CHAPTER 1

The *Science* in Social Science

1.1 Introduction

This book is about research in the social sciences. Our goal is practical: designing research that will produce valid inferences about social and political life. We focus on political science, but our argument applies to other disciplines such as sociology, anthropology, history, economics, and psychology and to nondisciplinary areas of study such as legal evidence, education research, and clinical reasoning.

This is neither a work in the philosophy of the social sciences nor a guide to specific research tasks such as the design of surveys, conduct of field work, or analysis of statistical data. Rather, this is a book about research design: how to pose questions and fashion scholarly research to make valid descriptive and causal inferences. As such, it occupies a middle ground between abstract philosophical debates and the hands-on techniques of the researcher and focuses on the essential logic underlying all social scientific research.

1.1.1 Two Styles of Research, One Logic of Inference

Our main goal is to connect the traditions of what are conventionally denoted "quantitative" and "qualitative" research by applying a unified logic of inference to both. The two traditions appear quite different; indeed they sometimes seem to be at war. Our view is that these differences are mainly ones of style and specific technique. The same underlying logic provides the framework for each research approach. This logic tends to be explicated and formalized clearly in discussions of quantitative research methods. But the same logic of inference underlies the best qualitative research, and all qualitative and quantitative researchers would benefit by more explicit attention to this logic in the course of designing research.

The *styles* of quantitative and qualitative research are very different. Quantitative research uses numbers and statistical methods. It tends to be based on numerical measurements of specific aspects of phenomena; it abstracts from particular instances to seek general description or to test causal hypotheses; it seeks measurements and analyses that are easily replicable by other researchers.

Qualitative research, in contrast, covers a wide range of approaches, but by definition, none of these approaches relies on numerical measurements. Such work has tended to focus on one or a small number of cases, to use intensive interviews or depth analysis of historical materials, to be discursive in method, and to be concerned with a rounded or comprehensive account of some event or unit. Even though they have a small number of cases, qualitative researchers generally unearth enormous amounts of information from their studies. Sometimes this kind of work in the social sciences is linked with area or case studies where the focus is on a particular event, decision, institution, location, issue, or piece of legislation. As is also the case with quantitative research, the instance is often important in its own right: a major change in a nation, an election, a major decision, or a world crisis. Why did the East German regime collapse so suddenly in 1989? More generally, why did almost all the communist regimes of Eastern Europe collapse in 1989? Sometimes, but certainly not always, the event may be chosen as an exemplar of a particular type of event, such as a political revolution or the decision of a particular community to reject a waste disposal site. Sometimes this kind of work is linked to area studies where the focus is on the history and culture of a particular part of the world. The particular place or event is analyzed closely and in full detail.

For several decades, political scientists have debated the merits of case studies versus statistical studies, area studies versus comparative studies, and "scientific" studies of politics using quantitative methods versus "historical" investigations relying on rich textual and contextual understanding. Some quantitative researchers believe that systematic statistical analysis is the only road to truth in the social sciences. Advocates of qualitative research vehemently disagree. This difference of opinion leads to lively debate; but unfortunately, it also bifurcates the social sciences into a quantitative-systematic-generalizing branch and a qualitative-humanistic-discursive branch. As the former becomes more and more sophisticated in the analysis of statistical data (and their work becomes less comprehensible to those who have not studied the techniques), the latter becomes more and more convinced of the irrelevance of such analyses to the seemingly nonreplicable and nongeneralizable events in which its practitioners are interested.

A major purpose of this book is to show that the differences between the quantitative and qualitative traditions are only stylistic and are methodologically and substantively unimportant. All good research can be understood—indeed, is best understood—to derive from the same underlying logic of inference. Both quantitative and qualitative research can be systematic and scientific. Historical research can be analytical, seeking to evaluate alternative explanations through a process of valid causal inference. History, or historical sociology, is not incompatible with social science (Skocpol 1984: 374–86).

Breaking down these barriers requires that we begin by questioning the very concept of "qualitative" research. We have used the term in our title to signal our subject matter, not to imply that "qualitative" research is fundamentally different from "quantitative" research, except in style.

