples of a broader trend. Alternatively, they may offer an unusual – and especially fecund – line of inquiry, opening up a new approach to the subject. In any case, a good literature review summarizes a vast literature with little expenditure of prose. It is synoptic. That is, it reduces the subject to a few fundamental parameters.

These parameters might be theoretical. For example, one might argue that the literature on democratization employs two very different types of explanations in the effort to explain why some countries democratize and others do not (or why they democratize at different times). The first sort of explanation is structural, focusing on distal causes of democracy such as geography, colonial heritage, social inequality, and modernization. The second sort of explanation is proximal, focusing on short-term causes or mechanisms such as elite bargains, social movements, civil conflicts, military interventions, and the sequences by which different political institutions develop over time. This

¹²¹ The foregoing list builds on Johnson and Reynolds (2008: 187).
distinction is fundamental, and allows one to categorize all studies as part of one or the other camp. Another way of dividing up the subject is methodological. One might point out that studies of democratization can be classified into three approaches: country case-studies (including comparative historical analysis), subnational studies, large-N crossnational studies, qualitative comparative analysis (QCA), and experimental studies.\footnote{Most of these approaches are reviewed in Coppedge (2012).}

One might also divide up a subject by looking at various ways of operationalizing its key concept(s). Thus, one might distinguish among studies of democracy that utilize binary measures of democracy (e.g., the Democracy/Dictatorship index), ordinal measures of democracy (e.g., the Freedom House indices), or interval measures derived from latent-variable analyses (e.g., the Unified Democracy Scores\footnote{http://www.unified-democracy-scores.org/}). If one is exploring country and regional studies of democracy, one might divide up the subject by geographic area – Latin America, Africa, Middle East, Asia, and so forth.

There are many ways to analyze any given subject. This is what distinguishes one literature review from the next, though both may be focused on the same subject. Different authors often identify different lines of cleavage, calling attention to different aspects of a topic.

**TAKING NOTES**

As you read, it is a good idea to take notes. Some readers take notes on whatever they are reading, regardless of whether they have a clearly designed purpose for the notes, ex post. For these readers, taking notes is a form of diary-keeping.

However, most readers are strategic in their note-taking. If one’s purpose is to prepare for a class discussion or an exam one’s notes will consist of points that are relevant to those tasks. If one’s purpose is to construct a paper – an original piece of writing – one’s notes should assist in that process. In either case, taking notes enhances your engagement with the text. It makes reading an interactive process. Thus, regardless of your immediate purpose in reading a text (whether it is to prepare for a discussion, an exam, or a paper), use note-taking to allow your mind to roam freely.

Nowadays, there are many ways to take a note. You may mark up a hard copy. You may mark up a PDF or Kindle text. You may insert notes in a text file. Or you may employ various citation managers (BiblioExpress/Biblioscape, Endnote, Mendelay, ProCite, RefWorks, Reference Manager, Zotero), most of which also allow you to maintain notes on each source in the database. Importantly, if you are using a text file to take notes make sure you have a good system for keeping track of which idea or quotation comes from which source (and what page in that source). Any sloppiness in your record-keeping is likely to lead to plagiarism if the notes are later incorporated into a paper. (See section on plagiarism in Chapter 14.) Also, make sure to mark clearly which ideas are copied verbatim from a source (within quotation marks), so they can be distinguished from ideas that you have paraphrased based on your reading of a source (suitably transformed so that they do not reflect the author's original rendition). In order to distinguish ideas that are clearly your own from those that need to be footnoted, you might want to preface the former with your initials. This will help when you return to your notes, whether this is a week later or several years later.
CONCLUSIONS

Social science offers an important set of resources on a wide range of topics. It is not the last word on these topics, and rarely is it the first word. However, social science sources are almost always worth consulting.

Consider the alternative. If you rely entirely on journalistic work on a topic – e.g., articles appearing in newspapers, popular journals, books produced for a general audience, or a nonacademic web site – you will be reading someone else’s digest of the social science literature on that topic. An article on crime will tell you what the author knows, based on his or her perusal of the literature and perhaps some interviews with academics who have conducted research on the topic. But the author of such an article is unlikely to have a deep knowledge of the scholarly literature, and – given space-limitations – must produce a brief and schematic report.

This is probably not good way to make up your mind on a complex subject. Without delving into the scholarly literature you will not be in full possession of the facts, even if – especially if – those facts are disputed.

Unfortunately, the literature of social science is not always readily accessible to the non-specialist – hence, the widening gap between social science and popular writings on most topics. This chapter is intended to help bridge that gap. We have discussed how to distinguish social science sources from other sorts of work (of a less scientific nature), how to locate and survey the literature on a chosen topic, how to read strategically, how to read critically, how to figure out complex arguments, how to construct a systematic review of the literature on a topic, and how to take notes as you go along. We hope that these tips will prove useful to you as you survey the literature on topics of your choosing.
KEY TERMS

- Meta-Analysis

INQUIRIES

1. Select a keyword or passage relevant to a social science topic. Conduct a Google search (or a search using some other general-interest search engine). Identify the first 20 hits that are relevant to your topic (eliminating those that are off topic). Then, classify the work or web page as predominantly (a) academic or (b) popular. Next, assign a score for each work or web page according to its status as an academic source, with 10 being the most reliable and 0 being the least. Use the criteria identified in the chapter to decide your scoring of each source. How reliable, and how helpful, is this method of source evaluation? How difficult is it to decide on a source’s “scientific-ness”? What are some problems or questions that arose?

2. Identify a topic and then pursue several methods of arriving at a source list for a topic, as described in the chapter. Which method worked best, and why? Would you expect the same outcome if you had chosen a different topic?

3. Choose an article or book in your field and give it a close, and critical, read. Now write a short review of that publication. What is the main argument? How does it compare and contrast with other work on this topic? What are the strengths and weaknesses of the author’s approach? (Is the author up-front about these?) In what ways might the author have improved his/her study? What areas call for further research?

4. Identify a topic that is not too broad, e.g., the role of colonial legacy in democratization or the impact of worker-training programs on employment prospects. Now put together a bibliography comprised of academic sources on that topic. The bibliography should be as comprehensive as possible, but should not include extraneous sources. (This could be a timed, in-class exercise.) An extra twist (requiring a good deal more time): write annotations to each source, indicating how it relates to the topic. Now, reflect on your experience. What elements of this exercise were easy, and which were hard?

5. With that same topic, or another (not too big, or too small), write a literature review, following the guidelines in the chapter. If other reviews have been written on this topic, make sure that you distinguish your own in some fashion (e.g., a different scope, a different angle, a different conclusion). What, for you, was the most challenging aspect of this exercise? Do you think it was helpful in identifying possible topics for original research?

6. Identify a prominent debate within a subfield of interest to you. What is at the foundation of this debate? Are they based on different methods? Different interpretations of the evidence? Different research sites? Different understandings of the theory? Different theoretical frameworks or traditions?
RESOURCES: JOURNALS AND BLOGS

This section includes journals that publish literature reviews or are oriented to the general public (non-specialists) as well as blogs, and is organized by field or subfield.

ANTHROPOLOGY

**Journals**
- Annual Review of Anthropology
- American Anthropologist
- American Ethnologist

**Blogs**
- Anthropology blog newspaper (http://www.antropologi.info/blog/)
- Anthropology Now
- Anthropology Works
- Grant McCracken (economics/anthropology)
- Savage Minds: Notes and Queries in Anthropology
- Tabsir (http://tabsir.net/) (Islam and Middle East)

ECONOMICS, BUSINESS/MANAGEMENT

**Journals**
- Annual Review of Economics
- Annual Review of Financial Economics
- Annual Review of Resource Economics
- Journal of Economic Literature
- Journal of Economic Perspectives

**Blogs**
- Conscience of a Liberal (Paul Krugman)
- Econcomix: Explaining the Science of Everyday Life
- Freakonomics
- Marginal Revolution

EDUCATION

**Journals**
- Review of Research in Education
- Review of Educational Research

**Blogs**
- Chronicle of Higher Education Blogs
- Dianne Ravitch
- Early Ed Watch (New America Foundation)
- Edunwok
• Inside Higher Education
• The Educated Reporter

ENVIRONMENT
Journals
• Annual Review of Environment and Resources
Blogs
• Dot Earth (NY Times)
• Environmental Economics
• Grist
• Real Climate: Climate Science from Climate Scientists
• ScienceBlogs.com Environment Channel
• The Oil Drum: Discussions about Energy and our Future

INTERNATIONAL DEVELOPMENT
Journals
• Journal of International Development
• Development and Change
Blogs
• Aid Thoughts
• Center for Global Development (CGD); http://www.cgdev.org/section/opinions/blogs/
• Chris Blattman
• From Poverty to Power
• Owen Barder

LAW
Journals
• Annual Review of Law and Social Science
• Journal of Law and Politics
Blogs
• Balkinization
• Brian Leiter’s Law School Reports
• IntLawGrrls: Voices on International Law, Policy, Practice
• PrawfsBlawg
• SCOTUS Blog
• The Faculty Lounge: Conversations about Law, Culture, and Academia
• Volokh Conspiracy

POLITICAL SCIENCE
Journals
• Annual Review of Political Science
• Perspectives on Politics [oriented to the general reader]
• PS: Political Science and Politics [oriented to the general reader]
Blogs
• Duck of Minerva
• Freethinking HQ
• Fruits and Votes (Matthew Shugart)
• Monkey Cage
• The Mischief of Faction

PSYCHOLOGY
Journals
• Annual Review of Clinical Psychology
• Annual Review of Psychology
Blogs
• Bering in Mind (Scientific American)
• Mind Hacks
• PsyBlog: Understand Your Mind
• The Situationist
• The Splintered Mind: Reflections in Philosophy of Psychology

PUBLIC HEALTH
Journals
• Annual Review of Public Health
Blogs
• Health Wonk Review
• Public Health Perspectives (The Public Library of Science)
• Shots: Health News from NPR
• The Public’s Health

SOCIAL WORK
Blogs
• Classroom to Capitol
• Musings of a Social Work Academic (Martin Webber)
• Political Social Worker: Politics Through a Social Work Lens
• The New Social Worker

SOCIOLOGY
Journals
• Annual Review of Sociology
• Contexts [oriented to the general reader]
Blogs
• Everyday Sociology
• Mobilizing Ideas
• Orgtheory.net
• Politics Outdoors
• Scatterplot
• The Society Pages: Social Science that Matters

URBAN PLANNING
Journals
• Cities: The International Journal of Urban Policy and Planning
Blogs
• Christopher Leo: Research Based Analysis and Commentary
• CoLab Radio (MIT Community Innovators Lab)
• Planetizen: The Urban Planning, Design, and Development Network
• Polis: A Collaborative Blog About Cities Across the Globe
• The Global Urbanist

RESOURCES: DATABASES

GENERAL DATABASES
• Academic oneFile
• Academic search premier (Online)
• Applied social sciences index and abstracts (Online)
• EconLit
• Humanities & social sciences index retrospective
• Index to legal periodicals & books (H.W. Wilson Company : Online)
• PAIS
• Social sciences citation index (Online)
• Social sciences full text (Online)
• Worldwide political science abstracts
• WorldCat local

SPECIALIZED DATABASES
• Annual reviews
• AP images
• Biography and genealogy master index
• CIAO Columbia international affairs online (includes think tank papers)
• Conference proceedings citation index social science & humanities
• Current index to statistics
• ICPSR Inter-University Consortium for Political and Social Research
• IMF eLibrary data
• International financial statistics (Online)
• LexisNexis academic
• Library PressDisplay
• National Bureau of Economic Research (NBER)
• OECD iLibrary
• ProQuest dissertations & theses (PQDT)
• ProQuest statistical datasets
• World news connection

DATABASES FOCUSED ON TOPICS RELATED TO THE UNITED STATES

• American National Election Studies ANES.
• Congressional bills
• Congressional committee prints
• CQ researcher (Online)
• Oxford encyclopedia of American political and legal history
• National Journal
• ProQuest Congressional
• ProQuest Digital U.S. Bills and Resolutions 1789-2013
• Roper Center For Public Opinion Research
• State Attorney General reports and opinions
• U.S. Attorney General opinions
• U.S. presidential library
• Vanderbilt Television News Archive.
12. **BRAINSTORMING**

How does one go about identifying a fruitful research question and, ultimately, a specific research hypothesis? This is the very early exploratory phase, when one quite literally does not know what one is looking for, or at. Arguably, it is the most crucial stage of all. Nothing of interest is likely to emanate from research on topics that are trivial, redundant, or theoretically bland — no matter how strong its research design components are.

Methodologists generally leave this task to the realm of metaphor – bells, brainstorms, dreams, flashes, impregnations, light bulbs, showers, sparks, and whatnot. The reason for this lack of attention is perhaps because beginnings are inherently unformulaic. There are few rules or criteria for uncovering new questions or new hypotheses. Methodologists may feel that there is nothing — nothing scientific, at any rate — that they can say about this process. Karl Popper states the matter forthrightly. “There is no such thing as a logical method of having new ideas,” he writes. “Discovery contains ‘an irrational element,’ or a ‘creative intuition.’”

Accordingly, what we have to offer in this chapter is more in the character of a homily than a framework. It reads like an advice column. We urge you to Study the Tradition, Begin Where You Are, Get Off Your Home Turf, Play with Ideas, Practice Dis-belief, Observe Empathically, Theorize Wildly, and Think Ahead.

This advice is largely commonsensical and by no means comprehensive. It cannot help but reflect our own views and experiences, though we have drawn extensively on the writings of other scholars. Nonetheless, it may help to orient those who are setting out on their first journey, or who wish to begin again.

**STUDY THE TRADITION**

Consider the state of the field on a topic. What are the frontiers of knowledge? What do we — collectively, as a discipline — know, and what don’t we know? We doubt if anyone has happened upon a really interesting research topic simply by reading a review of the extant literature. However, this is an efficient method of determining where the state of a field lies and where it might be headed.

In exposing oneself to the literature on a topic one must guard against two common responses. The first is to worship those that have gone before; the second is to summarily dismiss them. Respect the tradition, but don’t be awed by the tradition either. Try stepping outside the categories that are conventionally used to describe and explain a subject. By this we mean not simply arguing against the common wisdom, but also thinking up new questions and new issues that have not been well explored.

As you peruse the literature, be conscious of what excites you and what bothers you. Which issues are under-explored, or badly understood? Where do you suspect the authorities in a field are wrong? What questions have they left unanswered? What questions do you find yourself asking when you finish reading? Where does this line of research lead? Sometimes, typically in a conclusion or a review article, scholars will reflect upon the future direction of research; this, too, can be useful.

Sometimes, it is necessary, due to time-constraints, to limit oneself to the recent literature on a subject. However, time permitting, there is something to be gained by going deeper, delving into the “classics,” i.e., the founding texts of a field or subfield. This is useful because it sometimes prompts one to think about familiar subjects in new ways, because classic works tend to be evocative (raising questions), and because it is a reminder that some things have, in fact, been done before. Every subject has an intellectual history and it is worthwhile to familiarize yourself with this history.

As C. Wright Mills began his study of elites, he consulted the seminal contributions of Lasswell, Marx, Michels, Mosca, Pareto, Schumpeter, Veblen, and Weber. In commenting upon this experience, Mills reports,

I find that they offer three types of statement: (a) from some, you learn directly by restating systematically what the man says on given points or as a whole; (b) some you accept or refute, giving reasons and arguments; (c) others you use as a source of suggestions for your own elaborations and projects. This involves grasping a point and then asking: How can I put this into testable shape, and how can I test it? How can I use this as a center from which to elaborate – as a perspective from which descriptive details emerge as relevant?

Not every topic is blessed with such a rich heritage; but some are, and there it is worth pausing to read, and to think.

BEGIN WHERE YOU ARE

Charles Sanders Peirce points out, “There is only one place from which we ever can start … and that is from where we are.” The easiest and most intuitive way to undertake a new topic is to build upon what one knows and who one is. This includes one’s skills (languages, technical proficiencies), connections, life-experiences, and interests.

Hopefully, your chosen topic resonates with your life in some fashion. This is often a source of inspiration and insight, as well as the source from which sustained commitment may be nourished and replenished over the life of a project. C. Wright Mills writes,

You must learn to use your life experience in your intellectual work: continually to examine and interpret it. In this sense craftsmanship is the center of yourself and you are personally involved in every intellectual product upon which you may work. To say that you can 'have experience,' means, for one thing, that your past plays into and affects your present, and that it defines your capacity for future experience. As a social scientist, you have to control this rather elaborate interplay, to capture what you experience and sort it out; only in this way can you hope to use it to guide and test your reflection, and in the process shape yourself as an intellectual craftsman.

125 Snyder (2007).
127 Kaplan (1964: 86), paraphrasing Charles Sanders Peirce.
Because the business of social science is to investigate the activities of people, any personal connections we might have to such people – your subjects – may serve as useful points of leverage. It will be helpful if you can establish a personal connection – however distant – with the phenomena you are studying.\footnote{Gadamer (1975) refers to this as a fusion of horizons – us and theirs (the actors we are attempting to understand).}

Sometimes, our connection with a topic is motivated more by ideas than by personal connections. We are naturally drawn to subjects that are either horrifying or uplifting. Indeed, many research projects begin with some notion – perhaps only dimly formulated – about what is wrong with the world. What real-life problem bothers you?\footnote{Gerring and Yenowine (2006), Shapiro (2005), Smith (2003).}

The desire to redress wrongs also helps to keep social science relevant to the concerns of lay citizens. We all begin, one might say, as citizens, with everyday concerns. Over time, we come to attain a degree of distance from our subject. Thus, our roles as citizens and scholars inform one another.

Of course, at the end of a project one must have something to say about a topic that goes beyond assertions of right and exonerations of wrong. The topic must be made tractable for scientific inquiry. If one feels that the topic is too close to the heart to reflect upon dispassionately upon it then it is probably not a good candidate for study. As a probe, ask yourself whether you would be prepared to publish the results of a study in which your main hypothesis is proven wrong. If you hesitate to answer this question because of normative pre-commitments you should probably choose another subject.

As a general rule, it is important to undertake questions that one feels are important, but not projects in which one has especially strong moral or psychological predilections for accepting or rejecting the null hypothesis.\footnote{Firebaugh (2008: ch 1).} For example, one might be motivated to study the role of worker-training programs because one is concerned about the problem of unemployment. But one probably should not undertake a study of worker-training programs to demonstrate that they are a good or bad thing. To do this would be to prejudge the answer to your research question.

**GET OFF YOUR HOME TURF**

While the previous section emphasized the importance of building upon one’s personal profile (skills, connections), it is also vital for scholars to stray from what is safe, comfortable, and familiar – their home turf.

Consider that academia is not now, and likely never will be, a representative cross-section of humankind. At present, the practitioners of social science are disproportionately white, Anglo-European, and male. They will probably always be disproportionately privileged in class background. Evidently, if members of these disciplines restrict themselves to topics drawn from their personal experience little attention will be paid to topics relevant to underprivileged groups.

\begin{footnotesize}
\begin{itemize}
\item Gerring and Yenowine (2006), Shapiro (2005), Smith (2003).
\item Firebaugh (2008: ch 1).
\end{itemize}
\end{footnotesize}
Equally important, advances in knowledge usually come from transgressing familiar contexts. After all, local knowledge is already familiar to those who live it. Whatever value might be added comes from moving beyond established categories, theories, and ways of thinking. A good ethnography, it is sometimes said, renders the exotic familiar or the familiar exotic. The same might be said of social science at large. Try to think like a stranger when approaching a topic that seems obvious (from your “home turf”). Likewise, do not be afraid to export categories from your home turf into foreign territory – not willfully, and disregarding all evidence to the contrary, but rather as an operating hypothesis. Sometimes, the imposition of a foreign category is illuminating.

Novel descriptive and causal inferences often arise when an extant concept or theory is transplanted from one area to another. For example, the concept of corporatism arose initially in the context of Catholic social theory as an alternative to state socialism. It was later adopted by fascist regimes as a way of legitimating their control over important economic and social actors. More recently, it has been seen as a key to explaining the persistence and resilience of authoritarian rule in the developing world. There are endless ways of adapting old theories to new contexts. Sometimes these transplantations are fruitful; other times, not.

Most important, try to maintain a conversation with different perspectives on your subject. What would so-and-so say about X? If this does not drive you mad, it may serve to triangulate your topic.

Another sort of boundary crossing is that which occurs across disciplines, theories, and methods. The trend of the contemporary era seems to be towards ever greater specialization, and to be sure, specialization has its uses. It is difficult to master more than one area of work, given the increasingly technical and specialized techniques and vocabulary developed over the past several decades.

Yet, it is worth reflecting upon the fact that many of the works that we regard today as pathbreaking have been the product of exotic encounters across fields and subfields. Indeed, all present-day fields and subfields are the product of long-ago transgressions. Someone moved outside their comfort zone, and others followed. Note also that the social sciences are not divided up into discrete and well-defined fields. So, try reading inside, and outside, your area of training.

All academic work is theft of one sort or another. So long as the sources are well-documented one need not fear the charge of plagiarism. And thefts sometimes produce novel insights. Another word for this sort of theft is creativity.

**PLAY WITH IDEAS**

The literature on invention and discovery – penned by science writers, philosophers of science, and by inventors themselves – is in consensus on one point. Original discoveries are usually not the product of superior brainpower (i.e., the ability to calculate or reason). Robert Root-Bernstein is emphatic.

---

Famous scientists aren’t any more intelligent than those who aren’t famous. [Moreover.] I’m convinced that successful ones aren’t right any more often than their colleagues, either. I believe that the architects of science are simply more curious, more iconoclastic, more persistent, reader to make detours, and more willing to tackle bigger and more fundamental problems. Most important, they possess intellectual courage, daring. They work at the edge of their competence; their reach exceeds their grasp… Thus, they not only succeed more often and out of all proportion; they also fail more often and on the same scale. Even their failures, however, better define the limits of science than the successes of more conventional and safe scientists, and thus the pioneers better serve science.134

The key question, as Root-Bernstein frames it, is “How can one best survive on the edge of ignorance?”135

One way of answering this question is suggested by Richard Hofstadter, who describes intellectual life as a counterpoint of piety and playfulness. The first refers to the sober and dogged search for truth. The second, which saves the enterprise from dogmatism and which may be less obvious, is the intellectual’s capacity to play.

Ideally, the pursuit of truth is said to be at the heart of the intellectual’s business, but this credits his business too much and not quite enough. As with the pursuit of happiness, the pursuit of truth is itself gratifying whereas the consummation often turns out to be elusive. Truth captured loses its glamor; truths long known and widely believed have a way of turning false with time; easy truths are a bore, and too many of them become half-truths. Whatever the intellectual is too certain of, if he is healthily playful, he begins to find unsatisfactory. The meaning of his intellectual life lies not in the possession of truth but in the quest for new uncertainties. Harold Rosenberg summed up this side of the life of the mind supremely well when he said that the intellectual is one who turns answers into questions.

Echoing Hofstadter’s description, one might say that there are two distinct moments in any research project. The first is open-ended, playful; here, a wide variety of different ideas are generated and given a trial run. The second is filled with zeal and piety; here, one grips tightly to a single idea in the quest to develop it into a fullblown theory and test it against some empirical reality. This conforms to a distinction between discovery and appraisal, which has a long history in the philosophy of science. Whatever the shortcomings of this dichotomy, there is no question that the academic endeavor requires a crucial shift of attitude at some point in the enterprise. Since we are concerned here with the initial phase, we shall dwell on techniques of playfulness.

Although the art of discovery cannot be taught (at least not in the way that the technique of multiple regression can be taught), it may be helpful to think for a moment about thinking. The act of creation is mysterious; yet, there seem to be a few persistent features. Arthur Koestler, synthesizing the work of many writers, emphasizes that discoveries are usually “already there,” in the sense of being present in some body of work – though perhaps not the body of work that it had heretofore been associated with. To discover is, therefore, to connect things that had previously been considered separate. To discover is to think analogically.

This leads to the paradox that the more original a discovery the more obvious it seems afterwards. The creative act is not an act of creation in the sense of the Old Testament. It does not create something out of nothing; it uncovers, selects, re-shuffles, combines, synthesizes already existing facts, ideas, faculties, skills. The more familiar the parts, the more striking the new whole. Man's knowledge of the changes of the tides and the phases of the moon is as old as his observation that apples fall to earth in the ripeness of time. Yet the combination of these and other equally familiar data in Newton's theory of gravity changed mankind's outlook on the world.136

What frame of mind does this require? How does one think analogically? This trick seems to have something to do with the capacity to "relinquish conscious controls," to block out the academic superego that inhibits new thoughts by punishing transgressions against the tradition.137 Above all, one must feel free to make mistakes.

Just as in the dream the codes of logical reasoning are suspended, so 'thinking aside' is a temporary liberation from the tyranny of over-precise verbal concepts, of the axioms and prejudices engrained in the very texture of specialized ways of thought. It allows the mind to discard the strait-jacket of habit, to shrug off apparent contradictions, to un-learn and forget – and to acquire, in exchange, a greater fluidity, versatility, and gullibility. This rebellion against constraints which are necessary to maintain the order and discipline of conventional thought, but an impediment to the creative leap, is symptomatic both of the genius and the crank; what distinguishes them is the intuitive guidance which only the former enjoys.138

It might be added that what also distinguishes the genius and the crank is that the former has mastered the tradition of work on a subject. Her liminal moments are creative because they take place on a foundation of knowledge. In order to forget, and thence recombine features of a problem, one must first have a first-hand familiarity with the facts.

**PRACTICE DIS-BELIEF**

One can't think without words, but sometimes one can't think well with them either. Sometimes, ordinary language serves to constrain thought-patterns, reifying phenomena that are scarcely there. When we define, Edmund Burke commented, "we seem in danger of circumscribing nature within the bounds of our own notions."139 Language suggests, for example, that where a referential term exists a coherent class of entities also exists, and where two referential terms exist there are two empirically differentiable classes of entities. Sometimes this is true, and sometimes it is not. Just because we have a word for "social movement" does not mean that there are actually phenomena out there that are similar to each other and easily differentiated from other phenomena. Ditto for "social capital," "interest group," and virtually every other key concept in the social science lexicon. Words do not always carve nature at its joints. Sometimes, they are highly arbitrary ("constructed"). Likewise, just because we have a word for some phenomenon does not mean that cases of this phenomenon all stem from the same cause. It is not even clear that the same causal factors will be relevant for all members of the so-named set of phenomena.