Most research does not fit clearly into one category or the other. The best often combines features of each. In the same research project, some data may be collected that is amenable to statistical analysis, while other equally significant information is not. Patterns and trends in social, political, or economic behavior are more readily subjected to quantitative analysis than is the flow of ideas among people or the difference made by exceptional individual leadership. If we are to understand the rapidly changing social world, we will need to include information that cannot be easily quantified as well as that which can. Furthermore, all social science requires comparison, which entails judgments of which phenomena are "more" or "less" alike in degree (i.e., quantitative differences) or in kind (i.e., qualitative differences).

Two excellent recent studies exemplify this point. In *Coercive Cooperation* (1992), Lisa L. Martin sought to explain the degree of international cooperation on economic sanctions by quantitatively analyzing ninety-nine cases of attempted economic sanctions from the post–World War II era. Although this quantitative analysis yielded much valuable information, certain causal inferences suggested by the data were ambiguous; hence, Martin carried out six detailed case studies of sanctions episodes in an attempt to gather more evidence relevant to her causal inference. For *Making Democracy Work* (1993), Robert D. Putnam and his colleagues interviewed 112 Italian regional councillors in 1970, 194 in 1976, and 234 in 1981–1982, and 115 community leaders in 1976 and 118 in 1981–1982. They also sent a mail questionnaire to over 500 community leaders throughout the country in 1983. Four nationwide mass surveys were undertaken especially for this study. Nevertheless, between 1976 and 1989 Putnam and his colleagues conducted detailed case studies of the politics of six regions. Seeking to satisfy the "interocular traumatic test," the investigators "gained an intimate knowledge of the internal political maneuvering and personalities that have animated regional politics over the last two decades" (Putnam 1993:190).

The lessons of these efforts should be clear: neither quantitative nor qualitative research is superior to the other, regardless of the research problem being addressed. Since many subjects of interest to social scientists cannot be meaningfully formulated in ways that permit statistical testing of hypotheses with quantitative data, we do not wish to encourage the exclusive use of quantitative techniques. We are not trying to get all social scientists out of the library and into the computer center, or to replace idiosyncratic conversations with structured interviews. Rather, we argue that nonstatistical research will produce more reliable results if researchers pay attention to the rules of scientific inference—rules that are sometimes more clearly stated in the style of quantitative research. Precisely defined statistical methods that undergird quantitative research represent abstract formal models applicable to all kinds of research, even that for which variables cannot be measured quantitatively. The very abstract, and even unrealistic, nature of statistical models is what makes the rules of inference shine through so clearly.

The rules of inference that we discuss are not relevant to all issues that are of significance to social scientists. Many of the most important questions concerning political life—about such concepts as agency, obligation, legitimacy, citizenship, sovereignty, and the proper relationship between national societies and international politics—are philosophical rather than empirical. But the rules are relevant to all research where the goal is to learn facts about the real world. Indeed, the distinctive characteristic that sets social science apart from casual observation is that social science seeks to arrive at valid inferences by the systematic use of well-established procedures of inquiry. Our focus here on empirical research means that we sidestep many issues in the philosophy of social science as well as controversies about the role of postmodernism, the nature and existence of truth, relativism, and related subjects. We assume that it is possible to have some knowledge of the external world but that such knowledge is always uncertain.

Furthermore, nothing in our set of rules implies that we must run the perfect experiment (if such a thing existed) or collect all relevant data before we can make valid social scientific inferences. An important topic is worth studying even if very little information is available. The result of applying any research design in this situation will be relatively uncertain conclusions, but so long as we honestly report our uncertainty, this kind of study can be very useful. Limited information is often a necessary feature of social inquiry. Because the social world changes rapidly, analyses that help us understand those changes require that we describe them and seek to understand them contemporaneously, even when uncertainty about our conclusions is high. The urgency of a problem may be so great that data gathered by the most useful scientific methods might be obsolete before it can be accumulated. If a distraught person is running at us swinging an ax, administering a five-page questionnaire on psychopathy may not be the best strategy. Joseph Schumpeter once cited Albert Einstein, who said "as far as our propositions are certain, they do not say anything about reality, and as far as they do say anything about reality, they are not certain" (Schumpeter [1936] 1991:298–99). Yet even though certainty is unattainable, we can improve the reliability, validity, certainty, and honesty of our conclusions by paying attention to the rules of scientific inference. The social science we espouse seeks to make descriptive and causal inferences about the world. Those who do not share the assumptions of partial and imperfect knowability and the aspiration for descriptive and causal understanding will have to look elsewhere for inspiration or for paradigmatic battles in which to engage.