---

137 Koestler (1964: 169).
139 Quoted in Robinson (1954: 6).
The reader might respond that, surely, concepts are defined the way they are because they are useful for some purposes. Precisely. But it follows that these same concepts may not be useful for other purposes. And since one’s objective at early stages of the research game is to think unconventionally, it is important to call into question conventional language. For heuristic purposes, try assuming a nominalist perspective: words are merely arbitrary lexical containers. As an exercise, put brackets around all your key terms (“social movement”). Now try to re-describe the phenomenon of interest using different language.

A parallel skepticism must be extended to numbers, which also naturalize phenomena that may, or may not, go together in the expected fashion. Here, the claim is more complicated. First, the use of a number is explicitly linked to a dimension – e.g., temperature, GDP, number of auto accidents – that is thought to be relevant in some way. Moreover, the imposition of a numerical scale presupposes a particular type of relationship between phenomena with different scores on that variable – nominal, ordinal, interval, or ratio (Chapter 3). But is it really? More broadly, is this the dimension that matters for understanding the topic in question? Are there other dimensions, perhaps less readily quantified, that provide more accurate or insightful information? While GDP is the conventional measure of economic development it could be that other aspects of modernization are more relevant for explaining an outcome of interest. An obsessive focus on GDP may serve to obscure that relationship.

Another sort of conventional wisdom is contained in paradigm-cases. These are cases that, by virtue of their theoretical or everyday prominence, help to define a phenomenon – the way Italy defines fascism, the Holocaust defines genocide, the United States defines individualism, Sweden defines the welfare state, and the Soviet Union (for many years) defined socialism. Paradigm-cases exist in virtually every realm of social-science inquiry. They often provide good points of entry into a topic because are overloaded with attributes; they operate in this respect like ideal-types (Chapter 3). Yet, because they anchor thinking on these topics, they are also thought-constraining. And because they are also apt to be somewhat unusual – even extreme – examples of the phenomenon in question, they may present misleading depictions of that phenomenon.

With respect to words, numbers, and paradigm-cases – not to mention full-blown theories – it is important to maintain a skeptical attitude. Perhaps they are true and useful, perhaps only partially so, or only for certain purposes. To test their utility, try adopting the Socratic guise of complete ignorance (better labeled as thoroughgoing skepticism). Once having assumed this pose, you are then free to pose naïve questions of sources, of experts, and of informants. It is a canny strategy and can be extraordinarily revealing – particularly when “obvious” questions cannot be readily answered, or are answered in unexpected ways.
OBSERVE EMPATHICALLY

One technique of discovery is empathic, or (to invoke the philosophical jargon) hermeneutic. Here, one employs observational techniques to enter into the world of the actors who are engaged in some activity of interest – playing ball, drafting a bill, murdering opponents, casting a vote, and so forth – in order to understand their perspective on the phenomenon. Of course, this is easier when the actors are our contemporaries and can be studied directly (i.e., ethnographically). It is harder, and yet sometimes more revealing, if the actions took place long ago or are removed from direct observation, and must be reconstructed. In any case, non-obvious perceptions require interpretation, and this interpretation should be grounded in an assessment of how actors may have viewed their own actions.

Consider that the process of understanding begins with an ability to re-create or re-imagine the experiences of those actors whose ideas and behavior we wish to make sense of. Somehow, a link must be formed between our experiential horizons and the horizons of the group we wish to study. This may involve a form of role-playing (what would we do in Situation X if we were Person Y?). Some level of sympathy with one’s subjects is probably essential for gaining insight into a phenomenon. This may be difficult to muster if the subject is grotesque. No one wants to empathize with Nazis. But the hermeneutic challenge remains; some way must be found to enter into the lives and perceptions of these important historical actors in order to explain their actions, however repellant they might be.

THEORIZE WILDLY

Rather than working single-mindedly towards One Big Idea, you might consider the benefits of working simultaneously along several tracks. This way, you avoid becoming overly committed to a single topic, which may turn out to be unavailing. You can also compare different topics against one another, evaluating their strengths and weaknesses. “Have lots of ideas and throw away the bad ones,” advises Linus Pauling. At the same time, you should do your best to maintain a record of your ideas as you go along. Take a look at this idea-diary every so often and see which projects you find yourself coming back to, obsessing about, inquiring about. The objective should be to keep your mind as open as possible for as long as possible (given the practicalities of life and scholastic deadlines). “Let your mind become a moving prism catching light from as many angles as possible.”

Historians of natural science identify productive moments of science with the solving of anomalies – features of the world that don’t comport comfortably with existing theories. If these anomalies can be solved in a more than ad hoc manner, the frontiers of knowledge are pushed forward. Perhaps a new paradigm of knowledge will be created.

140 Gadamer (1975).
141 Quoted in Root-Bernstein (1989: 409).
142 Mills (1959: 196).
143 Mills (1959: 214).
Another technique for theorizing wildly is to juxtapose things that don’t seem to fit naturally together. Theorizing often consists of dis-associating and re-associating. One version of this is to examine a familiar terrain and think about what it resembles. What is “X” an example of? Charles Ragin calls it “casing” a subject.145

Another tactic is to examine several diverse terrains in order to perceive similarities. Can colonialism, federalism, and corporatism all be conceptualized as systems of “indirect rule”?146

A fourth tactic is to examine a familiar terrain with the aim of recognizing a new principle of organization. Linnaeus famously suggested that animals should be classified on the basis of their bone structures, a new principle of classification that turned out to be extraordinarily illuminating.147 In the realm of social science, scholars have provided organizational schemes for political parties, bureaucracies, welfare states, and other social phenomena.

A fifth technique for loosening the theoretical wheels is to push a conventional idea to its logical extreme. That is, consider an explanation that seems to work for a particular event or in a particular context. (It may be your idea, or someone else’s.) Now push that idea outward to other settings. Does it still work? What sort of adjustments are necessary to make it work? Or consider the logical ramifications of a theory – if it were fully implemented. What would the theory seem to require?

Theories are tested when they are pushed to their limits, when they are tried out in very different contexts. Root-Bernstein observes that this strategy leads, at the very least, to an investigation of the boundaries of an idea, a useful thing to know. Alternatively, it may help us to reformulate a theory in ways that allow it to travel more successful, i.e., to increase its breadth. It may even lead to a new theory that explains the whole empirical realm.148

In theorizing wildly, it is important to keep a list of all possible explanations that one has run across in the literature, or intuited. As part of this canvas, one might consider some of the more general models of human behavior as discussed in Chapter 5. Sometimes, these abstract models have applications to very specific problems that might not be immediately apparent. (How might the topic of romance be understood as an exchange? As an adaptation? As a product of diffusion?)

Once achieved, this list of possible explanations for phenomenon Y can then be rearranged and decomposed (perhaps some propositions are subsets of others). Recall that theoretical work often involves recombining extant explanations in new ways. Your list of potential explanations also comprises the set of rival hypotheses that you will be obliged to refute, mitigate, and/or control for (empirically) in your work. So it is important that it be as comprehensive as possible.

In order to figure out how to correctly model complex inter-relationships it is often helpful to draw pictures. If one is sufficiently fluent in graphic design, this may be handled on a computer screen. For the rest of us, pencil and paper are probably the best expedients. Laying out ideas with boxes and arrows, or perhaps with venn diagrams or decision trees, allows one to illustrate potential relationships in a more free-flowing way than is possible with prose or math. One can “think” abstractly on paper without falling prey to the constraints of words and numbers. It is also a highly synoptic format, allowing one to fit an entire argument onto a single sheet or white board.

146 Gerring et al. (2011).
147 Linsley and Usinger (1959).
THINK AHEAD

All elements of the research process are intimately connected. This means that there is no such thing as a good topic if that topic is not joined to a good theory and a workable research design. So, the choice of a topic turns out to be more involved than it first appears. Of course, all of the elements that make for a successful piece of research are unlikely to fall into place at once. And yet, one is obliged to wrestle with them at the very outset.

Recalling the elements of your topic – containing, let us say, a theory, a set of phenomena, and a possible research design – it is vital to maintain a degree of fluidity among all these parts until such time as you can convince yourself that you have achieved the best possible fit. Beware of premature closure. At the same time, to avoid endless cycling it may be helpful to identify that element of your topic that you feel most committed to, i.e., that which is likely to make the greatest contribution to scholarship. If this can be identified, it will provide an anchor in this process of continual re-adjustment.

Consider the initial decision of a topic as an investment in the future. As with any investment, the pay-off depends upon lots of things falling into place. One can never anticipate all of the potential difficulties. But the more one can “game” this process, the better the chance of a pay-off when the research is completed. And the better the chance that the research will be completed at all. (Really bad ideas are often difficult to bring to fruition; the more they advance, the more obstacles they encounter.)

Although the prospect may seem daunting, one is obliged to think forward even at the “just getting started” stage of research. Try to map out how your idea might work – what sort of theory will eventuate, what sort of research design, and so forth. If everything works out as anticipated, what will the completed project like?

An obvious question to consider is what results a study is likely to generate. Regardless of the type of study undertaken there will presumably be some encounter with the empirical world, and hence some set of findings. Will the evidence necessary to test a theory, or generate a theory, be available? Will the main hypothesis be borne out?

Sometimes, a failure to reject the null hypothesis means that the researcher has very little to show for her research. Conventional wisdom has prevailed. Other times, the failure to prove a hypothesis can be enlightening. Sometimes, a topic is so new, or a research design so much more compelling than others that came before, that any finding is informative. This is ideal from the perspective of the scholar’s investment of time and energy, as it cannot fail to pay off.

In any case, you should consider how your anticipated findings might be situated within the literature on a topic. How will it be perceived? What will be its value added?

In test-driving your idea you should also keep a close eye on yourself. See if your oral presentation of the project changes as you explain it to others (e.g., your friends and classmates). At what point do you feel most confident, or most uncertain? When do you feel as if you are bullshitting? These are important signals with respect to the strengths and weaknesses of your project. Indeed, the process of discussing your ideas – aside from any concrete feedback you receive – may force you to reconsider issues that were not initially apparent.

Most important, try to evaluate the feasibility of your project. Your work is probably subject to a hard deadline, so construct a realistic time-schedule with that deadline in mind. As you construct that schedule, be aware that research and writing generally demand more time than one imagines they will. Alternatively, you may identify a back-up plan – a shorter version of the project that can be implemented if the full-scale version turns out to be unrealistic.
CONCLUSIONS

Published work in the social sciences presents a misleading appearance of order and predictability. The author begins by outlining a general topic or research question, then states a general theory, and from then proceeds to the specific hypothesis that will be tested and her chosen research design. Finally, the evidence is presented and discussed, and concluding thoughts are offered.

This is nothing at all like the progress of most research, which is circuitous and unpredictable – hardly ever following a step-by-step walk down the ladder of abstraction. One reason for this is that knowledge in the social sciences is not neatly parcelled into distinct research areas, each with specific and stable questions, theories, and methods. Instead, it is characterized by a high degree of open-endedness – in questions, theories, and methods.

There is no Archimedean point of entry to this maze. One might begin with a general topic, a research question, a key concept, a general theory, a specific hypothesis, a compelling anomaly, an event, a research venue (e.g., a site, archive, or dataset), a method of analysis, and so forth. Accordingly, some research is problem- or question-driven, some research is theory-driven, some research is method-driven, and other research is phenomenon-driven (motivated by the desire to understand a particular event or set of events). These are obviously quite different styles of research – even though, at the end of the day, each study must be held accountable to the same methodological criteria, as laid out in Parts I and II of the book.

There is no right or wrong place to start; all that matters is where you end up. And yet, where one ends up has a lot to do with where one starts out. Students are rightly wary of the consequences of choosing a bad topic – one that, let us say, promises few interesting surprises, has little theoretical or practical significance, or offers insufficient evidence to demonstrate a proposition about the world. No matter how well-executed that research might be, little can be expected from it.

Because the selection of a good topic is difficult, careful deliberation is in order. In this arduous process, advice is welcome – from friends, family, teachers, advisors, experts in the field. Solicit all the feedback you can. But make sure that, at the end of the day, you are comfortable with the choice(s) you make.
**KEY TERMS**

- Paradigm-Cases

**INQUIRIES**

1. Within the scope of a general subject area (e.g., as defined by a course), identify ten possible research topics. In identifying a topic, you should state a general research question, e.g., Are countries with a heritage of English colonial rule more likely to democratize than other countries in the developing world? Ten topics may seem like quite a lot. However, the purpose of this exercise is to brainstorm, i.e., to get into the mode of coming up with ideas.

2. Next, explore these topics in a preliminary fashion, perhaps by getting feedback from a friend (who knows something about the subject area). In a classroom setting, each idea can be vetted by someone else in the class. The idea is to get an initial sense – without doing much actual research – of how promising each possible topic might be. Or, at least, what issues would need to be resolved in order for each topic to be worthwhile. Is the research idea new, or is there some new theoretical or empirical angle worth exploring? Is it feasible within your time and budget constraints? Is it too big, or perhaps too small? Is the topic properly bounded? What areas of theory and research are relevant to the topic? If the topic yields the result that you anticipate, will this be a contribution to the literature on this subject? And so forth.

3. Finally, which topic would you choose to pursue for a research paper? Why this one (and not the others)? What are the biggest challenges to this topic.

**RESOURCES**


**Examples:**
13. **Data Gathering**

Insofar as social science research rests on an empirical base it involves an encounter with data. Data may be qualitative or quantitative. It may be numerical, textual, visual, olfactory, auditory, or artifactual. It may be drawn from primary or secondary sources. It may be collected by the researcher or by someone else. “Data” is understood here in the broadest sense, including any sort of evidence that serves as the empirical basis for understanding and explanation.

In this chapter we lay out various methods of gathering data, a topic that is important for both producers and consumers of social science. Note that in order to judge the quality of a study you need to be familiar with the data-collection method that underlies the analysis.

Data-gathering methods may be classified broadly as obtrusive or unobtrusive. **Obtrusive methods** comprise standardized surveys as well as less structured interpersonal settings such as interviews, focus groups, and ethnography. **Unobtrusive measures** include those that are surreptitious and ex post. After outlining these techniques of data collection we discuss the problem of data assessment: how do we know that our data (or someone else’s) is true?

Before launching into the chapter, a few caveats and clarifications are in order.

First, each section of the chapter is cumulative. For example, many of the ingredients of a good survey are also ingredients of a good interview, focus group, or ethnography. This explains why the section on surveys is somewhat longer than these later sections of the chapter. It also means that this chapter should probably be read from front to back rather than browsed in a selective fashion.

Second, each method will be treated as a stand-alone technique of data collection. However, readers should be aware that in the course of conducting a study it is common to combine several data-collection strategies. Where a multimethod approach is adopted, the strengths and weaknesses of each strategy should be understood within that context.

Third, experiments and case studies are regarded in this book as research designs rather than data-gathering techniques and are therefore discussed in Part II rather than in this chapter. The distinction becomes clear when one considers that experiments and case studies may be implemented with a variety of data-gathering tools, as outlined here. Note also that strategies of gathering data might be understood as strategies of measurement, so the topics discussed here overlap those introduced in Chapter 3.

Fourth, any data-gathering effort must wrestle with fundamental tasks of all empirical analysis. Researchers must reflect upon (a) the basic unit of analysis (e.g., persons or countries), (b) the units that will be included in the sample and the size of the sample, (c) the population that the sample is drawn from, and (d) the extent to which the sample is likely to be representative of that population. These issues were presented in Chapter 4.

Finally, a note on terminology. People who gather data may be referred to as ethnographers, facilitators, interviewers, investigators, researchers, or surveyors – depending upon the context. People who provide data may be referred to as coders, informants, interviewees, participants, raters, respondents, or subjects.

---

149 Typically, either surveys or unobtrusive measures are used to measure pre-treatment covariates and outcomes of interest in an experiment. Surveys may even be used as the experimental intervention (the split-sample survey or survey experiment). There is usually a post-experiment interview in which the researcher debriefs the subject and attempts to understand his/her subjective experience. In field experiments it is common to apply ethnographic methods to the community under study.
– again, depending upon the context. Each of these terms has a somewhat different connotation. However, the distinction of importance for present purposes is between data **gatherers** and data **providers**. Terms that fall within each of these categories will be used interchangeably.

**STANDARDIZED SURVEYS**

A great number of methodological issues are sometimes grouped together under the rubric of **survey research**. However, many of these issues are not limited to standardized surveys, as they also apply to other styles of data-gathering. Thus we deal with issues of sampling in Chapter 4, issues pertaining to scaling and indexing in Chapter 2, and issues pertaining to causal analysis in Part II. The following section is limited to the construction of a survey instrument – understood as a standardized questionnaire with limited response-options – and its implementation. (Open-ended questions are discussed below under the rubric of **interviews**.)

There are many variations on the survey instrument. *Surveys...*

- May be conducted in person, by phone, by mail, by email, or by web site
- May be conducted by an interviewer, an automated voice, or self-administered (via hard copy or computer)
- May be administered in a public space (e.g., laboratory, street) or in a private space (e.g., the respondent’s home)
- May involve members of the general public, a subset of the public, elites, or experts
- May focus on the attitudes and behavior of the respondent or on some objective feature of the world (coding)
- May or may not offer compensation
- May incorporate large or small samples
- May be of any length and duration
- May be implemented once (producing a *cross-section*), multiple times with the same respondents (a *panel*), or multiple times with different respondents drawn from the same population (a *pooled time-series*).

Each variation in survey design involves distinct challenges and opportunities, and may require special strategies of implementation or analysis. You should consider carefully the impact of different choices on (a) response rates (the likelihood of those contacted agreeing to participate in the survey), (b) completion rates (the percent of respondents who successfully complete the survey), (c) sample representativeness (vis-à-vis the population of interest), (d) possible biases in responses, and (e) the time and cost involved in the implementation of a survey.

**MEDIUM, LOCATION**

A survey may be administered in person, by phone, by mail or email, or by web site. The administrator may be a “live” surveyor, an automated voice, a computer, or may be entirely absent (as in a pencil and paper survey). The location may be a public place (e.g., a laboratory or street) or a private place (e.g., a respondent’s home).

With the spread of computers and access to the web it is now possible to administer many surveys on-line. Here, the surveyor may take advantage of on-line sites like Survey Monkey that allow the researcher to set up a survey instrument using existing templates and to collect responses in a convenient data format.
However, choices among these various options have important consequences. For example, implementing a survey by email or web site effectively excludes those who don’t have access to the internet or who feel uncomfortable with the medium. It may, however, give the respondent a greater sense of anonymity and thus may prompt more honest responses than one would get in a person-to-person exchange (where an interviewer is administering the survey by phone or in person). Of course, the situation is probably reversed in societies where the internet is routinely monitored; there, respondents may feel more comfortable talking directly to a researcher (whose bona fides they can judge) rather than to a machine.

Wherever a survey is administered “live” by an interviewer there is a possibility of interviewer bias. That is, responses may be influenced by the interviewer. A good deal of research has been conducted on this point. It has been shown, for example, that responses are often affected by the sex and race of the interviewer, or by the power relationship between interviewer and respondent. Even subtle inflections of an interviewer’s voice may serve as a cue to the respondent. Issues such as these are often difficult to overcome, even if interviewers are well-trained and attentive. (Neither the interviewer nor the respondent may be entirely aware of such biases.)

Another problem is social desirability bias. The presence of an interviewer may prompt the respondent to answer a question in a way that s/he feels is most appropriate and acceptable, even when it does not accord with his/her true preferences. This is a common problem with surveys that touch on issues of race, sexual behavior, corruption, and other sensitive topics. Of course, this sort of bias may also be present in a self-administered survey. However, it is apt to be more problematic when an interviewer is present – in person or at the other end of a phone line.

A self-administered survey (via hard copy or computer) eliminates interviewer bias and may reduce social desirability bias. However, this format requires more of the respondent. If respondents are illiterate they will be unable to participate. And if they are bored or annoyed by the questions they are likely to quit before completing, or pay little attention to their answers. Accordingly, response-rates and completion-rates tend to be lower for self-administered surveys than for comparable surveys administered person-to-person.

Respondents

Respondents for a survey may be drawn from the general public or from some subset of the public, e.g., a minority group, teenagers, or those who have tested HIV-positive. These sorts of respondents may be regarded as members of the mass public, i.e., lay respondents.

A rather different sort of respondent is a member of an elite group, e.g., a business leader, political leader, or leader in a profession. Elites tend to be much less accessible, which is to say that response rates and completion rates will be lower. Moreover, elites are unlikely to respond to monetary compensation (or will require a great deal of compensation). Obtaining an elite interview may therefore require a lot of work, and perhaps some special connections or some quid pro quo arrangement (something you can do for them). Typically, elites are more willing to engage in an open-ended conversation than in a standardized survey, as they may not consider the latter a good use of their scarce time. If they are willing to submit to a survey it may be necessary for the researcher to administer it in person or over the phone.

A third sort of respondent is an expert in some area. This sort of respondent qualifies for inclusion in a survey because of his/her knowledge or experience, not her position or demographic characteristics. A wide range of surveys are focused on experts. Here, the goal is usually to obtain

151 Barath and Cannell (1976).
specialized knowledge of some topic rather than the personal predilections of the respondent. For example, researchers wishing to judge the ideology of a political party or the quality of democracy in a country often call upon experts to code these features of a polity. Typically, these experts are trained in social science or have some professional position that qualifies them as experts in this area. This sort of coding is useful whenever (a) it is difficult to reduce a subject to objective indicators and (b) when members of the general public are not very knowledgeable about that subject. While surveys geared for experts can assume a higher level of knowledge about the subject, one should not lose sight of the fact that experts are people too. That is, they may be subject to the same problems of non-response, non-completion, and bias as members of the general public.

** Compensation**

In order to recruit respondents, and to minimize non-response and non-completion, it may be necessary to offer some compensation to those who agree to undertake a survey. Respondents may be paid directly for their work or may be offered the opportunity to win a prize. One option that is increasingly popular among survey researchers is to recruit respondents through *Mechanical Turk*, a web site owned and maintained by Amazon. Here, one may find respondents willing to fill out surveys on a wide range of topics for a nominal sum.

Arguably, those who are paid for their work are more likely to take it seriously. However, one must also consider how it might affect the sort of people who agree to take the survey and the sort of answers they provide. Consider a survey that focuses on the respondent’s willingness to donate blood. Suppose that one of the questions is about monetary compensation for blood donations. The answer to this question may be affected by the structure of the survey, in which the respondent has already agreed to sell his/her services for a price.

** Introductory statement**

When contacting potential respondents, and/or at the start of a survey, several important issues must be covered. The potential respondent should be informed of…

- The topic of the survey. (This is a complicated issue, as you are under an obligation to inform the participant; at the same time, you don’t want to reveal anything about the survey that might compromise the possibility of obtaining unbiased responses.)
- The method by which the respondent was selected.
- Requirements for participation, e.g., the time required, the format, and any special instructions.
- Compensation (if any).
- Sponsorship of the survey, i.e., the identity of the principal investigator (PI), the identity of the interviewer (if different from the PI), and the funding agency (if any).
- Use to which the data will be put.
- Whether survey results will be made public or otherwise made available to the respondent.
- Whether the respondent will be granted anonymity (no one associated with the survey knows his/her identity) or confidentiality (no one except members of the project team know his/her identity), and what measures will be taken to assure this.
- Participation in the survey is voluntary. The respondent may refuse to answer any question and may exit the survey at any time.
- Specific Institutional Review Board (IRB) procedures that may apply, e.g., an explicit statement by the respondent that s/he is willing to undertake the survey.
- Researcher contact information (email, telephone, address), in case s/he has further questions. The potential respondent should also be encouraged to ask questions about any of the foregoing, if something is unclear.
Some of these points are commensensical. Others are informed by ethical imperatives, as discussed in Chapter 16.

**QUESTIONNAIRE**

The content of a survey – including all text, pictures, or other media as presented to the respondent – will be referred to collectively as the *questionnaire*. The questionnaire may consist of several questions or several hundred questions. Accordingly, it may require just a few minutes or several days to complete. A longer survey will probably result in lower response- and completion-rates, and may mean that respondents pay less attention to the questions and their responses. However, some respondents are willing to bear with a lengthy survey, especially if administered in-person and/or accompanied by some sort of compensation. Surveys conducted in the developing world sometimes last several hours. Members of the mass public in rich countries, as well as elite respondents (everywhere), are often less generous with their time.

The first few questions on a survey are usually designed to ease the respondent into the situation, to make them feel comfortable with the format, and to establish a rapport with the interviewer (if any). Sometimes, it is important to collect personal information about the respondent, e.g., age, sex, residence, occupation, and so forth. These are usually easy questions for the respondent to answer. However, these apparently bland factual questions are sometimes sensitive and may also serve to “prime” the respondent for what is to follow.

**Priming** (or framing) refers to a statement or question that affects later responses. Suppose, for example, that a respondent is asked about his/her age in an early question and then, later on, is asked about his/her feelings about the government's social security program. It is quite possible that the first question affects responses to the second question. This may be a desirable; sometimes, a researcher wishes to prime certain topics. Or it may be undesirable. Either way, it is important to be aware of the way in which question-ordering affects the responses one receives. This may be tested systematically by reversing the ordering of questions (or removing questions) and administering each version of the survey to randomized treatment and control groups – a *survey experiment* (as discussed in Chapter 7).

If one wishes to neutralize the priming effects of question-ordering one can also randomize this feature of a survey, assuring that equal numbers of respondents receive a survey in which Question \(A\) precedes Question \(B\) and Question \(B\) precedes Question \(A\). This solution, however, may interfere with a natural ordering of topics. If questions on a survey skip from topic to topic without any apparent logic, respondents may become confused.

Equally important is the way in which a question is posed. Consider the following statement intended to elicit a respondent’s position on abortion: Abortion is wrong. It seems obvious that one will obtain different sorts of answers to this question depending on how individual respondents interpret it. One might ask: Does abortion mean in all cases or just for particular situations? Some might respond to “wrong” in terms of it being a moral question while others may see it as a legal question. We have given a stark example of this, but even slight differences in question wording (e.g., substituting ‘reproductive choice’ for ‘abortion’) may be sufficient to alter the number of people who support a women’s right to abortion and thus may be claimed as political fodder for those who are fighting political battles on this issue. The answer may then be construed to indicate differing levels of support for abortion in the population.

Are some questions more correct than others? It is usually wise to steer clear of a question that seems to pre-judge the correct response, i.e., a question that is framed in a “biased” manner. One probably would not pose a question about abortion by asking whether the respondent is for or against “killing babies.” On the other hand, it may be revelatory to ask the question in this
inflammatory way – so long as the results are not systematically misinterpreted. Consider that someone who is willing to defend abortion when the question is posed in such an extreme manner is probably a very strong supporter. If the goal of the survey is to gauge varying levels of support, a “biased” question such as this may be quite fruitful.

Note also that some topics like abortion are difficult to address in an unbiased fashion. Supporters and opponents of abortion are unlikely to agree on the proper wording of a survey question. When polls on a subject appear to conflict it is often because they are asking a question in different ways or have not provided enough clarity in the question or response options to attain at a valid response from all respondents. A useful approach in this situation is to pose a series of statements or questions that offer a variety of opinions on a subject or a variety of ways of framing a subject. One may also preface these statements with “Some people say…” or “There is debate over the question of…” to indicate that the surveyor is not taking sides in the dispute. On the other hand, asking a range of questions on an issue will require a lot of space on the survey. Regardless, to avoid interpretation problems, it is important that the question is phrased so as not to lead respondents and that the question provide the full range of possible outcomes. The American National Elections Study wording of the abortion question attempts to do exactly that. Here they ask: “Which one of the opinions on this page best agrees with your view?” And they provide the following response choices:

1. By law, abortion should never be permitted
2. The law should permit abortion only in case of rape, incest, or when the woman’s life is in danger.
3. The law should permit abortion for reasons other than rape, incest, or danger to the woman’s life, but only after the need for the abortion has been clearly established.
4. By law, a woman should always be able to obtain an abortion as a matter of personal choice.

If the goal of the researcher is to measure variation across a population – rather than a point estimate for the population at-large – it may not be necessary to introduce multiple questions about the same topic. Consider the four questions on abortion listed above. Now imagine that there is a latent concept – degree of support for the right to abortion – that one is attempting to measure. If this latent concept underlies responses to all the questions then it ought to be manifested in answers to all of these questions. That is, those with strong support for the right to abortion will give the “pro-abortion” response regardless of how the question is posed; those strongly opposed to the right to abortion will give the “anti-abortion” response regardless of how the question is posed; and those who are ambivalent will give responses that lie in between the “pro” and “anti” positions. In this scenario, altering the question affects the level of support but not the distribution of support. (In regression analysis, the intercept changes but not the slope of the regression line. See Chapter 19.) In this manner, one may be able to ascertain variation in opinion across a population with a single question – so long as there is some variation in responses. One may also combine questions about the same general topic into an index, as discussed in Chapter 3.

Constructing a good question requires keen intuition about how respondents are likely to interpret and respond to that question. It is vital that respondents interpret the meaning of the question in the same way. So the survey researcher must consider how a question might be interpreted by respondents, bearing in mind possible differences in education, age, culture, or other background factors. If respondents understand a key term (e.g., “abortion”) differently, then one has a problem of non-equivalence. The same responses actually mean different things. Suppose some people understand abortion as any interruption of pregnancy prior to birth, while others understand it as an interruption prior to the first trimester. Here, people who say they oppose abortion may be opposing different sets of practices. Some may support right of a woman to interrupt a pregnancy prior to the first trimester; others may not.
Insofar as survey questions attempt to delve into difficult and contested issues they are inevitably subject to problems of interpretation. It is the responsibility of the surveyor to minimize such ambiguities, without sacrificing the topic of interest. To this end, here are some helpful tips.

- State questions and answers in the simplest and most natural way possible.
- Questions (and accompanying explanations) should be no longer than they need to be.
- Avoid technical jargon (unless your respondents are experts). If you must include a technical term, provide a definition.
- Be wary of terms that can mean different things (e.g., “inflammable”).
- Avoid double-negatives (e.g., “Do you oppose the abolition of taxes on cigarettes?”).
- Do not combine several disparate elements in a single question (e.g., “Do you support democracy and the rule of law?”).
- If you wish to employ an abstract concept (e.g., “liberalism”), also include some questions that probe into that concept in a specific context (e.g., “Do you favor increasing government support for education?”).
- Use vignettes (short stories that illustrate a concept) whenever a concept may be unclear. However, be aware that vignettes necessarily introduce a lot of background noise, which may affect responses in ways that diverge from the concept of interest.
- Adopt questions from other (reputable) surveys wherever possible – assuming they serve your goals and satisfy other requirements, as above. These questions have been vetted. Also, results obtained from your sample can be directly compared with those obtained from other surveys if the questions (and accompanying responses) are identical. (Make sure to acknowledge this borrowing, lest you be accused of plagiarism.)

**Response-Categories**

A standardized survey specifies a set of responses that are available to the respondent. These may reproduce any type of scale (see Chapter 3). The scale may be nominal – Yes/No. It may be ordinal (Strongly agree, Agree, Neither agree nor disagree, Disagree, Strongly disagree), a particular kind of ordinal scale known as a Likert scale. It may be interval, as in the “feeling thermometer” response which asks a respondent to register their agreement/disagreement on a scale of 0 to 100.

The interval scale is the most sensitive, and thus is generally most helpful in discerning small differences among respondents. However, an interval response is not always appropriate. One cannot ask someone to indicate whether they are male or female on a 100-point scale. Likewise, there are many circumstances in which ordinal-level categorical responses are more natural than interval-level responses. Arguably, it is easier for a respondent to understand a Likert scale than a feeling thermometer.\(^\text{152}\) If the respondent does not understand what a “10” on a 100-point scale means, or if different respondents interpret it differently, a good deal of noise will be introduced into the survey instrument. So, sensitivity to fine differences is not always the most important consideration. Consider also that a set of questions may later be combined into a single index, using techniques discussed in Chapter 3.

**Knowledge, Certainty, and Salience**

A survey is free to inquire about anything the surveyor wishes (so long as it doesn’t violate ethical norms). And many respondents will be happy to oblige the surveyor with an answer (perhaps because they believe that is what is required of them). But this does not mean that the respondent

\(^{152}\) A particular problem with the agree/disagree Likert scale is that some respondents may be more inclined to agree than to disagree. A study conducted some years ago found this bias to be strongest among less educated respondents (Schuman, Presser 1981: 233; cited in Converse, Presser 1986: 39).
knows or cares much about that issue, or is very certain about his/her response. Thus, when we read that X% of Americans support NAFTA (North American Free Trade Agreement) we must process this fact with a grain of sodium chloride (salt).

There are ways of ascertaining the knowledge, certainty, and salience of an issue to the respondent. One can add questions of a factual nature intended to gauge a respondent’s knowledge of an issue. (What does NAFTA stand for?) One can ask the respondent directly about his/her knowledge of a subject or its salience to him/her. One can offer a “Don’t know” (DK) or “No opinion” option. One can ask a follow-up question about the respondent’s confidence in their response.

Prior questions may also serve a screening function, i.e., only those who demonstrate knowledge of a subject are asked the next question. Alternatively, one may be interested in how knowledge or salience affects responses to a question, in which case the respondent would be required to answer both questions.

Note that including a DK option may dramatically change the results if the distribution of uncertain respondents across the response-categories is not uniform. Thus, if those who oppose NAFTA are strong in their opposition while those who claim to support NAFTA are not sure, many of the latter may opt for the DK option, thus altering the distribution of positive/negative responses.

**SENSITIVE SUBJECTS**

In the context of survey research – where there is ordinarily no opportunity to gain the trust of respondents or to judge their responses in a contextual fashion – there are nonetheless ways of accessing sensitive subjects.

Perhaps the most important step is to guarantee the anonymity of the respondent by omitting their name from the survey. Unfortunately, in an electronic age it is hard to convince a skeptical respondent that his/her anonymity will be respected. Consider that whatever mode of contacting is employed – in-person, mail, email, or web – involves a potential sacrifice of anonymity. In some countries, spies may pose as surveyors, and respondents are not generally aware that academic “IRB” protocol (discussed in Chapter 16) imposes strict requirements on any research on human subjects.

One can also adopt an anonymous setting for the survey, which may be administered by mail or on-line. This may do more to reassure the respondent of his/her anonymity, even though in reality the respondent probably has greater assurance of anonymity in a public setting like a mall or park (so long as she/he does not reveal his/her identity).

Another approach involves the construction of the questionnaire. One may frame sensitive subjects as questions about other people, e.g., “Do you think that other employers use race as a criterion for making hiring decisions?” The assumption here is that those engaged in activities that are denigrated by society (e.g., discrimination or corruption) will be inclined to see these activities as widespread, as this may assuage feelings of guilt or shame.

One may also enlist an experimental survey design in order to mask individual identities. The list experiment begins by sorting respondents randomly into two groups, each of which is given a small set of questions to ponder. The questionnaires are identical except that the treatment group is given one additional question of a sensitive nature (e.g., pertaining to racism or corruption). Respondents are then asked to report the total number of questions that they agree (or disagree) with, but not their answers to any specific question. Since the treatment and control groups are

---

153 Peeters et al. (2010), Warner (1965). Experiments employed to measure the concept of trust are reviewed in Nannestad (2008).
assumed to be comparable in all respects except the one additional question asked of the treatment group, any differences in responses across the two groups (i.e., in percentage of “agree” answers) may be attributed to this question. The innovation of the method is to allow for accurate aggregate-level results while avoiding any possibility of linking an individual with a specific answer.154

Another experimental survey research technique varies the questions on a questionnaire in small ways so as to gauge the effect of a carefully chosen treatment. For example, in order to probe hidden racism Paul Sniderman and colleagues construct surveys that inquire about respondent views of government’s responsibility to assist those in need. In one version of the split-sample survey the scenario involves an unemployed black worker, while in another version it involves an unemployed white worker. The scenarios (i.e., the questions) are identical, with the exception of the race of the worker, and so are the two groups (which have been randomly chosen). Thus, any differences in response across the two groups may be interpreted as a product of the treatment.155

LONGITUDINAL DATA

Most surveys are conducted once, producing a cross-section of those surveyed. However, longitudinal data may be generated from a one-shot survey by asking people about past events. For example, demographers may ask female respondents how many times they have given birth, whether any of these children died, and the approximate dates of these events. In this fashion, they may reconstruct a time-series of births and child mortality for each respondent (which can then be combined to construct fertility and mortality rates for a sampled population).

Of course, researchers must be wary of asking questions about the past that the respondent only dimly recalls. Indeed, studies have shown that many memories are false, or only vaguely rooted in reality. Selective memory may privilege events that are highly salient or that resonate with the image we currently have about ourselves and about the world.

Another approach is to survey the same people multiple times, e.g., twice (perhaps before and after a planned intervention), every day for several months, every month for a year, or ever five years for several decades. This is known as a panel design.156

Surveys may also sample repeatedly from the same population (but not the same people) over a period of time. For example, the well-known National Election Study (NES) and General Social Survey (GSS) draw national samples of the American public every few years. This is generally referred to as a pooled cross-section.

Note that a panel design provides information on individual-level changes. One can tell who changed their attitudes or behavior and when, which may offer insight into why they did so. However, over time, the panel may become less representative of the population of interest. Consider a panel of citizens observed over several decades. Let us suppose that the panel was chosen randomly in 1980 (using one of the techniques of probability sampling reviewed in Chapter 4) so as to represent the American electorate, and that respondents are asked every several years about their views on various issues pertaining to racial equality. We can easily ascertain who is changing their views in this issue-area and when (approximately), and we can also track other features of these individuals (e.g., their education, income, and place of residence), which may provide clues about their changes (when contrasted to respondents whose views remained constant). However, over time the panel becomes less and less representative of the general population. (Young people are missing, for one thing.)

154 Kane, Craig, and Wald (2004), Sniderman and Carmines (1997).
A pooled time-series, with observations taken at the same intervals (i.e., every several years), provides an accurate indicator of population-level changes. Each sample is representative, which means that the composition of the sample changes as the population changes. If we wish to know how the views of the American public at-large are changing on issues pertaining to racial equality, this is a preferred source of information. But it is often difficult to piece together the reasons for changes over time; for that purpose, a panel design is more useful.

SUMMARY
If you are developing your own survey you should be aware that a good survey takes a long time to develop. Think carefully about each of the issues raised in this section. If possible, conduct a pilot-test prior to scaling up to the full sample. If you don’t have time or money for a pilot test, ask your friends – or someone from your target population – to complete the survey and then get their feedback on it. Try to straighten out points of confusion. Gauge how long it takes to complete and decide whether this is a reasonable duration. See if other improvements can be introduced. And remember that once the full survey is initiated there is no opportunity for further revision.

When interpreting survey results (yours, or someone else’s) all of the issues raised in this section bear consideration. By way of a summary, it may be helpful to consider the strengths and weaknesses of surveys as instruments for the measurement of attitudes and behaviors (whether of elites or mass publics).

Surveys are powerful instruments for measuring changes in opinion or behavior over time so long as the survey instrument remains constant over time (i.e., the questions themselves, the available responses, question-ordering, and any other feature that might affect responses to the question or the representativeness of the sample). We can tell when a population is changing its views on matters of racial equality, for example (though it is possible that some of those changes reflect changing perceptions of what is appropriate behavior rather than changes in gut-level attitudes and behavior).

Surveys are also powerful instruments for comparing the attitudes or behavior of different groups with respect to the same issue-area, for we can directly compare their responses to the same questions. The caveat is that they must understand the questions in the same way.

Surveys are less impressive as tools for summarizing opinion when a topic is not salient to the population that is being surveyed. Although respondents can render an opinion on virtually anything, if that is what they think they are expected to do, it is unclear what this opinion might mean and how it should be interpreted. Any survey of the mass public in the US that focuses on NAFTA faces this problem. By contrast, a survey of the Mexican mass public on the same subject may be more meaningful, as it is a high-salience issue in Mexico.

Surveys are also problematic for summarizing a population’s view of very complex issues like racial equality. Even if a survey (or, better yet, a series of surveys) asks a battery of questions, exploring different facets of the subject and using both abstract concepts as well as specific scenarios, it will be difficult to summarize these findings in a concise manner. We are accustomed to reading survey reports that X% of the public supports racial equality, or abortion rights, or democracy. But it is exceedingly unclear what this means, given that slight changes in the wording of the question or the number of response categories may fundamentally alter the point estimate. The reason we face this problem is fundamentally conceptual rather than empirical. Since we have no agreement on what racial equality, abortion rights, or democracy means we are certainly not going to be able to measure these concepts with great precision.

One final weakness deserves emphasis. Because they are standardized in question and response-type, surveys can inform only on issues that are incorporated into the questionnaire. The
survey researcher must first identify a problem, craft a questionnaire to explore that problem, a sample to test it, and so forth. But the surveyor will hardly ever learn of a new problem that s/he hadn’t anticipated in the first place. As Donald Rumsfeld might say, surveys can tell you what you know and what you don’t know, but they cannot tell you what you don’t know you don’t know. This is because there is (generally) no place on a standardized survey for respondents to speak freely. And this, in turn, provides a nice segue to techniques of data-gathering that are more exploratory in nature.

**LESS STRUCTURED INTERPERSONAL SETTINGS**

We began this chapter with surveys because they are the most structured method of data collection (from persons) – and, hence, the easiest to discuss from a methodological perspective. Other methods of gathering data are less structured – a feature that may be regarded as a strength as well as a weakness.

The lack of structure also means that these methods are more difficult to generalize about. An interview, focus group, or ethnography may be successfully conducted in many ways, and different techniques may make sense in different settings. It is difficult to establish a set of instructions for such varied contexts beyond the obvious admonition to “be sensitive to local context.”

Another important difference between surveys and less structured interpersonal data-gathering techniques is that the former generally employ large samples – often a probability sample drawn from a known population – and the latter generally employ small samples, which are rarely drawn in a purely random fashion from a known population. The issue of sampling is discussed in Chapter 4 and will not be returned to, except to note that the sorts of generalizations one can make when a sample is small and perhaps unrepresentative are limited – or must be hedged with caveats. Of course, one must also bear in mind that large samples and random sampling are not always possible when data gathering depends upon less structured interpersonal settings.

A final difference between the survey and less structured modes of data collection is that the former generates quantitative (“dataset”) observations while the latter is apt to generate qualitative observations that do not fit neatly into a matrix format (these are sometimes referred to as “causal-process observations”).

Of course, qualitative data can sometimes be standardized so as to conform to a rectangular dataset. Indeed, a lot of data that we think of as quantitative is a product of qualitative judgments. For example, an index of democracy such those produced by Freedom House and Polity rests on judgments about how each country fits into a coding template. An index of ideology may rest on instances of natural speech, as recorded by newspapers or parliamentary records. Qualitative can become quantitative, as discussed in Chapter 16. In this section, however, we are concerned with the initial qualitative format of the data, which usually takes the form of natural language or visual images and is recorded by the researcher.

Despite their different formats, the reader should be aware that many of the points raised about surveys also pertain to these less structured settings. For example, any time data is obtained from human subjects the battery of topics addressed in the *Introductory statement* should be covered. Likewise, other considerations for surveys may also apply to interviews, focus groups, and ethnography. This section of the chapter thus builds on the previous.
INTERVIEWS

A survey interview involves a standardized questionnaire and a limited set of possible responses. Other sorts of interviews are less structured. That is, there may be a standard set of questions that one wishes to pose to all respondents; however, there are no pre-set response categories, and respondents are urged to respond in a way that makes sense to them (so long as it addresses the question). There may be a “yes” or “no” answer, but there is also room to explain – if necessary, at some length – one’s choice. It is this explanation, as much as the simple response that is of greatest interest. This sort of exchange may be referred to as a qualitative, unstructured, conversational, in-depth, or open-ended interview, as distinguished from a survey interview.

The best interviewers are probably journalists; this is, after all, what they do for a living. So, as you study the art of interviewing think of yourself in the role of an investigative journalist. The only difference – and it is an important one – is that the social scientist is not interested in a possible headline in tomorrow’s paper or in the misdeeds of individual but rather in piecing together a larger story and an explanatory theory. (This should help to put your informant at ease.)

In seeking out respondents, you may want to begin with those who are closest at-hand. That will allow you to gather background information that could prove useful in later interviews. It will also allow you to hone your skills as an interviewer and revise your interview schedule or protocol (the set of questions or topics that you wish to raise). Interviews are often difficult to set up and may be even more difficult to repeat. You may get only one shot to interview a big shot, so make sure you’ve done some preliminary interviews prior to entering that situation.

In constructing a sample of respondents, the researcher may be less concerned with representativeness (since the sample is unlikely to be sufficiently large to generate a precise estimate of population characteristics) than with other characteristics. For example, if the focus of the research is a particular event the researcher may want to locate those who are (or were) most directly involved in that event. For those focused on inter-state conflicts, the most useful respondents are usually those who are (or were) present when crucial decisions were made. As first-hand observers, they possess critical knowledge. Another sort of knowledge is possessed by those who experience violence directly, e.g., survivors of genocide. They, too, are first-hand witnesses to history, though it is history “from below” rather than history “from above.”

Typically, the researcher will be looking for the respondent’s sense of an entire situation or an entire community, not simply his/her own story or perspective. A key informant may prove especially useful insofar as that person has intimate knowledge of a community or a topic, is willing to share that knowledge, and has credibility as a source (a matter addressed below).

Many research projects have been ‘made’ by…that rare, reflective inside informant who seems to know just about everything…and has thought about it and reflected on it… They are often marginal to the setting or scene being studied and are often seen by others…as ‘lay intellectuals,’ thinkers, eggheads, or know-it-alls.157 One must be wary of relying too heavily on any single informant, or a group of informants who share a single perspective or set of interests. But the point remains, not all informants are equal. Some will reveal things, or a breadth of detail, that others are not privy to or are not reflective enough to recognize.

Triangulation is the key to successful investigative journalism, as it is to historical investigations and detective work. When facts are uncertain, try asking around. Robert Wade reports the following experience when investigating corruption in irrigation systems among several dozen south Indian villages.

157 Johnson (2002: 110)
Only gradually, from conversations with dozens of engineers, government officials from other departments and farmers did it become apparent that a ‘system’ was a work, which probably had an important impact on how canals are operated and maintained. In particular, once some degree of trust was established, farmers often volunteered information about how much they had to pay the Irrigation Department; and while one would discount their figures in one, two or three instances, the regularity in farmers’ statements across many villages did suggest that something more than wild exaggeration or generalisation was involved….This led to cautious, always informal enquiries of officers in other departments and of irrigation staff themselves, as part of wider conversations about the sorts of difficulties they saw themselves facing in doing their jobs well. These conversations, it should be noted, were with irrigation staff from outside the area of detailed fieldwork as well as with many serving within it, and covered the way ‘the department’ and ‘the government’ worked in the state as a whole, as well as in the specific district. Some of the engineers were thoroughly disgruntled at the situation they were caught in, and since disgruntled people tend to exaggerate the reasons for their discontent, one had to be cautious about accepting details from any one person at face value. Again, as with farmers, it is the regularities in the individual comments and incidents, and the consistency in the orders of magnitude (as between, for example, what a district Collector told me a Superintending Engineer had told him he had had to pay to get a one-year extension, and what an Assistant Engineer in one Division – in another district from the first case – said in strictest confidence his Executive Engineer had had to pay to get the transfer) that gives confidence in the correctness of the broad picture. 158

A sample of interviews should include those who are likely to have different factual knowledge and different viewpoints on a topic. If the topic is political, the sample should include those of varying political persuasions. If the topic includes a set of identifiable stakeholders, make sure to interview those with varying stakes in the outcome. If the topic is more general, make sure to include those who occupy varying positions within that community, organization, or network – the high and the low, the nodes and the spokes. In this fashion, and taking account of the proportion of these informants in the general population, one may construct a rough sense of how “representative” one’s sample might be.

Interviews may be formal affairs, where an appointment is secured, a designated time is allocated, and a formal setting is employed (e.g., an office). Alternatively, interviews may occur on the fly and in a serendipitous fashion, e.g., when a researcher encounters a valued informant in an elevator or at a night club.

The interviewer may present him/herself in a professional manner. This effect is enhanced by a business suit, a dignified comportment, and a formal style of address. (In some languages, formal and informal styles of address are marked by different verbs.) It typically commences with a formal handshake and the distribution of business cards. Alternatively, the interviewer may adopt a more personal approach, dropping formalities, talking “person to person,” and emphasizing similarities between his/her life experience and the informant’s.

Each approach has costs and benefits. The benefit of professionalism is that one may be more likely to receive a “serious” response, one that is thoughtful and considered. One is also better able to retain a dispassionate view of the subject at hand, thus retaining objectivity and not overly influencing the informant’s responses.

The cost is that the informant may be more guarded, less forthcoming. Likewise, the benefit of an informal approach to interviewing is the possibility of getting beneath the surface, i.e., to a piece of information or an emotion that the informant normally would not show (at least not to an

outsider). Of course, one should be careful about lending too much importance to chance utterances, especially if the informant is inebriated.

One solution to this dilemma is to pursue both approaches, seeing what works in a given context and whether the stories that one obtains converge or diverge. You may conduct the formal interview in a professional manner and then repair to a local bar or café, where you can continue the conversation in a more informal fashion.

The relative power and status of interviewer and respondent are important matters to consider. If the respondent has higher status it may be important for the interviewer to adopt formal dress and demeanor in order to gain respect. The interviewer needs to emphasize that, despite our different positions, “We are both professionals.” If the respondent occupies a lower status, it may be important for the interviewer to adopt a demeanor that is informal and sympathetic, not forbidding. The implicit message is, “We are both regular people.” Choices in dress and comportment thus serve as tools to level the playing field, establishing the possibility of trust and reciprocity between interviewer and respondent.

Bear in mind that the interviewer/interviewee relationship depends on many factors that the interviewer can scarcely control. In addition to social status, it depends on gender roles. It depends on the extent to which the interviewer occupies an “outsider” or “insider” position. And so forth. Background factors such as these are likely to establish parameters of expected and acceptable behavior on the part of the interviewer. Trespassing across these boundaries may involve risk and is likely to attract attention away from the interview itself, so it is not something to be trifled with. Equally important, it is likely to affect the sort of answers that you get and the degree of frankness and cooperation you receive. “Interviewer effects” are likely to be more intense than in a standardized survey since so much depends upon interpersonal dynamics between interviewer and interviewee.

If social boundaries prove too confining it may be wise to enlist a confederate with the desired social characteristics who can conduct the interview. In some cases, this will be required even to gain access to an informant. For example, in some Muslim societies it would be considered extremely inappropriate for a male to ask questions of a female.

A critical decision is whether to record the proceedings – generally with audio but conceivably with video as well. Technologically, this is as easy as placing your smart phone on the table in front of you (and making sure to meet in a quiet place without a lot of background noise). It offers the advantage of preserving a verbatim recording, relieving you of the necessity of taking notes, and allowing (for you) a more natural, conversational style. A verbatim recording may come in handy at some later date if the informant disputes what s/he has said or there is a question of how to interpret a passage.

On the other hand, your informant must approve of a live recording. Equally important, s/he must feel comfortable talking in front of a microphone. Most people, even if they are accustomed to speaking before a crowd, will choose their words more carefully if a live recording is being made. They may be less likely to share intimate thoughts, criticize others, or give out potentially damaging information. In short, the interview is likely to elicit less information, especially if the subject is a sensitive one.

Before beginning an interview, make sure you have reviewed any relevant background information. Go into the interview knowing as much as you can learn about person you are interviewing and the topic at hand so that you don’t waste the informant’s time (and yours) with background facts that could have been gathered from the web. You may wish to bring some notes if you don’t have these background facts at the top of your head.

You should also have (in your mind or in an accessible format) a list of questions or topics that you want the informant to address. Think about the natural order of presentation and the
direction the conversation might take. Generally, it makes sense to begin with easier questions of a factual nature and then to proceed to those issues that are either more sensitive or more complicated.

In order to establish rapport, one is well-advised to start an interview with some friendly banter—whatever you think will make the informant feel at ease. Wait until you think the informant is ready to begin. Then proceed to discuss the issues listed in the Introductory statement (above).

It is especially important to clarify for the informant why you are interested in interviewing him/her (taking up a considerable amount of their precious time) and why the topic of your research matters. This satisfies curiosity and allays suspicion. It also serves to establish reciprocity—the basis of trust. If the informant is convinced of the importance of your research and your seriousness as a researcher, s/he may feel an obligation to assist and is more likely to trust your professional integrity (e.g., to protect his/her identity and not to quote out of context). Likewise, the informant will feel that s/he is participating in a venture that may have some impact on the world—in addition to contributing to someone’s CV or helping them complete a class assignment. Instead of the informant helping the researcher (a one-way avenue), both of you can be engaged in a project, e.g., helping to better understand a subject of mutual concern such as democracy or environmental degradation. This is the relationship of reciprocity that you wish to establish.