In sum, we do not provide recipes for scientific empirical research. We offer a number of precepts and rules, but these are meant to discipline thought, not stifle it. In both quantitative and qualitative research, we engage in the imperfect application of theoretical standards of inference to inherently imperfect research designs and empirical data. Any meaningful rules admit of exceptions, but we can ask that exceptions be justified explicitly, that their implications for the reliability of research be assessed, and that the uncertainty of conclusions be reported. We seek not dogma, but disciplined thought.

1.1.2 Defining Scientific Research in the Social Sciences

Our definition of "scientific research" is an ideal to which any actual quantitative or qualitative research, even the most careful, is only an approximation. Yet, we need a definition of good research, for which we use the word "scientific" as our descriptor.[1] This word comes with many connotations that are unwarranted or inappropriate or downright incendiary for some qualitative researchers. Hence, we provide an explicit definition here. As should be clear, we do not regard quantitative research to be any more scientific than qualitative research. Good research, that is, scientific research, can be quantitative or qualitative in style. In design, however, scientific research has the following four characteristics:

1. **The goal is inference.** Scientific research is designed to make descriptive or explanatory *inferences* on the basis of empirical information about the world. Careful descriptions of specific phenomena are often indispens-

[1] We reject the concept, or at least the word, "quasi-experiment." Either a research design involves investigator control over the observations and values of the key causal variables (in which case it is an experiment) or it does not (in which case it is nonexperimental research). Both experimental and nonexperimental research have their advantages and drawbacks; one is not better in all research situations than the other.

able to scientific research, but the accumulation of facts alone is not sufficient. Facts can be collected (by qualitative or quantitative researchers) more or less systematically, and the former is obviously better than the latter, but our particular definition of science requires the additional step of attempting to infer beyond the immediate data to something broader that is not directly observed. That something may involve *descriptive inference*—using observations from the world to learn about other unobserved facts. Or that something may involve *causal inference*—learning about causal effects from the data observed. The domain of inference can be restricted in space and time—voting behavior in American elections since 1960, social movements in Eastern Europe since 1989—or it can be extensive—human behavior since the invention of agriculture. In either case, the key distinguishing mark of scientific research is the goal of making inferences that go beyond the particular observations collected.

2. **The procedures are public.** Scientific research uses explicit, codified, and *public* methods to generate and analyze data whose reliability can therefore be assessed. Much social research in the qualitative style follows fewer precise rules of research procedure or of inference. As Robert K. Merton ([1949] 1968:71–72) put it, "The sociological analysis of qualitative data often resides in a private world of penetrating but unfathomable insights and ineffable understandings. . . . [However,] science . . . is public, not private." Merton's statement is not true of all qualitative researchers (and it is unfortunately still true of some quantitative analysts), but many proceed as if they had no method—sometimes as if the use of explicit methods would diminish their creativity. Nevertheless they cannot help but use some method. Somehow they observe phenomena, ask questions, infer information about the world from these observations, and make inferences about cause and effect. If the method and logic of a researcher's observations and inferences are left implicit, the scholarly community has no way of judging the validity of what was done. We cannot evaluate the principles of selection that were used to record observations, the ways in which observations were processed, and the logic by which conclusions were drawn. We cannot learn from their methods or replicate their results. Such research is not a *public* act. Whether or not it makes good reading, it is not a contribution to social science.

   All methods—whether explicit or not—have limitations. The advantage of explicitness is that those limitations can be understood and, if possible, addressed. In addition, the methods can be taught and shared. This process allows research results to be compared across separate researchers and research projects studies to be replicated, and scholars to learn.

3. **The conclusions are uncertain.** By definition, inference is an imperfect process. Its goal is to use quantitative or qualitative data to learn about the world that produced them. Reaching perfectly certain conclusions

from uncertain data is obviously impossible. Indeed, uncertainty is a central aspect of all research and all knowledge about the world. Without a reasonable estimate of uncertainty, a description of the real world or an inference about a causal effect in the real world is uninterpretable. A researcher who fails to face the issue of uncertainty directly is either asserting that he or she knows everything perfectly or that he or she has no idea how certain or uncertain the results are. Either way, inferences without uncertainty estimates are not science as we define it.