Of course, you don’t want to reveal too much about the nature of your research, as this may prejudice your informant’s responses. If you tell the informant that you are investigating pollution caused by a power plant, you cannot follow this up with a question about whether the power plant has produced a lot of contaminants.

The issue of confidentiality must be treated with care. It is always preferable to cite a source by name rather than as Informant #32. However, people are often less inclined to speak for the record than off the record, for obvious and understandable reasons. If an informant is unwilling to be associated with specific statements you may ask permission to include his/her name among a set of informants, listed at some point in your study. If the list is long, and includes those who occupy similar positions (and thus can be expected to have a similar knowledge base), it will be difficult for anyone to link up a name with a specific statement. Of course, you will need to be careful about how you write the narrative; specific details may give away the identity of the informant. Another approach is to agree to go “on record” with some portion of the interview and “off record” with other portions.

Once these preliminaries have been completed, you will need to set an appropriate pace for the interview. This depends upon your (and your informant’s) time-constraints and on the topic you wish to address. Given the opportunity, some people will talk for hours on any given subject. This can be a blessing or a curse. If there is a specific topic that you wish to see addressed, and time is limited, you should make this clear at the outset. Don’t let the informant waste your interview hour regaling you with stories of his/her college days. Some people will not respond to subtle hints and will need to be interrupted.

On the other hand, if time is plentiful and you want the informant to free-associate, then let the conversation wander. The benefit of an open-ended interview format is in allowing the conversation to flow, so don’t try to impose too much order on the proceedings. Doing so may discourage the informant from speaking frankly, and will restrict your opportunity to discover things that lie outside the range of your interview schedule.

An open-ended interview is ostensibly a conversation, so protocol normal to a conversation applies. Make sure you are responsive to the speaker, nodding your head, uttering brief replies (“oh really?”), and showing your appreciation in other ways. Restate key points from time to time so as to confirm that you’ve correctly understood; this also serves to maintain focus and to show interest.
Express ignorance or confusion if you wish to elicit an explicit articulation of a key point. Encourage elaboration where more detail is required.

Signpost clearly when you wish to transition to another topic so that the respondent doesn’t feel jerked around like a dog on a short leash. Where appropriate, reveal information about yourself. This should serve to build rapport but should not exert pressure on your respondent to suppress his/her views.\(^{159}\)

Make clear to the respondent that there is no “correct” or expected answer, perhaps by pointing out that the issue has aroused controversy, with many respectable people supporting each position. The respondent should feel comfortable to express his/her own perspective. Likewise, the respondent should not feel that s/he is expected to know the answers to all questions posed by the interviewer.

As the conversation proceeds, probe for signs that the respondent is uncomfortable, or may be holding out or equivocating on a subject. This does not mean that you should necessarily press forward on these subjects; ethical constraints may require that you respect an informant’s privacy. However, if the informant occupies an elite position, and the subject of your query is directly related to official duties in that position, it may be appropriate to adopt a prosecutorial approach. In any case, pay close attention to the informant’s facial expressions, gestures, and tone of voice. See if you can “hear between the lines.” Likewise, if parts of a person’s narrative don’t hold together, or don’t square with what you have heard or read from other sources, you should be able to call attention to this in a friendly manner.

In some respects, the qualitative interview is poorly suited to deal with sensitive subjects, as there is no pretense of anonymity. On the other hand, there is an opportunity to build trust between interviewer and respondent and this may serve to reassure the latter that confidentiality will be maintained. It may also make the respondent more comfortable in revealing sensitive information. Likewise, the interviewer has an opportunity to judge the veracity of the respondent, to fill in information that is hinted at but not explicitly formulated. Needless to say, these options are not available to survey researchers.

As the interview proceeds, try to assess how credible your informant is – or, more specifically, on which topics s/he might be authoritative and on which s/he might be biased or uninformed. This judgment should rest on everything you know about the informant (position, age, training,…), on everything you have learned from other sources about the topic, and on your own intuition. You might want to take notes on these matters so that you don’t forget, as the credibility of sources is one factor that may incline you to weight one version of reality over another as you construct your own narrative of events.

When completing an interview, make sure to thank the informant for his/her time and ask whether it might be permissible to contact them for a follow-up interview or simply to straighten out some aspect of the interview that you, in retrospect, find confusing. You may also wish to ask about other potential subjects whom you might contact for further information.

Gratitude is obligatory, as the respondent has just spent a good deal of time with you and may have related some difficult material. Make sure that your last word or gesture is a positive one and that you don’t give the respondent the impression that they have been exploited for the purpose of advancing your research.

\(^{159}\) Gray et al. (2007: 164).
FOCUS GROUPS

A focus group comprises a set of respondents who are brought together to discuss an issue. Rather than a dyad – interviewer and respondent – there are a handful of people who interact with one another, with cues from the focus group leader, the facilitator.

As in a one-on-one interview, there is a schedule of questions the facilitator asks the group to reflect upon. However, there is more latitude for free-flowing conversation than there would be with just two people in the room. Each participant reacts to what others have to say, and the discussion is more likely to move off in unpredictable directions.

In a focus group the role of the facilitator is minimized, at least on the surface. He or she is just one of several in the group, occupying a less obtrusive position than s/he would in a one-on-one encounter. The hope is that participants are more conscious of each other than of the facilitator. In this way, interviewer effects – e.g., affirmation bias – may be mitigated.

Another goal of the focus group is to access thoughts and emotions that might not come to light in the dry, professional atmosphere of a survey or interview. A focus group should make a subject come alive, eliciting responses that reach beneath the surface. Granted, with some subjects people are more likely to talk openly in the confidential environment of a one-on-one interview. However, with other subjects the group may provide a degree of anonymity that is essential for free-flowing conversation. Participants can blend in with the crowd.

The size of a focus group varies, though 6-10 participants is typical. Beyond that number, the discussion format is apt to become unwieldy and the facilitator will have to play a more active role in keeping order and directing the flow of discussion. Likewise, participants may feel self-conscious since they are speaking in front of a large group. In a smaller group (say 3-4), the format is also more likely to be dominated by the facilitator. The ideal size of a focus group is probably the size at which participants feel most comfortable and in which discussion can sustain itself without frequent cues from the leader. Naturally, the leader will need to intervene if the discussion takes a dilatory turn (away from the chosen topic).

More important than the size of a group is its composition. Typically, the objective is to bring together persons with shared experiences or perspectives, people who are similar to each other in relevant respects. (Ideally, they do not know each other; but sometimes this cannot be avoided.) If one were studying ideology, one would look for people who share a similar ideology. If one were studying academics, one would bring together people in the same discipline. For many subjects, it makes sense to select a group with similar demographic characteristics – age, social class, ethnicity, gender, and so forth. The more homogeneous the group, the more likely it is that participants will trust each other and feel comfortable speaking freely with each other. Likewise, they may feel comfortable articulating thoughts that would be unacceptable, or incomprehensible, in a more heterogeneous setting.

A core assumption of the focus group methodology is that opinions are formed in interaction with others – particularly, those within one’s immediate social group (people like oneself). This is why a survey or interview may not get at the truth of the matter – because the respondent does not know the truth (i.e., his or her truth) until s/he has discussed the matter with those in his/her reference group.

However resonant a particular focus group experience might be, it is best not to rely too heavily on a single session. Bear in mind that the direction of discussion may be influenced by a particularly voluble or persuasive participant in the group. Likewise, the views of one group may be idiosyncratic, even if members are drawn from a seemingly homogeneous population. Thus, it is best to iterate the process several times, observing similarities and differences across the sessions.
Even if one assembles multiple focus groups, one should be cautious about drawing inferences about a larger population. First, it is important to appreciate that the number of participants in a set of focus groups is not the same as the number of participants in a survey. Because the former are not isolated from each other the expression of their views cannot be considered independent. And because they are not independent, the true “N” (sample size) is considerably less than the number of people who participated. Second, since the setting is not controlled by the facilitator each focus group is a little different. Specifically, the stimulus – that which participants are responding to – is inconsistent, and responses correspondingly difficult to summarize. This lack of standardization inhibits one’s ability to generalize. Finally, because participation in a focus group is a time-consuming and demanding exercise (by comparison with a short survey, let us say), and because it tends to draw on smaller and highly homogeneous (but perhaps hard to define) populations, it may be difficult to draw participants randomly from a known population (or the random selection procedure may be marred by very high non-response). Without a random sample, generalization to a population is problematic, as discussed in Chapter 4.

However, the goal of a focus group is usually more exploratory (hypothesis-generating) than confirmatory (hypothesis-testing). For example, several decades ago a group of researchers conducted focus groups in the US and the UK in order to explore citizens’ perceptions of citizenship. They discovered, among other things, that discussion within the American focus groups often centered on civil rights (freedom of speech, religion, movement). By contrast, discussion in British groups often emphasized social (economic) rights. This corroborates standard narratives about American and British political cultures. Noting the shortcomings of focus groups as a tool for generalizing across large populations, the authors regard the exercise as a useful first step in a larger empirical study.160

Once a key hypothesis has been identified it may be possible to test it in a more systematic fashion with a survey, an experiment, or with unobtrusive data. Sometimes, the main purpose of the focus group is to develop a survey instrument, i.e., to understand how best to frame a set of questions on a standardized survey. Sometimes, the purpose is to develop a marketing strategy for a product or a candidate. Sometimes, the purpose is to delve into deep-seated norms and values that might not be apparent from surveys or interviews. There are lots of uses for focus groups, both in the world of commerce and in the academic and policy worlds.

**Ethnography**

*Ethnographic research – aka participant-observation* may be understood as an extension of interviewing and focus-grouping, including aspects of both. As such, much of what we have said in previous sections applies here as well.

However, interviews are limited to one or two respondents and focus groups are limited to a small group. Both techniques are circumscribed, having a pre-set period of time and a schedule of questions to address. The role of researcher and informant are clearly delineated.

By contrast, ethnography is immersive. The researcher goes to a site – perhaps living there for a period of time and in any case observing and engaging in activities similar to those of the subjects under investigation. The time-period is indistinct; the location is “local” (wherever that happens to be). The delineation between researcher and subject is blurred, for the researcher functions as both participant and observer in the context s/he wishes to understand.

As such, ethnographic research is less structured. Naturally, a particular ethnographic context may be very structured, and most are. But the structure stems from whatever rules and norms apply in that setting. If the researcher is studying a construction site and serves as a

---

construction worker in order to gather information, his/her time will be rigidly monitored according to the responsibilities of the job. But there is nothing in the method itself that mandates a particular structure. The mandate is to work, to hang out, to blend in, to "go native" – whatever that may mean in the chosen context.

It follows that there is no technical manual laying out rules for how to succeed as an ethnographer. Nonetheless, certain challenges are ubiquitous and some vague rules of thumb for dealing with these challenges may be adduced. That said, one can probably learn more by absorbing stories “from the field” than from reading a list of do’s and don’ts. So, if ethnography is on your agenda, spend some time browsing through the list of suggested readings. And talk to people with experience in the setting you wish to explore. They can probably tell you more that is useful than a general textbook can hope to relate.

We have already quoted the folk wisdom that the task of a good ethnography is to render the strange familiar and the familiar strange. Classically, the topic of an ethnography is a setting that is exotic (to the anticipated readers of the ethnography). Anthropologists study naked tribespeople of the sort one is accustomed to see featured on the cover of National Geographic. Sociologists study poor and immigrant communities and distinct subcultures within our society. This is most people’s image of ethnography, and it illustrates something important about the enterprise. After all, those of us who inhabit the educated middle classes of the West need to understand other cultures around the world, if for no other reason than the fact that our actions vitally affect those who are different from us. Presumably, the field visits of anthropologists and sociologists also serve to inform other cultures of our way of life, and thus open a two-way avenue of cross-cultural communication.

A second sort of ethnography makes the familiar strange. That is, it focuses on a subject that is common enough (in our life-world), and which we think we understand. But it illuminates some aspect of that subject that surprises us. Into this category one might place Richard Fenno’s work on members of the US Congress, Robin Leidner’s study of fast food restaurants, and Amy Best’s study of American teens (cited at the end of the chapter).

Once one has chosen a research question and a site of investigation, the researcher must gain access to that site. Successful ethnographic work depends upon entry into the community under study – i.e., being treated as an insider (if not entirely as an equal). If people don’t accept you, they won’t trust you with their secrets. (This is also somewhat true of interviewing. But it is especially true for work of an ethnographic character.)

Gaining insider status does not necessarily mean that the researcher is “one of them.” After all, your role as a researcher and your (presumably temporary) status in the community probably sets you apart from the community you are studying. In addition, there may be differences of sex, class, education, language, religion, and so forth. But you must be considered enough of an insider so that informants can speak freely, without censoring their language.

Gaining this status is the first task of an ethnographer. It may be attained by an introduction from an insider. For example, Posner began his fieldwork with the Chewa and Tumbuka peoples, who inhabit the border area between Zambia and Malawi, by requesting permission from local chiefs. Without such permission, many villagers might have been reluctant to participate in the study. If you have connections, even distant ones, you may wish to exploit them to gain entry to your community of interest.

Insider status may sometimes be attained by signing up through some formal process as a member of a firm, a group, or an organization. It may be attained by hanging out in places where your subjects pass time, and doing what they do.

Sometimes, gaining entrée to a group happens by serendipity. In one of the most celebrated (and entertaining) accounts of ethnographic work, Clifford Geertz relates how he came upon a Balinese cockfight as it was raided by police (cockfighting being illegal in Indonesia). Fleeing from the authorities, he and his wife were sheltered by some villagers, thereby gaining their trust (and occasioning some degree of good-humored ribbing). 162

Whatever the method of entrée it will probably take some time before you are trusted. Becoming part of a corporate group is a complex process. On the one hand, you probably should avoid becoming too closely allied with a particular faction, as it will impede your access to (or bias your encounters with) other factions. Likewise, try to avoid taking strong positions on matters that members of the group care about.

On the other hand, if the group is factionalized, and perhaps in a hierarchical fashion, participating in that group may require identifying with one of the factions. And your official duties (if any) may also require it. You cannot be a bricklayer and a plumber on a construction site. You cannot be a lawyer and a secretary in a law firm. You cannot be a Crip and a Blood.

Collecting data at an ethnographic site should be accomplished in as unobtrusive fashion as possible, lest it mark you as an outsider and impede the natural flow of conversation. You may be limited to taking notes at the end of the day. But each setting is different. And in some settings – e.g., a law office – there may be nothing at all unusual about recording an interview or typing on a laptop.

In any case, keeping accurate and detailed fieldnotes is an essential part of the ethnographic exercise. These should include both the specifics of what you observe and learn from conversation, and also more general observations that seem to explain or tie together the threads of people’s lives and understandings. This includes causal understandings of a situation.

Determining how long your fieldwork at a given site should last is a delicate question. Evidently, you want to have obtained as much insight into your research question as possible before leaving. Specifically, you want to fully exploit your informants’ knowledge and experience. And you want to have fully plumbed the diversity of this knowledge and experience, contacting all members of the community who might provide new information or a new perspective on your subject.

A clue to this point of completion is redundancy. At a certain point in your fieldwork you will probably feel that you are no longer surprised by what you see and hear. Things that were once strange have become familiar. You can predict what informants are going to say on various subjects. And you have fully gained the trust of your informants, so that you are pretty sure they are revealing all that they know. This is the point at which further ethnographic research on this site is probably redundant. Of course, it may be necessary to return at some later date, or to re-initiate contact with key informants. Memories may fade, facts on the ground may change, and new ideas may germinate. Keep your links to the community strong.

If you have informants with a taste for social science – or for your particular topic – it is a good idea to get their feedback on what you have written while it is still in draft form. (The exception would be situations in which your study may be deemed offensive to some members in the community, or when you might face strong pressure to change elements of the narrative to suit particular individuals.)

---

162 Geertz (1973).
UNOBTRUSIVE MEASURES

A final category of data is gathered in an unobtrusive manner, i.e., those whose attitudes or behavior are being observed are not aware of the fact or are not aware of the scholarly purpose to which it is being put. This may also be referred to as nonreactive data collection since the subject has no opportunity to react to (or be influenced by) the data-collection process.

A chief advantage of this mode of data collection is the avoidance of one sort of researcher effect, i.e., bias introduced by contact with the researcher. We have discussed the problem of affirmation bias, which subjects answer questions in a way that they think the researcher will approve or in a way that they feel is appropriate, thus disguising their true beliefs or behavior. We have also discussed the problem that researchers may consciously or unconsciously influence the subjects of their research. Finally, we should mention the “Hawthorne effect” (see Chapter 6), that the mere fact of being studied may affect the way in which a subject behaves. These biases may result from any interpersonal method of data collection, i.e., from surveys, interviews, focus groups, or ethnography.

Unobtrusive data is free of these sorts of biases and on this account may be described as more objective than data collected by obtrusive techniques. However, it is not immune from other forms of researcher bias. Arguably, the most important form of bias arises from the selective use of data, i.e., when a researcher focuses on data that conforms to his/her hypothesis and ignores data that does not. Likewise, the researcher must interpret the meaning of chosen measures, which is rarely apparent. It is one thing to count up instances of a phenomenon and quite another to interpret it.

Finally, bias may be introduced by the original method of data collection or storage, whatever that may be. Recall that in order to reach the researcher the activity of interest must be generated and preserved in some fashion. Whatever persons and institutions are in charge of the collection and preservation of this data necessarily impose their own goals and purposes, which presumably affects the kind of data that is collected. Original data-collectors are not immune to bias or sloppiness. All of these factors must be carefully considered as one attempts to judge the veracity and representativeness of data gathered by others. So it is not the case that unobtrusive measures can be considered more objective than obtrusive measures, though they are subject to somewhat different biases.

What, then, does “unobtrusive” data consist of? The category is residual, including all data-gathering techniques in which subjects are unaware of their status as subjects. This is a large category, and no one can pretend to offer a comprehensive accounting of it. Nonetheless, all data gathered without the knowledge of participants can be classified as surreptitious or ex post.

SURREPTITIOUS MEASURES

An ethnographic style of research normally involves a direct – obtrusive – encounter between researcher and the subjects s/he is studying. However, it is sometimes possible to study behavior without direct engagement.

One might observe activity taking place in a public space – a park, bus station, café, or nightclub – without participating in that activity. Alternatively, one might talk to people but without revealing one’s purpose and while maintaining a low profile. Accordingly, the researcher’s identity remains obscured and subjects are presumably unaffected (or less affected) by the researcher’s presence. (Ethical considerations occasioned by surreptitious observation are discussed in Chapter 16.) For example, Tim Pachirat penetrated the “distant and concealed” world of the industrialized
slaughterhouse, working undercover for nearly six months in order to provide a firsthand account of
the kill floor from the perspective of those who work there.163

The same opportunity is sometimes afforded in experimental studies. For example, in a
“dropped-letter” research design the experimenter leaves addressed letters in a public place to see
how frequently they are taken by an anonymous bystander to a post office box. (Variations on this
design include leaving cash in the letter, forgetting to add a stamp, and so forth.) Response rates may
be interpreted as a sign of public-spiritedness. For example, in one set of experiments researchers
place letters in classrooms where economics is taught, comparing their response rates to those from
other disciplines in an attempt to determine whether studying economics minimizes public-
spiritedness.164

In Chapter 7, we discussed an experiment of employer bias conducted by Bertrand and
Mullainathan in which hundreds of resumes were sent out in response to jobs advertised in the
Boston and Chicago areas.165 The resumes differed in one key respect: some of the names on the
resumes are distinctively African-American and the others identifiably white. The researchers
discover that applications with “white” names like Emily and Greg are more likely to be contacted
by employers for a follow-up interview than applications with recognizably black names like Lakisha
and Jamal, suggesting employment bias by race. In such settings, subjects’ behavior under treatment
and control conditions may be observed unobtrusively. (Note, however, that most experiments
involve the explicit participation of subjects in a research protocol. Likewise, ethical considerations
usually require informed consent, as discussed in Chapter 16.)

EX POST MEASURES

A second genre of unobtrusive data involves observations that occur ex post (after-the-fact). The
event of interest may be broadcast, published, or distributed in some other fashion (e.g., on the
web). It may be produced for a variety of reasons, e.g., commercial, artistic, entertainment, political,
personal expression, or academic. It may be concurrent or may have been produced eons ago. It
may consist of text, sound, film, or some other artifact. All historical research is, by definition, ex
post.

All such data is properly classified as unobtrusive so long as the researcher has nothing to do
with its creation and so long as the creators are not aware of the researcher’s presence. Let us
consider some examples, focusing our attention on hard-to-measure phenomena such as corruption,
ideology, and alcoholism.

How corrupt are elites in polities around the world? Although one can gain a general sense
of the subject by interviewing citizens, it is difficult to arrive at a more precise measure of elite
corruption for the simple reason that this behavior is hidden from view. Crossnational indices such
as the Corruption Perceptions Index produced by Transparency International measure the
perceptions of citizens and businesspeople, which may or may not reflect actual practices. An
alternative approach to this question focuses on the behavior of diplomats. Top diplomats from
every country in the world are permanently stationed in New York City, the headquarters of
the United Nations. Here, they face a constant temptation. Because diplomats and their families
enjoyed immunity from parking offences (prior to 2000) they could park their cars anywhere they
liked in this crowded city – where legal parking spots are exceedingly scarce – without worrying
about parking tickets. Some took advantage of this perquisite, and others did not (or did so
irregularly). Noting that New York City keeps records of unpaid parking fines, Ray Fisman and Ted

163 Pachirat (2011).
164 Frank et al. (1993).
165 Bertrand and Mullainathan (2004).
Miguel compares the number of unpaid fines accrued by diplomats from countries around the world. Since payment is effectively voluntary, an unpaid fine may be interpreted as a sign that the diplomat (or a member of his/her family) did not feel a moral obligation to observe parking regulations or to compensate the city for infractions. This, in turn, may be interpreted as a measure of the degree to which elite corruption is practiced, and condoned, in countries around the world.166

What is the ideology of a legislator? It is a classic conundrum, for subjects – especially those holding elite positions – often have strong reasons to camouflage their true policy preferences. Moreover, the question itself may be difficult to articulate and hence open to multiple interpretations. One person’s definition of “conservative” may be quite different from another’s. One approach relies on behavioral measures, i.e., on what legislators do rather than on what they say. One of the most important things that legislators do is to vote on bills. Thus, when attempting to analyze the ideal-points of members of the United States Congress researchers may examine patterns of correspondence among voting records, under the assumption that those who vote together share the same ideology. This is the basis for the widely used “NOMINATE” score developed by Keith Poole and Howard Rosenthal.167

How widespread is alcoholism and how has the rate of alcoholism changed over time? Alcoholics do not always declare themselves as such; they may not even admit that they have a problem. And surveys on such topics do not extend back very far, in any case. So, if we want to chart the rate of alcoholism in a country we shall have to rely on more subtle measures. One option focuses on reported rates of liver disease. Since alcoholism is a principal cause of liver disease, when disease rates fluctuate it may be interpreted as a change in the rate of alcoholism.168

Potential sources for ex post data are many and various, as our discussion suggests. Amidst this variety, one important distinction – between primary and secondary sources – deserves discussion.

**Primary sources** refer broadly to material produced by the actors under study (without prompting by a researcher). These might consist of official papers, minutes from meetings, memoranda, letters, emails, blog posts, fieldnotes, diaries, autobiographies, newspaper reports, government documents, testimony in public hearings, works of fiction, photographs, films, handbills, flyers, pottery shards, and so forth. There is no limit, in principle, to what might be considered a primary source. Parking tickets, as employed in the Fisman/Miguel study, would be classified as primary, as would Congressional voting records or hospital records of liver disease.

**Secondary sources**, by contrast, are produced by those who are studying the actors of theoretical interest. Typically, this takes the form of published articles and books, though it may also take any of the forms traditionally associated with primary sources (as listed above).

Note that the distinction between primary and secondary sources depends upon the use to which a source will be put. Imagine a blog devoted to legislative activity in the US Congress. In the context of a study of the US Congress the blog is appropriately understood as a secondary source. In the context of a study of popular impressions of the US Congress the blog is rightly classified as primary material. The primary/secondary distinction is thus always relative to some specific research goal.

Unless you are studying some very new or very neglected topic it will be fairly easy to gain access to secondary sources. For help in identifying relevant secondary sources see Chapter 11. Primary sources, unfortunately, are often more obscure. Historical sources may exist only in

166 It should be noted that Fisman and Miguel’s (2007) primary purpose in this study is to assess a causal question – whether norms or sanctions are more important in influencing corrupt behavior. However, the strength of that causal assessment rests largely on the strength of the measurement instrument.


particular archive, and are thus accessible only to those who can travel to that archive. Even then, gaining access to archives may require personal connections, and finding what you want in a large archive can be a very long process – especially if it is not well-organized or the librarians working there are unresponsive. Of course, the challenges of archival work are also what make this sort of research rewarding. One never knows precisely what one is going to discover. Some documents may have been untouched for hundreds of years. Even the librarians working there may be unaware of their existence, or their import. So, for those with vast patience who enjoy combing through diverse documents (often hard to decipher) and artifacts looking for new insights to the past, an archive offers the pleasure of a scholarly treasure hunt.

In any case, historical material is increasingly accessible to non-archivalists. Internet archives such as Google Books are in the process of digitizing the world’s printed knowledge, putting them on-line for all to view. Works published before the 1920s are generally freely available, as they are no longer under copyright protection. Those under copyright may be partially accessible, or may be available for purchase. Sometimes, the most obscure sources are only a few clicks away. In addition, libraries contain vast holdings in microform or microfiche formats (which require special viewers). So, do not assume that lack of access to an actual archive means that you cannot do research with primary documents. Books, newspapers, and other printed materials (including artwork) from another historical era or another part of the world are often readily obtainable.

### DATA ASSESSMENT

Having introduced various techniques of data collection, we turn to the problem of assessment. Is the data, as collected and assembled by the researcher, valid? Does it accurately represent the phenomenon of interest? Or is it biased in some fashion?