4. **The content is the method.** Finally, scientific research adheres to a set of rules of inference on which its validity depends. Explicating the most important rules is a major task of this book.[2] The content of "science" is primarily the methods and rules, not the subject matter, since we can use these methods to study virtually anything. This point was recognized over a century ago when Karl Pearson (1892: 16) explained that "the field of science is unlimited; its material is endless; every group of natural phenomena, every phase of social life, every stage of past or present development is material for science. The unity of all science consists alone in its method, not in its material."

These four features of science have a further implication: science at its best is a *social enterprise*. Every researcher or team of researchers labors under limitations of knowledge and insight, and mistakes are unavoidable, yet such errors will likely be pointed out by others. Understanding the social character of science can be liberating since it means that our work need not to be beyond criticism to make an important contribution—whether to the description of a problem or its conceptualization, to theory or to the evaluation of theory. As long as our work explicitly addresses (or attempts to redirect) the concerns of the community of scholars and uses public methods to arrive at inferences that are consistent with rules of science and the information at our disposal, it is likely to make a contribution. And the contribution of even a minor article is greater than that of the "great work" that stays forever in a desk drawer or within the confines of a computer.

1.1.3 Science and Complexity

Social science constitutes an attempt to make sense of social situations that we perceive as more or less complex. We need to recognize, however, that what we perceive as complexity is not entirely inherent in phenomena: the world is not naturally divided into simple and com-

[2] Although we do cover the vast majority of the important rules of scientific inference, they are not complete. Indeed, most philosophers agree that a complete, exhaustive inductive logic is impossible, even in principle.

plex sets of events. On the contrary, the perceived complexity of a situation depends in part on how well we can simplify reality, and our capacity to simplify depends on whether we can specify outcomes and explanatory variables in a coherent way. Having more observations may assist us in this process but is usually insufficient. Thus *"complexity" is partly conditional on the state of our theory.*

Scientific methods can be as valuable for intrinsically complex events as for simpler ones. Complexity is likely to make our inferences less certain but should *not* make them any less scientific. Uncertainty and limited data should not cause us to abandon scientific research. On the contrary: the biggest payoff for using the rules of scientific inference occurs precisely when data are limited, observation tools are flawed, measurements are unclear, and relationships are uncertain. With clear relationships and unambiguous data, method may be less important, since even partially flawed rules of inference may produce answers that are roughly correct.

Consider some complex, and in some sense unique, events with enormous ramifications. The collapse of the Roman Empire, the French Revolution, the American Civil War, World War I, the Holocaust, and the reunification of Germany in 1990 are all examples of such events. These events seem to be the result of complex interactions of many forces whose conjuncture appears crucial to the event having taken place. That is, independently caused sequences of events and forces converged at a given place and time, their interaction appearing to bring about the events being observed (Hirschman 1970). Furthermore, it is often difficult to believe that these events were inevitable products of large-scale historical forces: some seem to have depended, in part, on idiosyncracies of personalities, institutions, or social movements. Indeed, from the perspective of our theories, chance often seems to have played a role: factors outside the scope of the theory provided crucial links in the sequences of events.

One way to understand such events is by seeking generalizations: conceptualizing each case as a member of a *class of events* about which meaningful generalizations can be made. This method often works well for ordinary wars or revolutions, but some wars and revolutions, being much more extreme than others, are "outliers" in the statistical distribution. Furthermore, notable early wars or revolutions may exert such a strong impact on subsequent events of the same class—we think again of the French Revolution—that caution is necessary in comparing them with their successors, which may be to some extent the product of imitation. Expanding the class of events can be useful, but it is not always appropriate.

Another way of dealing scientifically with rare, large-scale events is to engage in counterfactual analysis: "the mental construction of a course of events which is altered through modifications in one or more 'conditions'" (Weber [1905] 1949:173). The application of this idea in a systematic, scientific way is illustrated in a particularly extreme example of a rare event from geology and evolutionary biology, both historically oriented natural sciences. Stephen J. Gould has suggested that one way to distinguish systematic features of evolution from stochastic, chance events may be to imagine what the world would be like if all conditions up to a specific point were fixed and then the rest of history were rerun. He contends that if it were possible to "replay the tape of life," to let evolution occur again from the beginning, the world's organisms today would be a completely different (Gould 1989a).

A unique event on which students of evolution have recently focused is the sudden extinction of the dinosaurs 65 million years ago. Gould (1989a:318) says, "we must assume that consciousness would not have evolved on our planet if a cosmic catastrophe had not claimed the dinosaurs as victims." If this statement is true, the extinction of the dinosaurs was as important as any historical event for human beings; however, dinosaur extinction does not fall neatly into a class of events that could be studied in a systematic, comparative fashion through the application of general laws in a straightforward way.