Social science is focused on the attitudes and behavior of human beings. But attitudes are not directly observable, and behavior is not meaningful until it is given a context and an interpretation. Consider various events of interest to social scientists such as murder, mortality, voting, coups d’etats, and consumer purchases. These are countable – and hence readily measurable – phenomena. But our interest in these phenomena is not merely in their incidence but also in their meaning, and more specifically in their causes and effects. This requires us to interpret the meaning hidden in observable behavior, raising an oft-noted problem of social science that may be referred to as the interpretive (or hermeneutic) challenge.

Sometimes, the meanings of actions are relatively easy to intuit. But often – more often than not – they are hard. For example, in investigating the topic of corruption we face the problem that those who engage in corrupt activities try hard to conceal these actions and, perhaps equally important, are often informed by different understandings of corruption. It is quite different if an act of patronage is seen as a moral obligation (e.g., to help kith or kin) as opposed to an act of self-aggrandizement. Because questions of meaning and intentionality are often central to our understanding of a phenomenon they are also central to the task of data-gathering.

The same difficulties are encountered with many other social science subjects, e.g., clientelism, crime, democracy, discrimination, economic output, happiness, human rights, identity, ideology, intelligence, nationalism, prejudice, public opinion, utility, and wellbeing. We have trouble measuring these things because actors have strong incentives to misrepresent themselves and because actors often have differing understandings of their own actions or experiences. They lie and they disagree. Sometimes, they don’t have a good understanding of their own conduct, or they
misremember things they did or said or thought. (Beware of selective memories!) So the interpretive challenge is recalcitrant.

Regardless of the technique by which data for a study is gathered, it is important to keep sight of the active role played by the researcher. Data does not speak for itself. Structured data collection procedures such as surveys allow the researcher to set the agenda. Even when subjects are encouraged to articulate their sense of what is going on it is the researcher who must synthesize this information across sources and articulate it in a narrative. Although qualitative data-gathering processes generally allow more space for the subject to express him/herself it is not clear that, in the final product (i.e., in the study itself), the researcher plays any less of a role than s/he does in a study based on quantitative data. Arguably, the only difference is that the researcher intercedes in the process at different stages. When constructing a survey, the researcher’s role is apparent in the choice of questions to include in the survey. When utilizing in-depth interviews, focus groups, or ethnography, the researcher’s role is more prominent at a later stage – when sifting through the gathered information and deciding which is relevant to the study and which should be ignored. When utilizing unobtrusive measures, the intervention of researchers generally occurs at two stages. At the primary data-collection stage someone is probably responsible for collecting the data, thus imposing his/her imprimatur. If the data accrues as part of a natural process – e.g., archeological remains – then researchers impose their stamp on the process only after the fact.

Having acknowledged the active role of the researcher in collecting data, we must also acknowledge that researchers do not occupy a position of objectivity, separate and apart from the processes that they investigate or the world that we all inhabit. Researchers, like everyone else, have material interests, emotional bonds, prior commitments, and cultural baggage that allow them to see certain things in certain ways but not other things or in other ways. Naturally, researchers try to visualize as many aspects of a problem as they can. But human flexibility, and creativity, is limited. The point is that social science is a human product, created by humans. As such, it is appropriate to regard scientists, and scientific products, with a degree of skepticism. In particular, in light of the intimate – and often hidden – role that investigators play in the collection of evidence, we must look closely at the process of data-collection to see what things may have been missed, misstated, or wrongly stated. There is as much mischief in the process of data collection as in the process of data analysis.

In the following section, the task of data assessment is broken down into three problems: divergent sources (can divergent sources be integrated?), replicability (can the data be reproduced by others?), and validity tests (does the data conform to patterns that we believe to be true?).

DIVERGENT SOURCES

Any reconstruction of reality rests on the veracity of one’s data, and the veracity of data rests in part on the believability of one’s sources. This is true whether the source is qualitative or quantitative, contemporary or historical, primary or secondary.

Typically, different sources will tell somewhat different stories about a topic. One source claims that Actor A pulled the trigger; another says that it was Actor B. One source says that British colonialism was brutal and rapacious; another says that it was relatively enlightened. The problem of evaluating divergent social science evidence is no different from the problem of evaluating journalistic, historical, or criminal evidence. Sources matter, and because they matter social scientists must judge the quality of their sources.

In making these judgments the following considerations come into play:

• Relevance: The source speaks to the question of theoretical interest.
• **Proximity:** The source is in a position to know what you want to know. They are close to the action.
• **Authenticity:** The source is not fake or doctored, or under the influence of someone else.
• **Validity:** The source is not biased. Or they are biased in ways that (a) are readily apparent and can therefore be taken into account or (b) do not affect the theoretical question of interest.
• **Diversity:** Collectively, the chosen sources exemplify a diversity of viewpoints, interests, and/or data-collection methods, allowing one to triangulate across sources that may conflict with one another.

We shall now explore these issues in greater detail, with particular attention to problems of bias. Data gathered in an obtrusive fashion is subject to researcher bias. Subjects may tell the researcher what they think s/he wants to hear, or what they think is appropriate in a given context. Data gathered in an unobtrusive fashion is usually mediated by someone other than the researcher. This fact also warrants caution. In particular, if one is viewing an event through the eyes of later analysts one must be aware of whatever lenses (or blinders) they may be wearing. Their interpretation of the activity might not be the only possible interpretation, or they may have made errors of a factual nature.

Even where primary sources are available, one must be wary of the data collection process. Consider that the main source of information about crime, rebellion, and political protest in previous historical eras comes from the official records of police investigations. Police and military authorities have a natural interest in suppressing unrest, so it is not surprising that they keep close records of this sort of activity. Thus an extensive set of records accumulated by French authorities during and after the uprising of the Paris Commune, including interrogation of key actors in the rebellion, provide the most important primary source for our understanding of that key event. Likewise for other episodes of rebellion, protest, and crime throughout recorded history. Needless to say, one would not want to uncritically accept the authorities’ interpretation of these events (though one would not want to reject them out of hand either).

A combination of primary and secondary sources should give one a more complete view of what is actually going on than could be garnered from either genre on its own. Just as one should be wary of relying solely on secondary sources one should be equally wary of relying solely on primary sources. There may be secrets that later observers have uncovered that would help one interpret events occurring long ago or far away.

But the problem of interpretation stemming from source material is only partially captured by the hallowed distinction between primary and secondary sources. It is not simply a matter of getting closer to or further from the action. It is also a matter of the perspectives that each source brings to the subject under investigation. A contemporary example is offered by Christian Davenport and Patrick Ball (2002: 428) in their research on state repression in Guatemala. As part of this research, conducted over the past few decades, they reviewed “17 newspapers within Guatemala, documents from four human rights organizations within as well as outside of the country, and 5,000 interviews conducted by the International Center for Human Rights Research within Guatemala.” Sorting through this material, they find recurring patterns. Specifically, “newspapers tend to focus on urban environments and disappearances; human rights organizations highlight events in which large numbers of individuals were killed and when large numbers were being killed throughout the country in general; and…interviews tend to highlight rural activity, perpetrators, and disappearances as well as those events that occurred most recently.”

---

169 Bourgin, Henriot (1924).
each source has a distinct window on the topic, which sheds light on some particular facet of the topic. None are wrong, but all are partial. And this, in turn, stems from the position each of these sources occupies. The authors summarize,

[N]ewspapers, tied to both urban locales/markets and authorities, tend to highlight events that occur within time periods of excessive state repression (i.e., within years in which the overall number of killings is highest). This identification/distribution occurs predominantly in an environment where the regime is not overly restrictive. These sources become useful in documenting obvious behavior or that which is deemed politically salient within a specified political-geographic context. At the same time, journalistic sources may be relatively weaker at identifying events in more remote areas that occur during periods of relatively less state repression and that are relatively smaller in scale...In contrast, human rights organizations in Guatemala tend to highlight violations where they are most frequent, most destructive (i.e., where they injure the most individuals at one time), and where the context is most dire (i.e., during historical periods when individuals are generally being killed in the greatest numbers and when political openness is limited). As a result, these sources are useful in comprehensively trying to document human rights abuses—especially those of a particularly destructive nature...Finally, interviewees tied inexorably to their homes, loss, revenge, and/or healing tend to highlight events that took place in the area with which they are most familiar...Interviewees also favor highlighting the perpetrator who abused the victim(s) and specifically what was done during the violation. As a result, such sources are useful for identifying what happened and who did it within particular locales.171

Typically, diverse sources will reveal different aspects of a problem. These differences are “tied to where the observers are situated, how they collect information, and the objectives of the organization.”172 If these sources can be combined, as Davenport and Ball endeavor to do, the researcher will usually be able to put together a more complete picture of the phenomenon under study – in this case, the location, extent, and type of human rights violations occurring within Guatemala.

Sometimes, however, observers have frankly discordant views of a phenomenon, which cannot therefore be pieced together to form a coherent whole. Occasionally, this is the product of a false document, i.e., a document written by someone other than who the author claims to be, or at some other time or set of circumstances. The authenticity of sources must be carefully monitored. This old piece of advice becomes truer still in the electronic age, as the provenance of an e-document is probably easier to forge or mis-represent, and harder to authenticate, than hard-copy documents.

More commonly, discordant views of the historical record are rooted in divergent interests or ideologies. Consider that the interests of state authorities must have come to bear in their collection of data on crime and disorder, as discussed in our previous example. The potential biases of sources must therefore be carefully judged whenever a researcher uses those sources to reach conclusions on a subject.

This is not to suppose that some sources are thoroughly biased, while others are thoroughly reliable. More typically, each source is reliable on some features of an event but not on others. It is the researcher’s task to figure out who can be relied on, and for what sort of information. Figuring this out is a matter of understanding who they are, what they are likely to know (and not know), and what their stakes and pre-conceptions might be.

171 Davenport, Ball (2002: 446).
172 Davenport, Ball (2002: 446).
Sometimes, knowing the potential bias of a source is sufficient to establish an upper or lower bound for the information in question. For example, one might surmise that any human rights violations admitted by organs of the state, or organs closely affiliated with the state, would provide a lower bound. Likewise, estimates provided by zealous human rights advocacy organizations may be regarded as an upper bound. Somewhere in between these extremes (but not necessarily in the middle!), one might suppose, lies the true value.

Note that in searching for a “consensus view” on a particular question of fact or interpretation it is not sufficient to enumerate sources according to their views. Suppose that five sources take one view of a matter and three take another. This does not necessarily offer vindication of the first view. For one thing, it is never entirely clear when one has fully exhausted the sources on a subject. More important, some sources are probably in a better position to know the truth. Others may have no first-hand knowledge of the matter, and thus simply repeat what they have heard elsewhere. So, although it is good to keep tabs on who says what, do not imagine that all testimony can be weighted equally.

The issues raised in this section are often difficult to evaluate. How is one to know whether a source is biased, and in what ways? If you are having trouble reaching conclusions on these issues, consult someone who has worked intensively with the sources you are dealing with. This sort of source-expertise – even if they know little about your chosen topic – is immensely helpful, precisely because so much of the business of sourcing is context-specific. Someone with knowledge of one historical era may be unhelpful in elucidating another historical era, for example. Someone with experience working in a particular part of the world, or working with a particular sort of research subject (e.g., trial attorneys or wholesale merchandisers), may help you distinguish between reliable and unreliable sources.

Also, bear in mind that judgments about sources are rarely final or definitive. That is why every work of social science includes a long clarificatory section focused on the nature of those sources. It is long because it is complicated. And it is complicated because sources – through which we understand the world – do not speak for themselves. More precisely, they may speak for themselves but their speech requires interpretation.

**Replicability**

All knowledge should be **replicable**. This is a core goal of science. Narrowly construed, replication means that someone (other than the original researcher) should be able to access the data used in a study and repeat the procedures of the original analysis – whether qualitative and/or quantitative – working their way from the evidence to the author's conclusions. If this is not possible, i.e., if there is insufficient information to allow for this replication, then the value of a work is limited for it cannot be verified or falsified. One must simply take it on faith that the evidence gathered for a study supports the author’s conclusions.

In order to facilitate the goal of replication it is important for researchers to preserve their data and to make it available to others once the study is complete – either on their own web site, on a journal web site, or on the web site of a data repository (see Resources). It is also important to include careful notes about the sources of that data as well as various decisions the researcher makes in collecting, coding, and analyzing data. This may be contained in the study itself, in an appendix, or in a companion document. If the analysis is quantitative, a “do” file should be included showing the set of operations performed on the data to obtain the results shown in tables and texts of the study.

Achieving replicability is often harder in qualitative research than in quantitative research. If a study is based on qualitative interviews or ethnography it may be difficult to preserve the original data collected by the researcher. Likewise, it may be difficult to re-collect the data since informants
may be no longer available or the setting has changed so much that it is not possible to repeat the study in its original context. Even so, the researcher’s field notes should be preserved and made available to others (insofar as this is possible without compromising the identity of subjects). Clear and specific citations of primary and secondary sources should help to achieve replicability in work based on sources that are already in the public domain, e.g., books, articles, material posted on the web, and archival sources. Future historians should be able to follow the trail of footnotes left by their predecessors. (This process is likely to become much easier if the use of hypertext links located in on-line texts replaces hardbound texts.)

The obligations of a researcher extend to what can reasonably be achieved, and standards may differ for different styles of research. Likewise, the researcher must bear in mind ethical constraints such as informant confidentiality, as discussed below. This may require masking the identity of informants prior to releasing data to the general public.

**VALIDITY TESTS**

In order to assess the quality of data we may also apply validity tests. Broadly speaking, a validity test is one that compares the pattern of data under review to some other pattern that we believe to be true. Common approaches include face validity, case-based strategies, convergent validity, and causal validity.

Face validity refers to an obvious or intuitive appeal. If one is attempting to validate an index of democracy one might begin by considering countries that everyone (or almost everyone) agrees are democratic or non-democratic to see if they are correctly classified. If the index classifies Sweden as a democracy and North Korea as a non-democracy, this conforms to most people’s sense of the world and may be regarded as a simple validity test. If, on the other hand, the index classifies Sweden as a non-democracy or North Korea as a democracy, one might have serious questions about the index.

Case-based strategies examine key cases to see if the coding for these cases is factually correct. For example, scholars of Central America have shown that the scores allocated to these countries by Freedom House and Polity are often patently erroneous. This sort of investigation rests on a scouring of primary and secondary sources for the countries in question, including local newspapers, government documents, and US diplomatic correspondence, as well as interviews with local informants.  

Convergent strategies attempt to validate an indicator by comparing it with other measures that are deemed to be valid measures of the same concept. A high correlation demonstrates convergent validity; a low correlation suggests poor convergent validity. Convergent validity studies have shown that the leading indicators such as Freedom House and Polity are highly intercorrelated, rendering a Pearson’s r correlation of 0.88 across all countries in recent decades, and this has been interpreted as evidence that all such indicators are valid. Of course, the operating assumption is that additional indicators of a concept are themselves valid. If the measures that compose a convergent validity test are subject to bias the technique holds little promise.

Causal strategies attempt to validate a measure by looking at its relationship to an input or output to which it is presumed to be causally related. Suppose one is attempting to validate an index measuring the freeness and fairness of elections based on reports by election observers. One also has data on the percentage of the vote obtained by the winning party. If the index classifies an election as free and fair when the winning party obtains 100% of the vote one might have reason to doubt the veracity of the index. We expect that if an election is truly free and fair, some people will

---

CONCLUSIONS

Getting good data is critical to doing good social science research. *Garbage in, garbage out* – as the phrase goes. And judging the quality of data is critical to judging the quality of a piece of research. The more you know about various data collection processes the better you will be able to produce social science and judge the findings produced by others.

To this end, this chapter introduced a variety of techniques for data collection, distinguishing between those that are obtrusive (surveys, interviews, focus groups, and ethnography) and those that are unobtrusive (surreptitious measures and ex post measures). We then discussed the task of data assessment, including divergent sources, replicability, and validity tests. For further advice on these subjects the reader is referred to the extensive list of sources listed in the Resources section below.
KEY TERMS

- Data
- Obtrusive/unobtrusive methods
- Surreptitious (ex post) data
- Survey research
- Likert scale
- Interviewer bias
- Social desirability bias
- Questionnaire
- Priming
- List experiment
- Split-sample survey
- Panel design
- Pooled cross section
- Pilot-test
- Interview
- Focus group
- Participant observation
- Primary sources
- Secondary sources
- Interpretive challenge
- Replicability
- Validity test
- Face validity
- Convergent strategies
- Causal strategies

INQUIRIES

1. Consider a set of topics such as the following: (a) democratization, (b) the “9/11” bombing of the World Trade Towers, (c) the gap in test scores between social groups in the United States or in some other country, or (d) attitudes toward foreigners. Now consider how one could apply the various data-gathering techniques set forth in this chapter to each topic. Which techniques are most appropriate? What sorts of things would one be likely to learn, or not learn, from each technique? Reflecting upon your answers to these questions, what are the general strengths and weaknesses of each technique?

A number of exercises are outlined below. Each requires that you delve into a topic using a specific data-gathering technique. In developing a topic for these exercises all the usual advice applies, as articulated in the previous chapter and elsewhere in this book. In writing up the results of your
analysis, the advice on writing laid out in Chapter 14 may be helpful. These exercises, however, focus only on the data-gathering components of the study.

2. Conduct a survey-based study. To implement the survey you may wish to use the on-line application developed by Survey Monkey or Qualtrics. Respondents may be drawn from participants in a class, from your friends and family, or from Mechanical Turk (in which case some remuneration will be required). Naturally, you will need to consider the representativeness of your sample, however it is constructed.

3. Conduct an interview-based study, following the guidelines laid out in this chapter.

4. Conduct a focus-group based study, following the guidelines laid out in this chapter.

5. Conduct an ethnography, following the guidelines laid out in this chapter. You might choose a familiar setting such as a dormitory room or school cafeteria. If so, try to identify those features of behavior that might be puzzling to outsiders (and thus require interpretation). What “obvious” features of this setting are perhaps not so obvious?

6. Conduct a study based on unobtrusive measures, following the guidelines laid out in this chapter.

7. Consider a historical topic that you have already studied using secondary materials (e.g., works of historians). This should be a fairly narrow topic such as an election, battle, legal decision, or coup d’etat. (If it is a large topic it will be more time-consuming and more difficult to wrap one’s mind around.) Now, take a look at some of the primary materials available on that topic (see chapter discussion for clues about how to access this material). In what ways does your review of primary materials confirm or alter the view of the topic that was presented in the secondary material? How do primary and secondary accounts differ in their presentation of events? Which do you find more trustworthy (and for what)?

8. Compare reporting from several newspapers or other sources on a particular event, e.g., the terrorist attack on the World Trade Center in New York (“9/11”). Your sample of sources may include sources from other countries or widely varying political perspectives.

9. Compare eyewitness reports of an event in your vicinity. That is, identify those who witnessed that event, which could be a crime, a speech, a meeting, or anything that people are likely to have observed closely. Then interview these witnesses, asking them to recount what they remember. You may repeat these interviews a week or two later to see to what extent memories fade. At any rate, compare the responses to identify (a) who remembers more and (b) who remembers differently. Can you explain differing patterns of remembrance? Which witnesses do you feel are most credible (and why)?

10. Recruit several coders and have them code the same texts. Measure inter-coder reliability across the coders. Why do you think some questions elicited greater agreement than others? How confident are you of the results? Does it prove the point that you wished to make? What are the strengths and weaknesses of this form of content analysis?
11. Group project: Conduct studies of the same topic using different methods of interpersonal data-collection (survey, interview, focus group, ethnography) and the same basic set of questions. Try to keep other aspects of the data-gathering process fairly constant (e.g., focus on the same general population of respondents). Compare and contrast the results garnered from these parallel efforts.

12. Conduct “split-sample” surveys or interviews in which some aspect of the survey/interview is altered across two groups. This means that each group should be exposed to the same survey/interview experience except for the one or two features that are altered. This altered feature should be randomly assigned so that the two groups can be compared and potential confounders eliminated. The sample should be large enough to minimize threats to inference from stochastic factors. (See Chapter 7 for a general discussion of experiments.) The point of this exercise is to test whether the construction or administration of a survey or interview affects the sort of responses one obtains. Thus, one might alter the ordering of questions, the wording of questions, the framing of questions, the sex or race of the interviewer, or any other manipulable feature of the survey or interview. Finally, compare responses from the two groups to see if there is a (statistically significant) difference and whether that difference is interpretable in some fashion.

RESOURCES

DATA-GATHERING METHODS, GENERALLY CONSIDERED

SURVEY RESEARCH
• Stoker, Laura. 2003. “Is it Possible to do Quantitative Survey Research in an Interpretive Way?” *Qualitative Methods* 1:2 (Fall) 13-16.


**INTERVIEWING**


**FOCUS GROUPS**


275
Examples:


ETHNOGRAPHY


Examples:


UNOBTRUSIVE MEASURES


INTEGRATING SOURCES


Examples:


**Repli**


• Moravcsik, Andrew. 2010. “Active Citation: A Precondition for Replicable Qualitative Research.” *PS: Political Science & Politics* 43:1, 29-35.


*Web sites:*

• Qualitative Data Repository, Syracuse University: [https://qdr.syr.edu/](https://qdr.syr.edu/)

• Social Science Research Archive of the Public Affairs Video Archive (PAVA), Purdue University

• Publication-Related Archive of the Inter-University Consortium for Political and Social Research (ICPSR), University of Michigan
Sometimes, the task of writing is regarded as separate and independent from the study of social science. In this view, writing is a skill taught in English, rhetoric, or composition courses while theorizing and analyzing are reserved for courses in social science. People who hold this view may also regard writing as an art (and hence the province of the humanities) in contrast to the task of analysis, which they may regard as more scientific (and hence the province of the social sciences). They may even be a bit suspicious of eloquent prose, regarding it as a sign of sophistry (a subtle, superficially plausible, but generally fallacious method of reasoning). From this perspective, the cultivation of style is a substitute for sound analysis — “mere rhetoric,” as the phrase goes.

We can all agree that substance should trump style, at least in the realm of social science. Yet, in order to communicate ideas those ideas must be put into words. If an author is unable to do so the reader must put the pieces of an argument together him- or herself, with much effort and possible misunderstanding. A poorly written study is one that is poorly executed. It will have less impact on the world, and — if it has any impact at all — may have an impact that is different from what the author intended. Communication skills thus matter a great deal to the progress of social science.

Moreover, the act of communicating is impossible to separate from the act of reasoning. We think through prose. This recalls E.M. Forster’s question: “How can I tell what I think till I see what I say?” In our own experience, writing out an argument usually leads to a reconceptualization of that argument. One cannot express an idea cogently without first understanding that idea, and one cannot understand an idea until one has effectively expressed it. Bad writing is usually a symptom of a deeper malaise.

Of course, everything depends upon how one chooses to define “good writing.” Some purely stylistic components of writing such as spelling, grammar (norms of usage), and word-choice are indeed rather superficial in nature and do not necessarily reflect on the logic of an argument or the thoughtfulness of the writer. Non-native writers will always face difficulties in these areas, especially if their chosen language is English – arguably, the world’s most idiomatic language.

We urge you to think about writing in a more holistic way, incorporating spelling, grammar, and word-choice along with argumentation and organization. In this sense, writing is inseparable from thinking. And in this sense, writing is not an isolated skill, like penmanship, that can be separated from one’s overall expertise as a social scientist. Writing is a vital part of the skill-set that one expects all social scientists to possess.

The first section of the chapter identifies various writing genres that you may encounter, each of which follows somewhat different rules. The second section is about organization. The third section encompasses considerations of style, including grammar. The fourth section is about integrating sources clearly, honestly, and creatively. The fifth and final section is about editing.

Prior to reading this chapter, we shall assume that you have identified a topic for research – which may be assigned or may be of your own choosing (see Chapter 12). It is worth reiterating that the topic of a paper is the most consequential part of the resulting work; do not skip lightly over this crucial decision. Likewise, we shall assume you are familiar with the criteria pertaining to social

174 Forster (1927: ch 5).

280
GENRES

There are many kinds of written documents and hence many diverse – and occasionally conflicting – criteria for good writing. As an entrée to this chapter it may be helpful to identify some of these genres, and their distinctive aims.

Perhaps the most important distinction lies between fiction and non-fiction. Evidently, we are concerned with writing that aims to represent reality in a factual manner. Of course, fiction may contain valuable truths. A compelling novel, poem, or play may be more real than a non-fiction treatment of the same material, at least in certain respects. However, because the former is not intended to reproduce a factual reality its construction follows very different rules.

Within the vast category of non-fiction we can recognize a number of sub-genres based on the venue for which a work is produced. This includes (a) short papers or theses written in pursuit of a degree, (b) academic journal articles, (c) longer monographs (including dissertations and books) written for an academic audience, (d) books written for a general audience, and (e) memos, reports, press releases, and web site content produced for a business, government agency, or nonprofit organization. What we have to say in this chapter is relevant to all of these genres. But it bears most directly upon (a) short papers or theses written in pursuit of a degree.

Another way to categorize non-fiction genres hinges on the substance of the writing rather than its venue. Here, one may distinguish the following genres: (a) general accounts, (b) book reviews, (c) literature reviews, (d) policy reports, and (e) monographs. Examples of each of these genres are included at the end of the chapter. Here, we briefly review their goals and usual formats.

A general account seeks to encompass all aspects of a topic, or at least all those that might be of interest to a lay audience (non-experts, who are presumed to know little about the topic). For example, a general account of democracy might encompass the following topics: (a) the definition and measurement of democracy, (b) the origins and spread of democracy since ancient times, (c) the process by which countries democratize, (d) the exogenous causes of democratization, and (e) the effects of democracy. This is a lot to deal with, evidently, and even a booklength treatise is likely to skim lightly over these topics. Encyclopedia articles must be even more concise. For this reason, general accounts are often employed for textbooks or for books and articles addressed to a popular (non-specialist) audience. A general account offers a point of departure for more focused work.

A book review offers a synoptic overview and discussion of a chosen book. Book reviews appear on all subjects and in many academic and non-academic venues. Typically, a review begins by summarizing the main theme or argument of the book under review, along with the evidence. The review will also seek to place the book within a larger context, i.e., as part of an intellectual current, identifying elements that are novel or held in common with that tradition. Finally, there is an attempt to identify the book’s strengths and weaknesses. Laudatory reviews emphasize the former; critical reviews emphasize the latter. But generally one finds a mixture of both.

A literature review discusses work that has accumulated on a particular subject. For example, one might review studies of democratization or of social capital. Where a large literature has developed – as on these topics – the author may seek to further limit the scope of the review, e.g., to recent work, work published in a particular subfield, or work with a narrower focus. For further discussion the reader is referred to Chapter 11.

A policy report is written to provide direction and guidance for citizens, policymakers, or members of an organization who require guidance on a topic. Prior to reaching this goal the report
may contain a good deal of descriptive, causal, and/or predictive analysis. However, its primary goal is prescriptive, i.e., to suggest a concrete course of action. Thus, a report commissioned for a government office might begin by relating the history of a policy problem, including past attempts to deal with it and evaluations of their relative success, before concluding with a policy recommendation. Policy reports, unlike work in other genres, are meant to be acted upon.

A **monograph** encompasses studies with a highly focused topic and a concise argument or theory, which might be descriptive, causal, and/or predictive. This is the mainstay of academic work, as discussed in previous chapters. It also segues neatly into our discussion of essay organization.

# Organization

Grammar (rules of usage), spelling, and other niceties of the English language should be observed, as discussed in a later section of this chapter. But this is the most obvious, and in some ways least essential part of an essay’s style. More important is a clear argument (as discussed in Chapter 2) and a logical organization.

In order to facilitate this we suggest keeping an outline of how you think the paper will proceed. This outline might be extremely detailed — including, let us say, virtually every point that you wish to make — or it might be brief and schematic. You might experiment with both approaches to see what works best for you.

Of course, this outline will probably be revised as you work your thoughts out on paper (or on the computer). This is the thinking-through-writing-process that we have discussed. Nonetheless, at any given point in time you need to have a general idea of how all the pieces of your paper fit together. You may keep this outline in a separate document or on a separate screen (if you are working with two screens), and revise it continually as you go along. All your notes (including your ideas and your citations and quotations from the literature) should fit somewhere within this outline.

One way to move from a mass of notes towards a draft is to write headings and subheadings for each idea. This will then translate into sections, or paragraphs, of the paper. In any case, it will help you keep track of the flow of your narrative.

Sometimes, one has difficulty putting together an entire outline. It isn’t clear where the pieces fit, or even whether they all fit. Rather than spending hours and hours fiddling with the outline you might try another approach, based on the sections that you envision for the paper. Sometimes, you know that there will be a section on “apples” and a section on “oranges” but you don’t know which will come first, and the appropriate ordering does not become apparent until the sections are written. In this instance, you should probably write the sections first (or at least some of them) and then return to the outline, in the hopes that these sections can be fit together into a coherent narrative. This is equivalent to working on various sections of a puzzle separately and then putting them together at a later stage.

One of the hardest tasks is discarding good ideas and juicy quotes. However, this process of tossing things out is essential. It is often what distinguishes a focused, well-constructed essay from one that is disjointed and hard to follow. You should appreciate that any creative work will generate ideas that don’t fit within the rubric of a single topic. This is inevitable. It is akin to the sketches that an artist produces prior to the final work, or the false leads that a detective pursues prior to finding the culprit. There is always some “waste product.” Do not think of it as waste, however, for it is essential to the creative process. The point is that in order to create a well-crafted product you need to be able to recognize the parts that don’t fit and have the discipline to toss them out, or file them away for a future project.
When constructing an outline, bear in mind that most social science papers follow a similar organizational prototype. The format looks like this:

I. Introduction
II. Literature Review
III. Thesis
IV. Methods
V. Evidence and Supporting Arguments
VI. Conclusion
VII. End matter

In a longer paper – say, over five pages – it is helpful to separate different sections of an argument by headings, as below. (Additional sub-headings may be added, as needed.)

INTRODUCTION
Introductions contain the body of an essay in a highly abbreviated form. Typically, introductions (a) explain the meaning and importance of the chosen topic, (b) review the literature on that topic, (c) state the main argument along with the method of analysis and the evidence that will be brought to bear, and (d) outline how the topic will be addressed.

Let us discuss each of these objectives, in turn.

Introducing a subject means, first of all, explaining that topic to your reader. Some topics are self-explanatory. Others require a long preamble and perhaps the definition of key terms. Naturally, much depends upon the audience one is writing for. In any case, one must build on common ground, i.e., features of the world – and terminology – that your readers are likely to understand and to agree with.

Typically, one says something about the significance of the chosen topic. Why should we care about your topic (and by extension, your paper)? Why does it matter? Some topics are intrinsically interesting, touching on themes that lots of people care about. Others need to be connected to things that your reader is likely to care about. For example, the topic of “party identification” seems at first glance to be a relatively obscure preoccupation of political scientists. However, you may be able to claim that the strength or weakness of party identification in a country has important consequences for politics and policy.

Even with topics that are intrinsically interesting it may be important to establish the present-day relevance of that topic. For example, if you are writing about genocide some readers may assume that this refers to events that lie firmly in the past. It may be important to remind readers that genocides have occurred as recently as 1994 (Rwanda) and – according to some observers – are occurring right now (e.g., in Syria).

The significance of a topic may derive from a particular problem that it addresses. If so, it is natural to begin with a statement of this problem. For example, if you are writing about polarization in contemporary American politics, perhaps citing some authorities on the subject and making clear why you think it is a problem. Having set the context, and drawn your reader in, you can then state how your chosen topic addresses that problem, explains it, or perhaps suggests a solution to it.

At some point in an introduction you should outline briefly what others have written about your topic. This review of the literature might occupy several sentences or, at most, several paragraphs. If a longer review is required it must be postponed until a later section of the paper (see below).

Against this backdrop, tell the reader in a few sentences or a paragraph what your thesis is,
and what sort of evidence and method of analysis will be employed to prove it. Recall that a good thesis usually contains some element of novelty. This might be the argument itself if it is at variance with what most authorities on a subject have been saying or with what most people believe. If the thesis is not entirely new – and of course no thesis is completely unique – the novelty of an argument may lie in the evidence. Perhaps a new terrain is being explored, or an especially strong test of an established theory is on display. Or perhaps material is being synthesized in a new way. In any case, a thesis is more interesting and useful insofar as it points out things that are not readily apparent. A paper should add something to the sum-total of human knowledge. You should make this clear in the introduction.

Finally, provide an outline of the paper. Don’t shy away from straightforward sign-posting techniques: “First, I will address the question of X. Next, I will…”

Introductions generally occupy at least several paragraphs, perhaps as much as two pages. They should in any case consume no more than one-tenth the length of your paper. They are, after all, introductions. If your introduction stretches beyond that consider either cutting the excess or moving it to another section.

**LITERATURE REVIEW**

Every paper has a section in which the literature on a subject is reviewed. Sometimes, this is incorporated into the Introduction or the Thesis section (above and below). Sometimes, it stands alone.

In any case, it is vital to establish what sort of work has already been conducted on a subject before the author can introduce his/her own perspective on that subject. Literature reviews establish what they say and what they have done, thus situating the author’s own work and the originality of that contribution.

The author must acknowledge his/her forbears graciously; there is nothing as discrediting as a crass and stingy review of the literature that pours ridicule on everyone who has worked on a subject. By the same token, the author must be careful to distinguish his/her work from those who came before. If one is meekly following in a well-established tradition, one cannot claim to be making much of a contribution. One should honor, but not worship, one’s forbears. This is the delicate balance – between hubris and timidity – that every writer must strike.

There are at least two dimensions to every social science literature review. The first focuses on the argument, i.e., the theory. The second focuses on the analysis, i.e., the empirics. A contribution can be theoretical and/or empirical. Most reviews of the literature offer a little bit of both.

Thus, an author might begin by tracing the lineage of the chosen theory – to what extent is it original or derivative? To what extent is he/she agreeing or disagreeing with the standard view of a subject?

Next, an author might tackle the empirical elements of previous research. What sorts of material have been examined by scholars, and with what methods? What are the strengths and weaknesses of these empirical efforts? In what ways are they limited or open to question?

The literature review may seem daunting, and it commonly occupies a good deal of time. However, it is usually the least important part of a paper. As such, you should take pains not to get too wrapped up in it. Try to summarize the extant literature as succinctly as you can. If the literature is substantial, this will require a good deal of bundling. Since you cannot afford to discuss each study individually you will need to group them according to some schema. “Group A focuses on Topic 1, Group B focuses on Topic 2,…” and so forth.

Another approach is to capture important characteristics of many studies in a table, with
studies listed across each row and their characteristics listed across each column. An example is provided in Table 14.1. Here, salient features of recent studies focused on the relationship between development and democracy are summarized, including the outcomes employed to measure democracy, the research design, the period of analysis, and the main finding.

Table 14.1: Crossnational Studies of Development and Democracy

<table>
<thead>
<tr>
<th>Study</th>
<th>Democracy Indicator</th>
<th>Analysis</th>
<th>Period</th>
<th>Finding: Development affects…</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acemoglu et al. 2008</td>
<td>Polity2; PR</td>
<td>TSCS with country FE, IV</td>
<td>1500-2000</td>
<td>0</td>
</tr>
<tr>
<td>Boix, Stokes 2003</td>
<td>BMR</td>
<td>TSCS</td>
<td>1850-1990</td>
<td>+</td>
</tr>
<tr>
<td>Epstein et al. 2006</td>
<td>Based on Polity2</td>
<td>TSCS with Markov estimation; survival analysis</td>
<td>1960-2000</td>
<td>+</td>
</tr>
<tr>
<td>Przeworski, Limongi 1997</td>
<td>DD</td>
<td>TSCS</td>
<td>1950-1990</td>
<td>0</td>
</tr>
</tbody>
</table>

Democracy indicators

- **BMR**: Boix, Miller & Rosato (2013)
- **DD**: Democracy-Dictatorship (Cheibub, Gandhi, Vreeland 2010)
- **PR**: Political Rights (Freedom House 2007)
- **Polity2**: Marshall, Jaggers (2007)

Analysis

- **TSCS**: Time-series cross-section design
- **FE**: Fixed effects model
- **IV**: Instrumental variable analysis

Finding

- **0**: No consistent relationship between development and democracy
- **+**: Positive relationship between development and democracy

**Thesis**

If the thesis (aka theory, main argument) of a paper is very simple it may be incorporated into the introduction. If it is more complex, it deserves a section of its own, or may be combined with the literature review. Here, you have space to lay out the theory in its entirety. Make sure to be clear about what you are arguing. Any ambiguity on this score will injure your cause. Recall that the purpose of a social science paper is not to leave the reader dangling, as one might in a work of rhetoric or fiction. Readers of social science have very little patience. If you wait until the middle or end of the paper to reveal the punchline you will lose some readers and annoy the rest.

Sometimes, one does not become fully aware of the thesis until one has already written a rough draft of a paper and put it aside for a few days. It is common to see thesis statements in concluding paragraphs. Once you realize this, a simple reorganization of the paper should be possible (swapping text from back to front).

If the thesis is complex, with many interacting parts, it may be helpful to construct a diagram, summarizing the key features, as suggested in Chapter 2. Regardless of how many moving parts it has, the argument should be summarizable in a few sentences. If it is longer than that, we
suspect it is too long—which is to say it needs to be pared down, simplified, or presented in a more unified manner.

Although parsimony is important, you should not feel pressured to over-simplify the argument. Where caveats, clarifications, and scope-conditions are needed, make sure that they are fully articulated. It is important not to claim too much as this will make your case less persuasive. Sometimes it is helpful to distinguish between a “home turf” where you are pretty sure your argument is correct and a larger turf where its application is more speculative. (This distinction may conform to the distinction between sample and population, as discussed in Chapter 4.)

METHOD
If the method employed in a study is fairly straightforward it can be folded into the following section (Evidence and Supporting Arguments). If it is more complicated it should be accorded a section of its own. This is where you explain how you collected your data, how you analyze it, what methodological problems you face, and how you intend to overcome them.

Note that these methodological questions are equally important in quantitative and qualitative research. In a case study (see Chapter 9), you need to explain how you chose your case(s) and what method(s) you are employing for analyzing data drawn from that case(s).

EVIDENCE AND SUPPORTING ARGUMENTS
The body of the paper is composed of your defense of the thesis. Here is where you present evidence and supporting arguments that are intended to convince the reader that you are right. Evidence is broadly interpreted, including any species of “data” discussed in the previous chapter.

Each portion of the paper should address a different facet of the author’s thesis. Generally, one saves the most important and/or the most complicated parts for last. But this is a matter of taste. Sometimes, a cumulative logic is at work, demanding that some issues be presented first and others later.

In arguing for your thesis imagine possible responses from those who might be inclined to skepticism. How might you convert this sort of reader to your argument? Remember that in order to convince the skeptics you will need to deal not only with the evidence and arguments that support your case but also those that do not. Omission of contrary evidence is generally damning to an argument for it suggests that the writer is not aware, or has not fully considered, the facts of the case. Thus, you need to show why these points are wrong, overstated, or counterbalanced by opposing arguments or evidence. Since the thesis is your purpose for writing, if you do not argue your thesis effectively you have not achieved your stated objective.

Note, however, that a social science paper is not a legal brief, a debate, or an exercise in rhetoric. One wishes to convince, naturally, but not by misrepresenting the truth. One’s purpose is to shed light on a subject, honestly and with as much completeness as you can muster (given time and space constraints). This is how science advances. There is no honor in convincing readers of a false thesis. The writer, therefore, has a strong professional obligation not to overstate the evidence in support of a thesis and to acknowledge arguments and evidence that contradict it.

CONCLUSION
One is obliged to sum things up in some manner. In a short paper, this summation should be brief. Remind the reader what you have argued and what you have proven. Clarify, as well, the limits of your thesis, its scope-conditions, if you have not already done so. If there are weaknesses in the argument that you have not already addressed, now is the time to acknowledge them.

Traditionally, the concluding section of a paper approaches the subject from a broader
Perspective, exploring possible implications of the thesis. What does your argument imply? If true, what predictions flow from it? What debates does it relate to? What additional topics might it be applied to?

Conclusions are often speculative, as you can see. They set forth ground for future research by pointing out various extensions of the subject. Here, it may be appropriate to discuss some of the thoughts and questions you had as you conducted your research – things that couldn’t be proven or that didn’t fit neatly into your paper but which are nonetheless connected to your subject and might be of interest to readers.

END MATTER
At the end of a paper one generally finds a References section, providing full citations for all in-text references. (Naturally, if you choose to employ full references in the body of your paper – as footnotes or endnotes – you do not need a separate References section.) There may also be an Appendix, or even several appendices. An appendix is typically used to provide further detail on sources, descriptive statistics, robustness tests, or other information pertaining to the analysis. Sometimes, Tables and Figures are listed at the end of the document, rather than in the body of the paper. This is a matter of choice, though our personal preference is to integrate tables and figures into the body of the paper, where they are more accessible.

VARIATIONS
Having presented the prototypical organization of a social science paper it is important to note that these features can be aggregated or disaggregated in various ways. For example, one might merge the Literature review and/or Thesis sections into the Introduction. Alternatively, one might split the Evidence and Supporting Arguments section into several sections, each devoted to a separate analysis. Much depends upon how much you have to say about each of these topics. Sections should be roughly similar in length, though this is not a strict standard. It is much more important to carve up the paper into logically distinct parts than it is to create parts of equal length.

STYLE

Social science is similar to rhetoric insofar as its goal is to persuade. However, unlike other genres, the work of persuasion in social science is carried by the logic of the argument and the strength of the evidence. Fancy turns of phrase, evocative metaphors, a compelling narrative, provocative observations – these sorts of adornments are not essential, and may detract from the presentation of a theory and evidence to support that theory.

It follows that expository styles appropriate for popular journals, or even highbrow journals like the New Yorker, are not always appropriate for social science. The job of a social-scientific study is to contribute to the development of a body of knowledge, not to entertain. Our stylistic motto might be summarized as follows: To hell with beauty, let’s try to communicate some truth.

Of course, there is no reason why the art of evocative writing must be sacrificed on the altar of clarity. One can be clear, organized, and also entertaining. However, wherever the two might conflict, intelligibility should take precedence.
The use of technical language – including mathematical symbols – also imposes a sacrifice of intelligibility, at least for those without the requisite technical knowledge to follow the argument. To remedy this problem, we suggest that everyday language be employed in works of social science wherever possible.

We recognize that a technical vocabulary is often essential insofar as it is clearer (less ambiguous) or more concise than the analogous term or phrase in everyday language. That said, it is important that writers summarize technical issues in everyday language at some point in a study – perhaps in the introduction or conclusion – so that the latter is accessible to lay readers.

Note that if social science is to have any effect at all on society we must be able to translate our wisdom into the vernacular. It is no use discovering the benefits and drawbacks of an electoral system if one cannot influence public debate on electoral reform. Knowledge about the effects of public and private investment do not bring any benefits at all if economists are the only holders of that knowledge. Whatever sociologists may learn about the sources of racism will not help anyone overcome this condition if sociologists are the sole repositories of this truth.

More generally, whatever arguments are developed in specialized venues of social science must eventually filter down to a broader audience. In order to make sure that this occurs, or at least has some chance of occurring, social science must be intelligible to the lay reader. We must do our best to bring social science to the people.

RULES

Language, in common with math, chemistry, music, and any sporting event that you can imagine, has rules. Without rules, language is meaningless; indeed, it is no longer language at all but simply a random set of words without meaning (or with a wide range of possible meanings). When writing emails, tweets, texts, and in other contexts we may apply these rules loosely. However, in a formal setting it is important to abide by the formal rules of the English language, perhaps with an occasional change of pace to provide dramatic or comic relief.

Spelling, usage, word choice, and all the delicate mechanics of language are essential to effective communication. You will not convince your reader that you know what you are talking about and have thought seriously about the subject at-hand if there are careless mistakes of punctuation or spelling. This is a serious “image” problem, and you need to protect your credibility. Matters of form are also likely to affect the substance of the argument. But even if they are peripheral, stylistic mistakes will affect the rhetorical power of your paper – your ability to persuade.

As you think about grammatical rules bear in mind that writing is not a paint-by-numbers exercise. Regrettably, it is not possible to issue a set of rules that would tell you everything you need to know about proper sentence structure. Good writing is a matter of developing sensitivity to the English language, a process that develops over a lifetime. English is also a highly idiomatic language, so rules of grammar don’t take one very far. In any case, good writing in any language involves much more than following correct rules of grammar. It involves choosing the best word from among several near-synonyms. It involves finding the right way to phrase an idea, the right organization for a set of related ideas, and the proper mix of general statements and supporting examples. This is what differentiates a persuasive and powerful essay from one that is merely grammatically correct.

While it is impossible to learn good writing by memorizing a set of rules, here are some bits of advice – drawn from a variety of sources – that are worth paying close attention to.

• Good writing is possible only if one has a good idea of the audience one is writing for. For most purposes, you may assume an audience of your peers. As you sit before the computer screen, imagine

175 “The economist who wants to influence actual policy choices must in the final resort convince ordinary people, not only his confreres among the economic scientists,” notes Gunnar Myrdal (1970: 450–1).
yourself writing to other members of your class.

- Don’t assume knowledge of specialized topics. Do your best to explain things in a way that non-specialists can understand. Then, you may proceed to technical details that only specialists will be able to appreciate. (We have already discussed the rationale for this prescription.)
- Avoid jargon wherever possible. “Jargon” refers to technical or abstruse vocabulary for ideas that can be communicated just as accurately and parsimoniously with everyday words. If you wish to vary the vocabulary in an essay by introducing an unusual word, use this word only once or twice. *Chew* can be repeated; *masticate* should be used sparingly. Granted, some technical words have no ordinary-language counterpart. These are permitted, and indeed are often indispensable. Make sure that these technical terms are carefully defined at the outset of the paper, lest you lose your audience.
- Don’t talk down to your reader. Fancy words and phrases often come out sounding pretentious. Likewise, explaining the obvious suggests that you have a low opinion of the reader. Find an appropriate voice, one that conveys respect.
- Don’t let your prose get in the way of the logic of the argument. Overly long sentences with multiple clauses are hard for the reader to follow. Try to write as simply as possible – without sacrificing the complexities of your topic.
- Colloquial phrases are sometimes funny (dammit). But they should be kept to a minimum.
- Avoid a polemical or conversational style. The tone should be even, measured, and scholarly.
- Use examples wherever a statement might not be entirely clear, or simply to avoid the arid effect of an unrelieved series of generalizations.
- Avoid deterministic language, unless it is clearly justified. Most things in the social-science universe are probabilistic, rather than invariant.
- Don’t overstate your argument. In debates and in courtroom arguments one is enjoined to give no quarter, to contest every point. Academic writing is different. Here, you are enjoined to acknowledge the limitations of your own position and the possible utility of arguments offered by others. Your purpose is to reach the truth, not vanquish opponents. And reaching the truth is usually a communal endeavor. This does not mean that compromise is always warranted, or that the truth always lies in the middle. It means, very simply, that you should worry about getting things right, not about settling scores. It means that you should indicate uncertainty wherever uncertainty is indicated, using appropriate qualifiers and caveats.
- Turn on, and pay attention to, the grammar promptings that your word processor provides. Also, take a look at this page, which clarifies a basket of words that are often confused with each other (such as *there*, *their*, and *they’re*): [www.englishchick.com/grammar/grconf.htm](http://www.englishchick.com/grammar/grconf.htm)
- Thesauruses are now available on-line or as part of word-processing programs. Don’t hesitate to use them. But don’t use them too often as it will slow down your writing and, worse, may encourage you to write in an ungainly manner. Words drawn from a thesaurus tend to be poorly chosen and stick out inappropriately in a paper. In order to use a word correctly you need to be familiar with it, which is to say you need to have seen that word in a natural context several times. In this way, you maintain control over the medium.
- Avoid brackets (“”) wherever possible. Irony is not well-conveyed by the use of a scare-quote. A new term, if questionable in some way or if under definition, may be placed in brackets when it is first introduced. (Note the use of “jargon” above.) Afterwards, it should be used without the brackets.
- Each paragraph should contain a single idea. Typically, this idea is expressed in the first sentence of the paragraph. This allows your reader to skim your paper by reading the first sentences of each paragraph, which function as headings in an outline. Of course, no one follows this format slavishly, and to do so would probably end up sounding rather stilted. However, as a rule topic sentences should not be hidden in the middle of paragraphs. They should be placed at the beginning, or (occasionally) at the end.
• The length of a paragraph is less important than its logical coherence. Some will be long and others will be short, depending upon how much verbiage needs to be packaged within that container.
• You may think of paragraphs as separate slides within a PowerPoint presentation; each should address a different facet of the argument. Establish breaks between paragraphs when you move to a new idea.
• Work hard on your transitions from one paragraph to the next and from one section of the paper to the next. If there are no transitions, your reader will have difficulty following the narrative.
• State your points as concisely as possible and avoid redundancy. Delete words that are not needed, perhaps because they are implied by other words. Replace phrases with words, where possible. Alter sentences so that they are stated positively rather than negatively (eliminate not).\footnote{Williams and Bizup (2014: 186).}
• The main point of an essay should appear in the introduction, in the conclusion, and—in varying ways—within the body of a paper. This is justifiable redundancy—although each appearance should be phrased somewhat differently. Other points should appear only once in the course of your essay. If you find that a given issue is treated on several occasions, you should think about reorganizing the essay to eliminate this redundancy. Another way to deal with this problem is to refer back to earlier points (“as stated above”). This relieves you of the necessity of repeating a point ad nauseam while allowing you to point out continuities and connections.
• Differentiate clearly between (a) what others say, or what is generally understood to be true about a subject (background knowledge), and (b) what you—the author—say about that subject (the argument). The first is established in the introduction or the literature review section of a paper; it provides the point of departure. The second is the author’s contribution to that subject—where s/he extends or contests established wisdom. If you confuse (a) and (b) your readers will be confused and perhaps also annoyed. Consider the difference between “Alcoholism is a principal cause of unemployment,” and “I argue that alcoholism is a principal cause of unemployment.” The first statement suggests that this is a truth that most people—or at least most knowledgeable experts on the subject—accept. The second statement suggests that this is the author’s perspective, which knowledgeable observers may contest, and which the author will try to prove in the course of the essay.
• Length is probably the least important element of any paper. Granted, one is often constrained to work within arbitrary page or word limits set by a journal or by an instructor. However, bear in mind that this is an entirely arbitrary matter, established for convenience. Above all, do not confuse length with quality. Longer is not necessarily better; indeed, it may be worse. Pascal once apologized to a correspondent, saying “The present letter is long, as I had no time to make it shorter.” Writing concisely usually requires more care and attention than writing at length about a subject. One must pick and choose.

To summarize, the object of your paper is to persuade the reader, to communicate. So try to be as clear and straightforward as you can, without trivializing your ideas or patronizing your reader. The secret to what is generally regarded as good nonfiction writing probably has less to do with prose style than with clear thinking.

**Sources**

Ideas come from somewhere; they are not invented out of whole cloth. (If they were, they would probably be pretty absurd, and wouldn’t constitute good social science.) So don’t be ashamed of taking ideas from other places. Everyone does it. Social science is theft.

The point is to give full attribution wherever borrowing occurs if the piece of information is...
not already common knowledge. (You don’t need to footnote that the sun rises in the East and sets in the West.) This is a matter of honesty, as well as a matter of scholarly cumulation. Recall that your paper builds on a skein of existing knowledge. Only through accurate documentation can the reader distinguish the writer’s original contribution from those of others. Likewise, your arguments build on evidence, and in order to establish the veracity of the evidence you present readers must be able to trace it back to a source. Correct use and acknowledgment of source materials is therefore vital to any research project.

Thus, if you use material drawn from something aside from your own first-hand experience, and the material is not common knowledge, give credit to your source. If you quote directly, even a word or phrase, use quotation marks and a citation. If you paraphrase (i.e., take the ideas and put them into your own words), cite the source.

If you take ideas or words from sources without attribution you may face disciplinary action – including expulsion – from a college or university, or you may lose your job. You should also bear in mind that the current state of information technology assures that a paper that you write today may remain in the public domain for the rest of your life. This means that decades from now someone may discover an episode of plagiarism, committed in your youth, which jeopardizes your position, not to mention your standing in the community. There is a lot at stake. So, if you are in doubt about whether a citation is necessary, play it safe by citing the source or consult your instructor for further guidance. Once your paper is turned in, the reader has the right to assume that whatever appears in the paper, unless otherwise indicated, is your own work or is common knowledge.177

Naturally, even with appropriate citation you do not want to take your entire argument from someone else. What you should be taking from your sources are bits-and-pieces: a fact here, a point there – whatever bears upon your argument. You will need to refer to multiple sources; otherwise, you can hardly avoid relying excessively on one person’s work. Creativity, in this context, means putting together the material presented in the text in a new way in order to answer a question that is at least slightly different from the authors’ point of view.