Nevertheless, dinosaur extinction can be studied scientifically: alternative hypotheses can be developed and tested with respect to their observable implications. One hypothesis to account for dinosaur extinction, developed by Luis Alvarez and collaborators at Berkeley in the late 1970s (W. Alvarez and Asaro, 1990), posits a cosmic collision: a meteorite crashed into the earth at about 72,000 kilometers an hour, creating a blast greater than that from a full-scale nuclear war. If this hypothesis is correct, it would have the observable implication that iridium (an element common in meteorites but rare on earth) should be found in the particular layer of the earth's crust that corresponds to sediment laid down sixty-five million years ago; indeed, the discovery of iridium at predicted layers in the earth has been taken as partial confirming evidence for the theory. Although this is an unambiguously unique event, there are many other observable implications. For one example, it should be possible to find the metorite's crater somewhere on Earth (and several candidates have already been found).[3]

The issue of the cause(s) of dinosaur extinction remains unresolved, although the controversy has generated much valuable research. For

[3] However, an alternative hypothesis, that extinction was caused by volcanic eruptions, is also consistent with the presence of iridium, and seems more consistent than the meteorite hypothesis with the finding that all the species extinctions did not occur simultaneously.

our purposes, the point of this example is that scientific generalizations are useful in studying even highly unusual events that do not fall into a large class of events. The Alvarez hypothesis cannot be tested with reference to a set of common events, but it does have observable implications for other phenomena that can be evaluated. We should note, however, that a hypothesis is not considered a reasonably certain explanation until it has been evaluated empirically and passed a number of demanding tests. At a minimum, its implications must be consistent with our knowledge of the external world; at best, it should predict what Imre Lakatos (1970) refers to as "new facts," that is, those formerly unobserved.

The point is that even apparently unique events such as dinosaur extinction can be studied scientifically if we pay attention to improving theory, data, and our use of the data. Improving our theory through conceptual clarification and specification of variables can generate more observable implications and even test causal theories of unique events such as dinosaur extinction. Improving our data allows us to observe more of these observable implications, and improving our use of data permits more of these implications to be extracted from existing data. That a set of events to be studied is highly complex does not render careful research design irrelevant. Whether we study many phenomena or few—or even one—the study will be improved if we collect data on as many observable implications of our theory as possible.

1.2 Major Components of Research Design

Social science research at its best is a creative process of insight and discovery taking place within a well-established structure of scientific inquiry. The first-rate social scientist does not regard a research design as a blueprint for a mechanical process of data-gathering and evaluation. To the contrary, the scholar must have the flexibility of mind to overturn old ways of looking at the world, to ask new questions, to revise research designs appropriately, and then to collect more data of a different type than originally intended. However, if the researcher's findings are to be valid and accepted by scholars in this field, all these revisions and reconsiderations must take place according to explicit procedures consistent with the rules of inference. A dynamic process of inquiry occurs within a stable structure of rules.

Social scientists often begin research with a considered design, collect some data, and draw conclusions. But this process is rarely a smooth one and is not always best done in this order: conclusions rarely follow easily from a research design and data collected in accordance with it. Once an investigator has collected data as provided by a research design, he or she will often find an imperfect fit among the main research questions, the theory and the data at hand. At this stage, researchers often become discouraged. They mistakenly believe that other social scientists find close, immediate fits between data and research. This perception is due to the fact that investigators often take down the scaffolding after putting up their intellectual buildings, leaving little trace of the agony and uncertainty of construction. Thus the process of inquiry seems more mechanical and cut-and-dried than it actually is.

Some of our advice is directed toward researchers who are trying to make connections between theory and data. At times, they can design more appropriate data-collection procedures in order to evaluate a theory better; at other times, they can use the data they have and recast a theoretical question (or even pose an entirely different question that was not originally foreseen) to produce a more important research project. The research, if it adheres to rules of inference, will still be scientific and produce reliable inferences about the world.

Wherever possible, researchers should also improve their research designs before conducting any field research. However, data has a way of disciplining thought. It is extremely common to find that the best research design falls apart when the very first observations are collected—it is not that the theory is wrong but that the data are not suited to answering the questions originally posed. Understanding from the outset what can and what cannot be done at this later stage can help the researcher anticipate at least some of the problems when first designing the research.