In citing evidence, don’t simply cite an author’s view that such-and-such is true. Research is not a polling of authors. If four out of five authors say something is true it still may be false. Of course, it may be helpful to establish what the prevailing wisdom on a topic is. However, in bringing evidence to bear you must be sensitive to whether a particular source is authoritative. An authoritative source is a source that is, for one reason or another, well-suited to weigh in on a given topic – an eye-witness, an expert, and so forth. For these sources, direct quotations may be appropriate.

Even so, such authorities must often be viewed with suspicion. In general, you should avoid quotations, especially long ones. Try to paraphrase instead (put things in your own words). This, of course, still requires a citation.

Note that the purpose for which a source is being cited determines what sort of source is most useful or most authoritative. If you are attempting to demonstrate that a certain mood pervades a society, or that a certain event received a great deal of attention then citing popular media (newspaper, wide-circulation magazine, best-selling novel, television report) may be the best source. You might also cite an academic study that studied these popular media in a systematic fashion. If, on the other hand, you are trying to demonstrate a non-obvious point about the world – a descriptive, predictive, or causal inference – then an academic source is probably more authoritative than a popular source.

177 This statement is adapted from a document entitled “Use of Source Materials,” Pomona College Department of Government, Claremont, CA.
Although most sources are in written form they may also be in the form of personal communication with the author (interviews, discussions, and so forth). This raises a final, but extremely important, point: if you talk with friends and classmates about your paper and this discussion leads to an exchange of ideas (substantive ideas, not just stylistic/organizational ones) you need to cite these sources just as you would a book or article. If you got the idea for an argument from Cindy Walker, Cindy Walker should appear in a citation where this argument is presented. Otherwise, you are plagiarizing.

How many sources/citations are necessary? This is an oft-repeated question, to which our oft-repeated response is: it depends. It depends on what it is you are trying to prove, on what sources are out there, and on whether it is possible to cite one or two sources as examples of what is out there. Citations, like pages of text, are not to be judged by their quantity. More is not necessarily better. There are over-referenced papers and under-referenced papers. However, you are more likely to be sanctioned for the latter than for the former. So, if you must err, err on the side of over-referencing.

Note that if you find a well-referenced article or book that reviews the academic literature on a subject it may be sufficient to cite this one source, rather than all the additional sources that are cited therein. You may indicate in your citation that this particular source offers a good review of the literature (“for a comprehensive review of the literature see Smith 1989”).

**Quotation Formats**

If a direct quotation exceeds a sentence or two you should set it off in the text in a block quotation, such as the following:

>This is a block quotation, with larger margins than the rest of the text. Sometimes, a smaller font is also employed, as it is here. In any case, no quotation marks are necessary. Simply type the quotation into the block, and include the citation as you normally would (Smith 1989: 45).

In adapting a quotation for use in a paper you may need to alter it in small ways. If you drop words from the quoted passage, indicate the missing text with ellipses. For example, Smith (1989: 45) writes, “Ellipses are important…but sometimes ignored.” The deleted words are situated in between ‘important’ and ‘but’.

If you insert words into a quotation, this is signaled by square brackets. For example, Smith (1989: 45) writes, “Ellipses [in a published paper] are important…but sometimes ignored.” The added words are in brackets.

If a quotation includes a misspelled word or grammatical error you should indicate this by inserting (sic) after the error. For example, Smith (1989: 45) writes, “Ellipses are important… but sometimes ignored (sic).” This indicates to the reader that the error is in the original.

Occasionally, you may need to employ quotations marks within a quotation. This is handled with single quotes. For example, Smith (1989: 45) writes, “You should remain financially independent. Following the words of Shakespeare, ’Neither a borrower nor a lender be.”

**Citation Formats**

There are many citation formats. You should learn the format that is most common in your field, or that which your instructor advises. As a default, you may follow the formatting style used in this book – described below – which is similar to most formats used in the social sciences today.

In the text, or in a footnote, list the author’s name in parentheses, followed by the year your edition of the work was published, followed by the page number of the quotation or idea you are citing (Smith 1989: 45). If you are citing a whole book, which is to say an idea or argument that
consumes an entire book, then you may omit the pagination (Smith 1989). Information from several sources may be combined in a single parenthetical note (Smith 1898; Washington 1945). Within a parenthetical citation, author last names are alphabetized (Smith comes before Washington). A work with multiple authors should cite each author (Smith, Wilson, and Crane 1989) unless the number of authors is greater than three, in which case cite only the first author followed by et al. (Smith et al. 1989). A citation drawn from a source without an author may be cited by the name of the organization or journal that published the work (The Economist 1989: 45). Information obtained from a personal communications should be cited in a footnote and needn’t appear in your bibliography. For example, Footnote: Roger Smith, personal communication (5/31/1989). When a person has authored several pieces in the same year these may be distinguished by letters (Smith 1989a, 1989b, 1989c). Very long lists of citations, or extensive substantive comments of a parenthetical nature, should go into footnotes. Do not use endnotes, unless instructed to do so (they are hard to follow).

At the end of your paper include a References section including all works cited, with complete citations, as follows.

Books:

*The same author with several works published in the same year:*

*Edited books:*

*Book chapter in edited volume:*

*Works with multiple authors:*

*Works without authors (collective authorship):*
“Great Ideas that Aren’t Really so Great.” 2004. The Economist (August 15), 44-46.

*Translated works:*

*Newspaper articles:*

*Journal articles:*

*For the web:*
Work that has appeared, or will appear, in printed form (e.g., the New York Times on the web), can be cited as if it were printed material (as above). If there is no printed version, or the printed version has different pagination or is otherwise altered from the web version, construct a bibliographic entry that approximates your entry for books and articles. At the end of the entry, give the exact web address from which you downloaded the material and the date that you downloaded it.
In handling citations you may wish to employ software that is either incorporated in your word processing program or can be imported to it. Popular citation software (aka bibliographic software, citation managers, or reference managers) includes BiblioExpress/Biblioscape, Endnote, Mendelay, ProCite, RefWorks, Reference Manager, and Zotero. These programs import citations from databases and websites, build and organize bibliographies, format citations (according to your choice of format). They may also allow you to take notes on articles and to save other files (e.g., PDFs). Note that most of these programs are proprietary, so you will need to purchase the software or employ a site license.

EDITING

The skill of writing is learned primarily by writing, not by reading about writing. That is why this is a short chapter rather than a long one (though there is no harm in reading longer treatises such as those listed at the end of this chapter).

However, the act of writing, by itself, is unlikely to advance your skills. You also need feedback. You should look closely at the comments you receive from your teachers. Request that they comment on the form of your essay, not simply its content.

Be aware that very few writers – even professional writers – get it right the first time. Good writing depends upon re-writing. Usually it is better to write quickly, in a stream of consciousness, rather than to slave over sentence structure the first time a sentence is formed. Editing, in any case, is essential. Careful proofreading can tell you where things work and where they’re not so clear or convincing. Reading the paper aloud to yourself may also alert you to stylistic problems that look innocent enough on the written page but sound awkward or confusing when spoken.

If you can enlist a friend to read the paper – one who will give honest feedback without sparing your feelings – this is often extremely helpful. By the time you have written a paper you are perhaps too familiar with the subject matter to be an objective judge of your communication skills. Someone with no knowledge of the topic is in a better position to tell whether you have done a good job of getting your point across. If he or she doesn’t get it, or has to struggle to understand it, then you have done a poor job.

As you proofread, make sure that you are actually addressing your thesis in some way in every paragraph. You may wish to keep a sentence-long copy of your thesis taped to a wall next to you as you write so that you can remind yourself of your argument. This is a good way to identify and eliminate dilatory points.

CONCLUSIONS

By way of conclusion, let us review the most important criteria of all social science papers, whether written for a classroom assignment, for the completion of an advanced degree, or for publication. These may be summarized in the following questions.

1. Is there a thesis and is it clearly stated?
2. Is the thesis significant? Does it matter?
3. Is the thesis adequately argued or proven (within the constraints imposed by time, resources, and available sources)?
4. Is the study innovative with respect to theory, method, or findings? Does it expand our
knowledge of the chosen subject?

5. Is the essay well-written, i.e., cogently organized, grammatically correct, stylistically elegant, and citing all appropriate sources?

Flaws in one of these categories may be compensated by virtues in others, but generally speaking an essay must satisfy all five in order to qualify as excellent. This chapter has focused primarily on the fifth, though we have stressed repeatedly that matters of form can never be neatly separated from matters of substance.

Writing is the perhaps most important basic skill that the social sciences and humanities impart, and the skill upon which the liberal professions (law, medicine, academics) and business still depend. Getting good grades, getting into the graduate or professional school of your choice, getting a job, and succeeding in that line of work will rest, in part, on your ability to put your thoughts on paper (read: computer) clearly and persuasively.

Freshman English classes give you a start on the road to good writing. However, writing must be practiced continually or the skill atrophies. It is in your interest to pay attention to writing as a craft, and to practice that craft as frequently as possible.

Whatever its role in enhancing your career, good writing is also inherently rewarding. If you wish to influence the world – to change the way people think about something, to offer a new idea or solution – you will need to find a way to communicate that idea to the public, or to some portion of the public with a special interest in what you have to say. Doing so will probably require you to put your words into a coherent, organized format, e.g., a memo, article, essay, or book. If you think that your argument matters then you should be convinced that the communication of that argument matters. One without the other will not go very far.
KEY TERMS

- Book review
- Literature review
- Policy report
- Monograph

INQUIRIES

1. Exchange papers with a friend and comment on each other's work. Or, as a class, exchange papers with each other (perhaps with names removed, so that the review process is anonymous) and comment on each other's work, following the guidelines set out in this chapter.

2. Find an unpublished paper on the web on a topic of interest to you. Make sure it's a paper that's written in a manner that is understandable to you (not too technical or requiring a great deal of background information). It may be written by an established scholar or a lay scholar. Comment on the paper following the guidelines set out in this chapter.

3. Find a paper that you have written recently and that you feel represents your best work, or best work to-date. Grade your paper, that is write comments (both positive and negative) and assign a grade, relying on the criteria laid out in this chapter. (Your comments may repeat, or depart from, those you may have received from an instructor or a reviewer.) Now rewrite the paper, making improvements as suggested by your comments. In this exercise, we are primarily concerned with stylistic issues, though you should try to address substantive issues as well (so long as it doesn’t require a lot of additional research).

RESOURCES

RESEARCH AND WRITING


Web sites:

- Harvard Writing Center: fas.harvard.edu/~wricntr/html/tools.htm
- Bartleby.com (various resources)
- Dictionary.com (on-line dictionary and thesaurus)

CITATION STANDARDS FOR DATASETS


REFLECTIONS ON WRITING


WRITING CENTERS

- Most colleges have a writing center, devoted to improving students writing. Do not think of this as a place only for “remedial” cases. If the staff is good, they should be able to offer assistance at all levels. Likewise, do not wait for an instructor to advise you to go there. If you are interested in improving your writing, go there for yourself and check it out. See what they have to offer (and bring a paper that you have written).
EXAMPLES OF WORK IN VARIOUS GENRES

The following examples are drawn mostly – though not exclusively – from the three topics we have followed throughout this book: worker-training programs, democracy, and social capital (see Chapter 1).

GENERAL ACCOUNTS

- See also Wikipedia articles on various subjects.

BOOK REVIEWS


LITERATURE REVIEWS


POLICY REPORTS

MONOGRAPHS

15. **SPEAKING**

Having surveyed the craft of writing, we turn to the craft of speaking, the oral form of language. Good speaking is somewhat different from good writing. Of course, one can simply read a prepared text, in which case the difference virtually disappears. But this does not generally qualify as good speaking in an academic or professional context – unless, that is, it is carefully honed to appear as if it is extemporaneous.

People expect a “live” performance. Your presence should be real, un-mediated. Naturally, it may be mediated by various technologies and it may be pre-recorded; but it should feel as if it is happening right here and right now. This is the dynamic quality of public speaking. It is a special quality that is probably hard-wired in our brains and therefore carries a resonance that cannot be simulated with prose.

Public speaking is inextricably linked to comportment, i.e., how you carry yourself. Every time you make a public intervention you convey a vision of yourself. It is this persona that people tend to remember. Thus, when we say “speech” in this chapter we intend to include all the visual cues that accompany speech – dress, gaze, posture, gestures, and so forth.

In ancient times, speech was the preeminent art of persuasion. Rhetoric meant speech, and only secondarily prose or poetry. That prioritization is easy to understand in the context of a predominantly oral culture.

Nowadays, the craft of public speaking has fallen into desuetude (though some writing courses also include a component devoted to public speaking). People still talk, but speech is no longer cultivated as a professional activity, with a few exceptions such as moot courts in law school. This is unfortunate because speech is no less important today than it was a century ago. Perhaps destined to become more important over time as on-line lectures and Youtube videos replace written texts, and video calls and video conferences replace email. The spoken word may turn out to be mightier than the written word.

In any case, whether you are a good speaker or a poor speaker is likely to affect how successfully you can get your ideas across – not to mention getting good grades, landing a job, and succeeding in your chosen profession. To be sure, college courses generally don’t allocate much credit for participation. Nor is an employer likely to admit that s/he hired someone because they present themselves well. However, we suspect that oral presentation matters a lot more than instructors, or employers, are willing to admit – or are even aware of.

Following is some advice culled from various sources and from our own experience. This advice is framed primarily around the task of delivering a formal presentation – a talk or lecture. At the end of the chapter we discuss the complementary role: how to participate in discussion as an audience member.

Be aware that everyone comes to the task of public speaking with different strengths and weaknesses. For some people, fear of speaking before an audience is the main obstacle; for others, unfamiliarity with English. Some have lots of experience; others do not. Tailor the following advice to your circumstances.
YOUR PUBLIC IMAGE

Your public image must fit the setting, i.e., the audience and the topic. If your topic is genocide, don’t wear something garish. If the setting is professional, dress professionally. If you’re not sure, it probably won’t hurt to over-dress. Likewise, adopt a form of address that is consistent with the setting. Don’t speak colloquially unless the occasion is sufficiently relaxed to allow for it.

People are sometimes offended if the speaker is more informally attired than the audience, or if the speaker adopts an informal tone. These visual and verbal cues may be taken as a sign of disrespect, especially if you are younger than most members of the audience or from a different social class or cultural background. The greater the distance (chronological, cultural, economic, whatever) between you and them, the more important it is to convey respect in your choice of clothes, your bearing, and your words.

At the same time, the most important part of establishing a public image is finding a stage presence that feels comfortable to you. After all, you’re the one that has to wear the suit (or whatever it is that you choose to wear) and you’re the one who must inhabit the persona that you establish. If you are comfortable you will be relaxed, and the audience will likely be too. What makes audiences squirm is when they can see that the speaker is uncomfortable. So, be comfortable with yourself and try to convey that comfort to the audience – without being arrogant. A fine line separates confidence from conceit.

FORMAT

Most talks in academic and professional settings last 30-40 minutes and are followed by a question-and-answer period, which may last a half hour or as long as people remain in the room. Even if your host tells you that there is no time-limit you should still impose a limit on your talk. Our own experience suggests that 30-40 minutes is about the upper limit of what most people are willing to sit through in an academic or professional setting. After that, they get fidgety. If you go on longer you are likely to lose the audience – literally and figuratively. You will also lose the opportunity hear their input because you have used up all the time available.

So time your talk carefully. And remember that it usually takes longer to do things “live” than in a practice session. Cut material mercilessly, honing things down until you can get through your talk in an unhurried fashion in the allotted time.

Naturally, the protocol obtaining at a talk will vary according to the setting. Among economists, it is common to pepper the speaker with all sorts of questions – often critical – which may be launched at any point during the talk. In other settings, audience members generally limit themselves to points of information, waiting for the speaker to finish before offering comments and criticism.

The smaller the group the greater the opportunity for informal back-and-forth between speaker and audience. Sometimes, the assembled group is so small that there is little separation between speaker and listeners; it is like speaking to friends in a living room.

Make sure that you know what the established format is and that your presentation conforms to it (more or less). If the ground rules are unclear and you wish to establish some parameters, make sure you do so at the beginning of the talk.
Nowadays, presentations are usually accompanied by Powerpoint slides or some equivalent technology. For most settings, including most academic settings, this is strongly advised. Powerpoint allows you to present the main ideas along with other visual materials such as graphs, data tables, pictures, perhaps even video clips. This is all to the good. But don’t let the slides get in the way of the talk, and don’t let it replace you as the center of attention. It’s ok to glance at the screen from time to time, to draw the audience’s attention to it. But we suggest focusing mostly on the audience, encouraging them to watch you along with the projected images.

It is worth spending a bit of time putting your slides together. Think carefully about each slide and try to minimize the number of words and images on each one. Concision is even more important for slides than it is for prose. Try to group related ideas on the same slide, so as to limit the number of slides you show.

You may prepare additional slides as back-up, in preparation for specific questions that your presentation does not cover. These can be kept in a separate file or at the end of your slide-show. They are the presentational equivalent of an appendix.

Graphs, figures, and pictures are often helpful – if they contribute to your argument. But don’t feel obliged to insert colorful filler if it does not relate to the argument. This will detract from your task and you are less likely to be taken seriously.

Likewise, if the statistical tables you are presenting are too large to fit on a slide (and be read by those in the back of a room), consider printing out copies for members of the audience, or at least for those who may be interested. In some settings, this is important. You don’t want to find yourself in a situation where you have to answer queries by saying “Trust me, it’s in the fine print.”

There is an old saying, variously attributed, about the game of golf. “Golf,” says the pro, “is a game of luck. And the more I practice, the luckier I get.”178 This is true of most things, and it is certainly true of public speaking.

If you are faced with an important occasion such as a class presentation, you probably should practice your talk multiple times. You may even wish to write it out as a script. However, you certainly don’t want to memorize the script, as it will probably come out in a “canned” fashion (unless you are an extremely skilled performer). Instead, make sure you become familiar with the key points, the key terms and phrases, and have your notes on-hand as back-up if you freeze on the spot. The best way to practice is to give the talk to friends, a mock-up of the real thing. Try to make this as realistic as possible so that you can simulate the experience. If you expect questions from the audience, instruct your friends to ask questions (without sniggering).

178 See http://quoteinvestigator.com/2010/07/14/luck/
KNOW YOUR AUDIENCE

With speaking, as with writing, it pays to know your audience. In a classroom setting, you have the advantage of intimate familiarity. Of course, you don’t want to pander, even if it pleases the room. While your classmates may be looking for diversion, your professor is probably looking for you to grapple with a social science topic in a serious fashion. So, bear this in mind.

Likewise, in a public setting where you don’t know everyone – perhaps you don’t know anyone – think carefully about the sort of person that is likely to be present at this event. What is their background? How much do they know about your subject? What will they find interesting, and perhaps unexpected? What will they find humorous? What, more importantly, might they find offensive?

It helps to know in a general way what your audience is expecting to get from your presentation. Of course, you may not be able to give them precisely what they wish. But knowing what they want will help you craft your talk. At the very least, you can address their expectations and try to re-frame them to suit your own interests and capacities, and your own sense of what is important.

In any case, you should be clear at the outset about what you aim to deliver in the talk, and make sure you achieve those goals. Set your own expectations. This is usually sufficient. Un-met expectations are much more problematic when a talk fails to deliver on its own terms; that is, questions or problems raised at the outset or suggested by the topic are not adequately addressed.

SPEAK NATURALLY (AND, IF POSSIBLE, GRAMMATICALLY)

Spoken English is different from written English (as, we assume, the spoken version of any language is different from its written version). When speaking extemporaneously one can scarcely avoid grammatical mistakes. One may not find exactly the right word at the tip of one’s tongue. This is perfectly understandable. However, your goal should be to emulate correct speech.

Minimize your use of “ugggh” and other space-fillers. Don’t be afraid of dead air in your talk. No one should talk continually for a half-hour, and the introduction of a few open spaces will give your audience a chance to breathe (figuratively) and will heighten the drama of what follows.

Try to suppress colloquial expressions. For example, employ the locution “like” only when comparing things – not, like, when pausing in the middle of a sentence.

Emulating correct written English does not mean that you must speak in long sentences with lots of polysyllabic words. Extemporaneous speech should be natural, not stilted. If you use words that are usually encountered only in a written form your audience is likely to be confused.

RELATE TO THE WHOLE ROOM

Do your best to relate to the audience (and to the camera if it is being filmed). This means scanning the room to make eye contact with as many people as you can as you speak. Of course, you may not actually be making eye contact. Generally, when one is speaking to a large room one tries not to focus on anyone in particular, as it tends to be rather distracting. Instead, let your gaze wander about the room so that everyone is made to feel involved. If you look only at one part of the room, those
sitting elsewhere are bound to feel neglected. They may even feel that the speaker does not care for them, or that the speaker is playing favorites. Of course, there are generally some persons in the room who are more active in asking questions and making comments than others, and these windbags may be sitting in the same place. This is all the more reason to scan the rest of the room, so others feel that they are part of the enterprise and that the speaker is not responding only to the blowhards.

Another way to engage members located far away is to wander up and down the aisle, or pace back and forth at the front of the room. A moving target enlivens the show. (But there is no need for stage antics.) Make sure that those in the back of the room can hear you. One may even start a talk by asking this question: “Can you hear me?” This also gets the back of the room engaged, sending a message that the speaker recognizes them and cares about their participation.

**ENTERTAIN (A LITTLE)**

In a public speech (as opposed to an academic paper), people generally expect some divertissement – a joke, a story, a bit of narrative to relieve an otherwise tendentious argument. This will help you win your audience over, and keep them awake. Typically, the speaker opens with a joke.

Years ago, one of the authors attended a talk by Peter Skerry, a noted political scientist. After being introduced, Skerry remarked: “Some of you are probably expecting me to open with a joke. Well, I don’t have one. So, if anyone feels obliged to laugh, now is your opportunity to do so.” This brought guffaws. He had won us over by calling attention to the absurdity of the begin-with-a-joke protocol.

Of course, there is a well-grounded reason for beginning with a joke. Like most rituals, it plays a sociological function. Humor is well-suited to establishing rapport and relieving situations of stress. The beginning of a talk is a situation in which these two functions must be performed. Otherwise, everyone feels awkward.

Jokes are funny when they play off context. Canned jokes generally fail, especially when the teller is not a professional comedian and the setting is not a comedy club. So, in introducing humor to your talk, think about what is funny in the setting in which you find yourself.

Further back in time, one of us attended a talk by Mervin Field, the founder of the Field Poll, a leading polling firm in California. Field was being introduced by Merrill Shanks, and Shanks was reading from a page with a long list of accolades, which seemed to include everything Field had accomplished since he was toilet-trained (which was quite a lot). This went on for some time. Finally, Shanks finished and gave the floor to Fields. Fields shuffled up to the front of the room, leaned over to Shanks with his hand cupped in a mock attempt to shield his words from the audience, and stage-whispered: “I hope you didn’t have any trouble reading my handwriting.”

Jokes are fun when they’re funny. However, you should not allow humorous comments to upstage your argument. An academic or professional setting is quite unlike other settings insofar as the main focus must remain the subject matter. Some speakers are able to integrate humor and story-telling seamlessly into a talk without drawing attention away from the subject matter. But this is a finely-honed skill and it is easy for the novice to get carried away. When in doubt, opt for a sober presentation rather than a hilarious one. Otherwise, people will come away from your talk guffawing, but they will not take you – or your ideas – seriously. You want them to remember the punchline of your talk, not the punchline of your joke.
CALM YOUR NERVES

For those who are inclined to get a bit queasy when we look out over a hundred heads, here is some advice.

If you mis-speak – forgetting a noun or mispronouncing a word – correct yourself and move on. There is no need to call attention to it. If appropriate, you may wish to joke about it. No one will care about these verbal slip-ups, so do not allow them to distract you.

As you look out over the room people will be in various stages of attention and inattention. Some will be eyeing you closely; others will be fiddling with their phones or reading. There is not much you can do about this, and you shouldn’t imagine it reflects upon you or your performance. Likewise, some members of the audience may leave while you are speaking. Again, this may have nothing to do with you; they may have a previous engagement, or they may have discovered they are in the wrong room. Any number of factors might be at work. But you should be aware that this reflects badly on the audience, not on you, and other folks in the room are likely to be sympathetic to you if they observe it.

Another feature of public speaking that you may find disconcerting is that most members of the audience will not react to what you are saying. They will stare, passively, as if watching TV. Those in the back of the room may feel invisible, exactly as they do when watching television. This is likely to strike the speaker as odd. When we speak to people we generally expect them to respond, verbally or facially. Non-response is generally interpreted as a sign of extreme rudeness, if not autism. But public speaking is different. Although it is live, and although there should be some interaction between speaker and audience, the speaker often drones on for some period of time without cues from the other side. Indeed, the speaker may have no idea how the audience is receiving the talk until it is over. Part of becoming a practiced speaker is getting used to speaking into a vacuum. Alternatively, locate a few people in the audience who appear to be following you closely, nodding sympathetically, laughing at your jokes. Pretend that you are speaking to them and forget about everyone else in the room. (Of course, you still need to cast your eyes across the whole room so no one feels excluded.)

Foster exchange

The most important piece of advice for effective public speaking in academic or professional contexts is not to think too much about yourself. As you stand in front of the room, think instead about the reason you are there. That is, focus on the subject matter. Your function is not to display yourself for public acclaim (or disapprobation). This is not a beauty contest, nor is it a contest of brains. Your purpose is to communicate important information about a subject that you know something about (more, at any rate, than most members of the audience). This is your reason for being there.

Likewise, members of the audience are not antagonists – even if some of them behave in an antagonistic way. They are your allies in the search for the truth, which is always, and necessarily – a communal venture. So, when you start your talk enlist these allies. Tell them that you hope your presentation leads to a lively discussion and that you welcome their input.

Granted, you might prefer a quiescent audience that passively accepts whatever you serve up – better yet, an audience that erupts into applause the moment you have finished speaking.
However, this sort of response is not going to contribute to your thinking on the topic. Your quest for knowledge will not be enhanced.

Forward movement in a field usually begins with debate. And one debate leads to another. This is what academic and professional discourse is all about, or should be about. Thus, debate is what you should expect and what you should look forward to when you make a presentation in an academic or professional setting. It is a sign that your audience is taking you seriously. By contrast, if the room falls silent when you are done speaking, or if you receive polite responses but no searching questions, this is a sign that people are not engaged with your work. They don’t get it, or they don’t care.

If debate erupts after a talk you can assume that you have successfully engaged your audience. They are convinced that the subject is important and that what you have to say about it is important, even if they don’t agree with you or if they agree with the general argument but are critical of some aspect of your presentation. This is the sign of a successful talk. It is hoped that the discussion leads to insights, for you and for the members of the audience. In this fashion, our collective knowledge of a subject is advanced.

BE GRACIOUS

When handling debate, be gracious. Even if they hold views that you find repugnant, remember that they are human beings, and as such worthy of respect and understanding. They may have good reasons to hold the views they hold. In any case, to respect a person does not mean that you respect everything that s/he says or does. It is possible – and, arguably, essential – to accord respect to persons who have violated fundamental human rights. It follows that you should also respect those with whom you have ideological, substantive, or methodological disagreements.

It is tricky, to be sure, to convey a sense of respect for someone while arguing vehemently with them. You may need to monitor your words and gestures carefully, for it does not come naturally. Generally, when we argue about something that argument is accompanied by a degree of hostility towards the person we are arguing with. We strike an antagonistic pose. We roll our eyes, look away, or find ways to express contempt for the person we are confronting. These ingrained responses must be controlled, for they tend to personalize the conflict and lead away from any possibility of rational debate or reconciliation. Each side digs in and refuses to cede ground.

Remind yourself that the purpose of academic and professional debate is to find the right (or best) answer, not to win an argument or to prove one’s intellectual superiority. The true intellectual is one who seeks truth, not one who seeks victory. Learn how to argue with someone (or with a whole room) without conveying contempt or dismay.

You may find it helpful to preface your response to a critical comment with a disarming comment like, “That’s an interesting point,….” This turn-of-phrase should not be over-used or applied in situations where it is patently inapplicable. However, it is always good to begin by acknowledging the interlocutor’s perspective. Everyone wants to be acknowledged. And as the speaker you have special powers in this domain.

Another point to bear in mind is that disagreements are less invidious when they are openly acknowledged. So whenever you think that you might be saying something that others would take issue with, call attention to this fact. In doing so, you confer respect upon those who hold this opposing view. It is a rhetorical bowing of the head or doffing of the cap.

The most noxious statements are unacknowledged assumptions. Indeed, people generally take offense not when they disagree with your argument but when they disagree with your premises.
This recalls the witticism: *When did you stop beating your wife?* What is annoying about the question is the assumption contained in the predicate. Anyone addressed in this manner is apt to be driven into frenzy.

Likewise, do not assume that members of the audience share your ideology or your worldview. It is common in social science settings to find a predominance of liberals and leftists, so much so that speakers may easily assume that everyone in the room shares this political perspective. Imagine how uncomfortable someone with more conservative views might feel if the basis of a comment – perhaps intended to be humorous – is the wrongness of their politics.

**AUDIENCE PARTICIPATION**

Thus far, our comments have been addressed to situations in which you are called upon to make a formal presentation (a talk or lecture). Most of this advice is also applicable to situations in which you participate as a member of an audience, e.g., in a classroom setting where someone else (a teacher or fellow-student) is holding the floor. Be respectful of your audience; be mindful of how you present yourself; consider the audience and the context carefully; use various techniques for combatting nervousness; and so forth. However, there are some distinctive features to audience participation that deserve special consideration.

Participation is a formal component of some courses. Even where it is not accorded a portion of your grade, your participation is likely to affect the teacher’s opinion of you as a person and a scholar. As such, it may influence your grade in subtle ways and it will certainly influence any letter of recommendation that you are able to extract from the instructor. (Nota bene: this letter may be much more important for your future prospects than the grade you receive in the course.) Instructors, like the rest of us, tend to remember personal interventions more vividly than written work. Or, perhaps, they all blend together.

In any case, participation is a vital part of your education. If you participate regularly in class discussion you are more likely to enjoy the course, to remember its content, and to think creatively about the subject.

Talking extemporaneously on a subject works the brain, just as making a structured presentation does, but in somewhat different ways. Talking “from the floor” allows you to think in a relatively unstructured fashion about a topic, without having devoted countless hours of preparation. In formulating your words you may find yourself thinking about things differently. You may come to new insights. This is what the classroom experience is all about, or should be all about. (Regrettably, some classes consist of canned lectures delivered to passive audiences, who stare at their notes or computer screens, struggling to stay awake.)

You may think of questions and comments as an unwanted intrusion on a well-crafted presentation. You may wonder about your capacity to engage a complex subject, which the speaker has been studying for many years and about which others in the room may have superior knowledge. You may fear embarrassment.

This is entirely the wrong attitude. Bear in mind that an academic talk is not a finished work of art meant to be observed from a distance like a dance performance, a symphony, or a play. It is a participatory art form where little distinguishes the speaker from the audience, except that the former happens to be in front of the podium and the latter are seated around a table or in some arrangement in the hall. Where eminent people sit in the audience, their response to the presentation is often more eagerly anticipated than the speaker’s own presentation. Regardless of who is in attendance, the hall generally comes alive in the Q&A (question-and-answer) period. This is what
academics look forward to for it is the dynamic part of the talk, where an argument is put to test and
where new ideas are vetted. By contrast, the talk itself may be a formality, especially if the speaker’s
argument is wellknown. So, rather than thinking of participation as auxiliary to the talk you might
think of the talk as auxiliary to the discussion. A formal presentation facilitates productive debate,
focusing the discussion — usually on matters that are in dispute, leaving other issues (those that
everyone more or less agrees upon) in the background, where they belong. A talk without Q&A is
like a symphony without a crescendo.

In an educational environment, where the speaker is an expert and the audience are novices,
the situation is slightly different. There is more separation between speaker and audience, for one
thing. However, it is still the case that the format of live presentation is served only if there is some
interaction. Indeed, without such interaction between speaker and audience (or among audience
members) there is nothing to distinguish a live lecture from one that is viewed on video or streamed
on-line. There must be a purpose for bringing speaker and audience together in one place. That
purpose is for the two parties to interact.

Likewise, intervening regularly in class discussion will serve your educational goals. You are
likely to feel more involved, more motivated, and more engaged in the class. (At the very least, your
instructor will have that impression, and this will serve you well at the end of the semester.)

Note also that instructors generally appreciate questions and comments, even critical ones.
There is nothing so discouraging as talking to a silent — and apparently indifferent — room. If you
show the lecturer that you are following what is going on, and thinking independently about the
material, s/he is likely to be grateful. And you are likely to be rewarded — and certainly not punished
— for having spoken your mind.

Naturally, it is more difficult to participate in a cavernous hall, where the number of
audience members precludes a lot of talk from the floor. Some courses are structured so as to
facilitate a good deal of student participation; others are not. With that caveat, we would encourage
you to participate whenever it is possible and appropriate to do so.

In planning your interventions, here are some guidelines to follow:

• If there is an assigned reading for that day, make sure that you have read it.
• Listen carefully to the lecture and discussion prior to intervening.
• Avoid repeating points that have been raised previously (unless you have something new
to say about them).
• Make sure that your question or comment is germane. It should follow the previous
point. If it refers back to something covered earlier, point this out. And if others are
wishing to engage the current topic, let them have their say before changing the subject.
(Sometimes it is helpful to employ the one finger/two finger rule. Raising one finger indicates
that you have a new point to make. Raising two fingers indicates that your comment
follows the previous comment. Two fingers trump one finger, facilitating the logical flow
of discussion.)\textsuperscript{179}
• Do not think of participation simply as a matter of “saying something.” There are
insightful comments and questions, and there is also drivel. Consider what you want to
say before raising your hand, especially in a large hall. You may even jot down a few
notes, which you can refer to as needed.
• Be judicious in the use of anecdotes drawn from your personal life. We all have stories,
but these stories are not always relevant to the subject matter. In an informal setting

\textsuperscript{179} There are additional “finger rules”; but, to our knowledge, they are infrequently employed. See
http://www.utn.uu.se/best/aboutbest/fingerrules.html
(e.g., a coffee-house) it is appropriate to indicate your personal connection to what a speaker has to say. But it is not good form in an academic setting – unless, that is, the anecdote moves the discussion forward. There must be a point to any story – whether drawn on your experience or someone else’s.

- Don’t dominate the discussion. If you have spoken already and others have not yet had a chance, wait for a few moments before raising your hand or cede the floor if you see that others wish to be involved.
- Don’t be afraid to make a mistake. Making mistakes is part of the learning experience. If you do not allow yourself to make mistakes you are preventing yourself from learning. We make mistakes all the time. There is nothing wrong with a wrong answer. Oftentimes, the most productive sort of answer or question is one that reveals what is not clear in a presentation, and is thus on everyone’s mind.

**CONCLUSIONS**

Some people enjoy speaking to a crowd. Others detest it. We suspect that the difference between the two is that some are more comfortable in the spotlight than others, and this in turn is largely a product of familiarity. After all, most of us like attention. And this is an opportunity to get a lot of attention, and perhaps some accolades as well.

There is a common view that good speakers are born rather than raised. We don’t think this is any truer than the idea that good writers or good mathematicians are genetically endowed. In any case, whatever your native capabilities, there is probably a lot that you can do to improve your speaking ability. Indeed, the fact that public speaking is neither taught nor systematically practiced in most academic settings suggests that, for most of us, there may be great room for improvement. We can become much better speakers if we put our minds (and voices) to the task.

To that end, this chapter has introduced a number of tips that may prove useful to you as you work to improve your presentation of self, and your presentation of ideas. Here, as elsewhere, practice is important. Unlike some other skills, this one cannot be mastered on your own.
INQUIRIES

1. Reflect upon the last time that you gave a public talk – in a class or in some other (perhaps nonacademic) setting. What were your feelings about standing up in front of a group? Did you feel comfortable or uncomfortable? (Why?) What would have enhanced your level of confidence? Thinking more generally, what sort of public speaking situations seem manageable to you, and which seem most challenging? (Why?)

2. Think about people you know who address academic or political issues by speaking publicly about them, and do so successfully (i.e., they are known for being effective spokespersons). They might be professors, public intellectuals, doctors, lawyers, business-people, journalists, politicians, or entertainers. The venue in which they address the public might be small (20 people) or large (a large auditorium); it may be filmed for wider distribution or not. What is it that makes these people, and their public presentations, effective? Why have they (and not others) been chosen as spokespersons for an issue? What do you find personally appealing about them and about their presentations?

3. Obtain a video of an academic talk on a subject in your field of interest. (This can usually be accomplished by searching YouTube or the WWW.) After watching the talk, write a brief review of its strengths and weaknesses, following the guidance in this chapter. What did the speaker do well? What would you advise him/her to do differently?

4. The best way to improve your public speaking is to speak publicly. This can be integrated into class by having students present on various topics, e.g., the readings assigned for a specific class meeting, or their own project. Other students should serve as audience members and critics, offering suggestions for improvement after the presentation is complete. Other opportunities to speak publicly are provided by debate societies, Toastmasters, seminar classes, political meetings, and meetings of organizations. Sign up, show up, and speak.

RESOURCES

16. ETHICS

Ethics refers to formal and informal codes of behavior, i.e., the rightness or wrongness of an action as judged by a community. In the present context, we are of course concerned with ethical norms that apply to the activity of social science.

Ethical concerns are often raised in conjunction with experiments, where some subjects are assigned to a treatment condition and others to a control condition. If the treatment condition is potentially dangerous – or potentially beneficial – there are obvious ethical concerns. Why should one group of subjects be singled out in a wholly arbitrary fashion? What could justify such an intervention?

While they may be less apparent, ethical concerns also impinge upon other aspects of social science, e.g., to the collection of data from human subjects by observational methods, data analysis, and the dissemination of data and results. Arguably, everything that one might do, or refrain from doing, has ethical implications. Consequently, we take a broad view of the subject.

The chapter begins by discussing the juridical conception of ethics – centered on statutes, legal precedents, and institutional rules. Next, we discuss the central importance of protecting human subjects. Finally, we explore the activity of social science as an ethical sphere of action. What is right (moral) conduct for a social scientist?

JURIDICAL ETHICS

The juridical conception of ethics refers to rules one is obliged to follow as a member of a university, a profession, or a country. Universities in the United States are obliged, as a condition of receiving federal funding, to establish a set of procedures to govern research with human subjects. This involves the formation of an Institutional Review Board (IRB) to apply the protocol and to approve specific research proposals submitted by students or university employees. (IRB’s may also be referred to as ethics review boards.) All research on human subjects is subject to approval by the IRB. Even research that does not involve human subjects – such as historical research – may require IRB exemption. Thus, IRBs have become an integral part of scientific research at all American universities – at least, all universities that receive federal funding.

Those interested in the general goals of the IRB process are advised to take a look at the Belmont Report issued by the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research in 1979. This document informs all IRB processes.

Similar protocols have been developed in universities outside the United States and within businesses and research organizations, though there is considerable variety in procedures and methods of enforcement. Likewise, since every university in the United States sets its own procedures, and empanels its own IRB, there is considerable variety even across American universities.

Researchers are well-advised to familiarize themselves with IRB requirements wherever they are situated. The penalty for not abiding by these regulations may be severe, as it constitutes a very explicit form of professional misconduct and may be grounds for dismissal. In addition, professional associations generally publish a code of ethics (for a short list, see Resources). Membership in an
association implies an acceptance of this code, and violation of the code may, in principle, result in expulsion from the association.

**PROTECTION OF HUMAN SUBJECTS**

The vital importance of protecting human subjects became apparent to the world in the wake of the Tuskegee syphilis experiment in which black victims of the disease in rural Alabama were studied by the US Public Health Service for several decades without treatment and without even being informed that they had contracted the disease.

At the present time, protection of human subjects is vested primarily in IRBs, as discussed above. The principles underlying the IRB process are laid out in the *Belmont Report*, which reviews two main topics: informed consent and assessment of risks and benefits. (A third topic, the selection of subjects, is somewhat less important and is largely subsumed under the former. Quotations in the following paragraphs are from the Belmont Report.)

Informed consent refers to the principle that “subjects, to the degree that they are capable, be given the opportunity to choose what shall or shall not happen to them.” Consent procedures may involve signing a form or otherwise giving explicit consent to the researcher.

Evidently, subjects cannot make informed choices unless they are informed of the nature of the activity they are participating in. This involves, at the very least, information about who is sponsoring the study, what its goals are, how long the process will take, what will be required of them, and any costs or benefits that they might realize as a result of participation. In addition, they should be given an opportunity to ask questions and to withdraw their participation at any point.

The quandary of informed consent is that, quite often, a study may be compromised if the participants in that study are fully informed of its research goals. For example, in any experiment subjects’ choices are likely to be affected if they are told what the experiment is about. Imagine conducting a study of cheating in which subjects are given the opportunity to cheat on an exam. If subjects are informed of the purpose of the experiment no one is likely to cheat (and if they do, one may wonder about how to interpret this behavior).

The issue of informed consent is thus a delicate one. Sometimes, it is deemed sufficient if the participant is informed of the general topic undertaken by the researcher, described in a diffuse manner that is unlikely to affect how the informant responds to stimuli. Regardless of prior knowledge, participants should be fully informed (de-briefed) about the specific purposes of the experiment when their participation is completed.

Special consideration must be taken for subjects who are minors, mentally impaired, unfamiliar with the language, or for some other reason unable to exercise informed consent. Likewise, if subjects feel special obligations or pressures to serve as participants in a study, perhaps because they are being enlisted by someone in a position of authority (e.g., their professor) or perhaps because their peers are participating, the goal of informed consent may be thwarted. Informed consent “requires conditions free of coercion and undue influence,” whatever those may be.

A somewhat lower bar for informed consent may be granted for research subjects who occupy positions of prominence and power (e.g., political leaders). The assumption is that by virtue of the public nature of their profession and the generally recognized need to subject such individuals to special scrutiny, prominent individuals have less claim to privacy than others (at least as regards features of their life that are relevant to their public duties).
The assessment of risks and benefits is perhaps the most complex aspect of human subject research. Risk may involve possible “psychological harm, physical harm, legal harm, social harm [or] economic harm.” In a large study, with thousands of participants and (let us imagine) a long questionnaire, it is quite possible that some untoward consequence will result to a few of the participants. The most likely sort of harm is psychological. Insofar as research involves topics that are sensitive — at least to some subjects — those subjects may experience distress as a result of their participation. If they are already in a condition of anxiety, fear, or depression, that distress may be especially severe and may have consequences for them or those they are intimate with.

It is fatuous to suppose that social science research — or, for that matter, medical research — could be conducted without doing any harm to anyone. The avoidance of harm must be considered a relative matter, and must be judged against the possible benefits to be gained from the research.

The Belmont report specifies:

- Brutal or inhumane treatment of human subjects is never morally justified.
- Risks should be reduced to those necessary to achieve the research objective. It should be determined whether it is in fact necessary to use human subjects at all. Risk can perhaps never be entirely eliminated, but it can often be reduced by careful attention to alternative procedures.
- When research involves significant risk of serious impairment, review committees should be extraordinarily insistent on the justification of the risk (looking usually to the likelihood of benefit to the subject -- or, in some rare cases, to the manifest voluntariness of the participation).
- When vulnerable populations are involved in research, the appropriateness of involving them should itself be demonstrated. A number of variables go into such judgments, including the nature and degree of risk, the condition of the particular population involved, and the nature and level of the anticipated benefits.
- Relevant risks and benefits must be thoroughly arrayed in documents and procedures used in the informed consent process.

One important mechanism for limiting harm to subjects is to guarantee anonymity or confidentiality. **Anonymity** means that the researcher has no knowledge of the identity of subjects under study, who remain anonymous. Most surveys do not require the name of the respondent, so a limited form of anonymity is usually possible. However, there may be identifying information contained in the survey, so researchers will need to be careful about releasing this information if they have guaranteed anonymity to their subjects.

**Confidentiality** means that the researcher is aware of the identity of subjects but promises not to reveal it to others. This may be accomplished by keeping names in a separate database (perhaps even on a separate computer) and referring to subjects only by an assigned number or pseudonym. Again, the problem of identifying information must be handled carefully.

The Belmont report has been quoted extensively because of its broad impact on the conduct of IRBs and on the consideration of ethical norms pertaining to the protection of human subjects. However, the diffuse nature of the mandate to protect human subjects, and the largely autonomous work of IRBs within institutions throughout the world, leads to considerable unevenness of application. Inevitably, some valuable studies with (arguably) few negative implications for human subjects never gain approval, and some not-so-valuable studies with negative implications for human subjects gain approval. The process is not perfect.

Moreover, everyone covered by an IRB protocol is obliged to spend a lot of time satisfying these procedures, a (relatively) unproductive use of academic resources. At the same time, those living outside IRB regimes — such as researchers working for international NGOs such as the World Bank, private sector businesses, or journalists — have very little oversight.

It is troubling that studies with an academic objective receive much closer ethical scrutiny than activities undertaken for commercial or journalistic purposes. For example, if a company
wishes to conduct a survey, focus group, experiment, or participant-observation study in order to market-test a new product the researchers engaged by that company are unlikely to undergo a rigorous IRB process. If the same study were taken by a student or professor within a university they would be required to be IRB compliant. This divergence of standards is difficult to justify.

**BROADER ETHICAL ISSUES**

Ethics pertaining to social science begins with the protection of human subjects. That is why it is commonly associated with the process of data collection, and especially with obtrusive approaches to data collection such as experiments with a manipulated treatment. However, ethical considerations do not end there.

In conducting social science research one is (hopefully!) affecting others in society. Some research is directed explicitly at policymakers or at the general public. Even if it is directed at other academics the general expectation is that it will, eventually, have an echo in the broader society of which academics is a part. If this weren’t the case – if we engaged in social research purely for our own edification or entertainment – the enterprise wouldn’t be very satisfying. For most of us, there is an expectation that our ideas will change the world, even if only modestly.

It follows that a complete consideration of social science ethics involves a consideration of social science’s role in society and how that role might be fashioned in the best possible way.

Arguably, the contribution of social science to society lies in its capacity to elucidate useful truths, i.e., insights into topics that citizens and policymakers care about. To the extent that members of the various social science tribes are successfully performing this task we might say that they are acting ethically. To the extent that they are not, we might say that they are violating an implicit ethical obligation to be of service to society.

The question then becomes, which features of social science – which norms and rules – are conducive to elucidating useful truths? The answer, we would suggest, are all those features generally regarded as part of social science methodology, as laid out in this book.

Of course, some norms are not universally agreed upon. The point remains that insofar as a norm promotes useful truths it is rightly accompanied by an ethical imperative. Consider a norm that most practitioners would probably agree upon such as *replicability*. Replicability is generally regarded as promoting the advancement of science. Likewise, violating this ideal by failing to cite sources, to make available one’s data, or to leave a trail that other scholars can follow, is regarded as destructive of scientific advance. If we cannot replicate studies we cannot cumulate knowledge on the subject those studies address. Consequently, a violation of this methodological principle is generally regarded as a gross violation of academic norms. This is just one example of how methodological rules carry normative freight. They are not just bees in a methodologist’s bonnet.

It may seem odd to associate methodology with ethics. After all, methodologists occupy an academic niche far removed from the field of moral philosophy. But methodologists are the standard-setters and standard-keepers of “good behavior” in the social sciences. They are the makers and enforcers of professional norms. We are not referring to norms like the prohibition against sexual encounters with students or against plagiarism; these we will classify as general norms insofar as they apply to everyone, not just academics. By professional norms we mean norms that are specific to social science.

So, if good social science is our contribution to society, then the methodological norms that contribute to good social science have ethical sanction. In this respect, it may be argued, the ethics
of social science extend far beyond the obligation to respect the rights of human subjects in our
data-gathering efforts. They extend, in a diffuse fashion, to all the other subjects raised in this book.

THE PURVIEW OF SOCIAL SCIENCE
One final point relates to the purview of social science, a theme raised initially in Chapter 1. In
proposing that social science should make a positive contribution to society we are not supposing
that we are the sole contributors to human advancement – or, for that matter, the sole possessors
of truth. There are many ways to make a contribution and it would be vain to arrogate a greater role in
human affairs than social scientists actually play, or have a right to play. Policing this line – between
the appropriate and inappropriate applications of social science – is vital if social science is not to
overstep its role.

In particular, a distinction must be maintained between partisan politics and social science.
Of course, one hopes that social science is enlisted in partisan politics – not as a weapon to
bludgeon opponents but as a means of bringing systematic evidence and reflection to bear on
problems that matter. Likewise, one hopes that when social scientists engage in partisan politics (as
they have every right to do) their activity is properly labeled – as partisan rather than scientific.

Noam Chomsky, a prominent linguist, served as an influential leader of the antiwar
movement during the Vietnam War. However, he did not claim the mantle of social science in order
to argue that his cause was just.

Naturally, it is often difficult to distinguish work that is “polemical” from work that is
impartial, objective, dispassionate. After all, our motivation to study social science, as well as our
choice of subjects to study, may be animated by a desire to influence political and social affairs.
However, these background motivations must be separated, as much as is possible, from the
conduct of inquiry.

Let us put a fine point on this. A researcher is likely to choose a topic, and identify a
hypothesis, because s/he has a strong “prior”: s/he believes this is the way the world works. This is
well and good. Intuitions about the world must come from somewhere. However, s/he must put
this hypothesis to an honest test, and must be willing to accept its refutation if that is what the
evidence suggests. One’s obligation to tell the truth must outweigh all other considerations, even if
the results are unpalatable.

In this context, one might consider the rise and fall of academic fashions. A century ago, it
was common to regard human nature and geography as wellsprings of social behavior. By mid-
century, these views were regarded as passé, if not racist. Now, we find a return to these venerable
explanatory foundations (purged of their racist content, one hopes).

Scholars must feel free to think and theorize freely, so long as they are also able to put their
ideas to the test. This, it seems to us, is the principal ethical responsibility of social scientists.

CONCLUSIONS

In this chapter we have endeavored to cover a very large and complex subject in a synoptic fashion.
We did so by dividing the subject into three areas – (a) juridical ethics, (b) the protection of human
subjects, and (c) broader ethical issues. While (a) and (b) are often regarded as the sum-total of
professional ethics we believe that the latter deserves equal consideration when social scientists are
constructing and presenting their work. Ethical considerations do not end with a stamp of approval
from an IRB. Nor should the approval of an institutional review board constitute the final and
authoritative judgment on the ethical nature of a piece of research – though it may, in practice, settle the matter (by allowing or preventing a study) in a particular context.
KEY TERMS

- Ethics
- Institutional Review Board
- Belmont Report
- Informed Consent
- Anonymity
- Confidentiality

INQUIRIES

1. What does “informed consent” usually mean in an IRB context? What should it mean? (How strictly should it be interpreted?)

2. When (if ever) should deception be allowed in a study with human subjects?

3. Consider the following study (or some other study in your field of interest). Does it violate the rights of human subjects? What are the relevant issues and why might people view them differently?

4. How does one (or should one) assess “risks and benefits” of a study? Is this standard meaningful? Or is it so open-ended that it can mean whatever one wishes it mean?

5. What are the IRB requirements at your institution? Follow whatever training or instruction is provided by your institution. If you have a project that involves human subjects, submit it to the board for approval (making sure to allow sufficient time for them to process it). Now, reflect on your experience. Do you think IRB requirements at your institution are (a) reasonable, (b) too onerous, or (c) insufficiently demanding? Are the questions they ask, the training they provide, relevant to your research? Are there important ethical issues that are left out, or are given short shrift?

6. Do you think that some types of research are impossible to carry out without violating IRB regulations? Is this a problem?

7. What do you think is the best procedure for ensuring that ethical procedures are followed by social scientists? Is the current system, as it operates in your country, effective?

8. Are the ethics of social science reducible to the maxim “Do no harm”?

9. What is the distinction between a “narrow” and “broad” conception of ethics? (Some might refer to the latter as morality.) Why does this distinction matter?
10. Is it ethical to have strong views (“priors”) about the issues one is researching?

11. Is it ethical for a social scientist to be strongly identified with a particular issue, ideology, or political party and to engage in partisan politics? How should one balance one’s obligations as a social scientist and as a citizen?

RESOURCES

PUBLISHED WORK


WEB SITES

• Ethics Core digital library http://nationalethicsresourcecenter.net/index.php/home


PROFESSIONAL ASSOCIATIONS


• Applied Research Ethics National Association (ARENA).
References


Ross, Michael L. 2014. “What have we learned about the resource curse?” Unpublished paper, Department of Political Science, UCLA.


Index