For analytical purposes, we divide all research designs into four components: the *research question*, the *theory*, the *data*, and the *use of the data*. These components are not usually developed separately and scholars do not attend to them in any preordained order. In fact, for qualitative researchers who begin their field work before choosing a precise research question, data comes first, followed by the others. However, this particular breakdown, which we explain in sections 1.2.1–1.2.4, is particularly useful for understanding the nature of research designs. In order to clarify precisely what *could* be done if resources were redirected, our advice in the remainder of this section assumes that researchers have unlimited time and resources. Of course, in any actual research situation, one must always make compromises. We believe that understanding the advice in the four categories that follow will help researchers make these compromises in such a way as to improve their research designs most, even when in fact their research is subject to external constraints.

1.2.1 Improving Research Questions

Throughout this book, we consider what to do once we identify the object of research. Given a research question, what are the ways to conduct that research so that we can obtain valid explanations of social and political phenomena? Our discussion begins with a research question and then proceeds to the stages of designing and conducting the research. But where do research questions originate? How does a scholar choose the topic for analysis? There is no simple answer to this question. Like others, Karl Popper (1968:32) has argued that "there is no such thing as a logical method of having new ideas. . . . Discovery contains 'an irrational element,' or a 'creative intuition.'" The rules of choice at the earliest stages of the research process are less formalized than are the rules for other research activities. There are texts on designing laboratory experiments on social choice, statistical criteria on drawing a sample for a survey of attitudes on public policy, and manuals on conducting participant observation of a bureaucratic office. But there is no rule for choosing which research project to conduct, nor if we should decide to conduct field work, are there rules governing where we should conduct it.

We can propose ways to select a sample of communities in order to study the impact of alternative educational policies, or ways to conceptualize ethnic conflict in a manner conducive to the formulation and testing of hypotheses as to its incidence. But there are no rules that tell us whether to study educational policy or ethnic conflict. In terms of social science methods, there are better and worse ways to study the collapse of the East German government in 1989 just as there are better and worse ways to study the relationship between a candidate's position on taxes and the likelihood of electoral success. But there is no way to determine whether it is better to study the collapse of the East German regime or the role of taxes in U.S. electoral politics.

The specific topic that a social scientist studies may have a personal and idiosyncratic origin. It is no accident that research on particular groups is likely to be pioneered by people of that group: women have often led the way in the history of women, blacks in the history of blacks, immigrants in the history of immigration. Topics may also be influenced by personal inclination and values. The student of third-world politics is likely to have a greater desire for travel and a greater tolerance for difficult living conditions than the student of congressional policy making; the analyst of international cooperation may have a particular distaste for violent conflict.

These personal experiences and values often provide the motivation to become a social scientist and, later, to choose a particular research question. As such, they may constitute the "real" reasons for engaging in a particular research project—and appropriately so. But, no matter how personal or idiosyncratic the reasons for choosing a topic, the methods of science and rules of inference discussed in this book will help scholars devise more powerful research designs. From the perspective of a potential contribution to social science, personal reasons are neither necessary nor sufficient justifications for the choice of a topic. In most cases, they should not appear in our scholarly writings. To put it most directly but quite indelicately, no one cares what we think—the scholarly community only cares what we can demonstrate.

Though precise rules for choosing a topic do not exist, there are ways—beyond individual preferences—of determining the likely value of a research enterprise to the scholarly community. Ideally, all research projects in the social sciences should satisfy two criteria. First, *a research project should pose a question that is "important" in the real world*. The topic should be consequential for political, social, or economic life, for understanding something that significantly affects many people's lives, or for understanding and predicting events that might be harmful or beneficial (see Shively 1990:15). Second, *a research project should make a specific contribution to an identifiable scholarly literature by increasing our collective ability to construct verified scientific explanations of some aspect of the world*. This latter criterion does not imply that all research that contributes to our stock of social science explanations in fact aims directly at making causal inferences. Sometimes the state of knowledge in a field is such that much fact-finding and description is needed before we can take on the challenge of explanation. Often the contribution of a single project will be descriptive inference. Sometimes the goal may not even be descriptive inference but rather will be the close observation of particular events or the summary of historical detail. These, however, meet our second criterion because they are prerequisites to explanation.

Our first criterion directs our attention to the real world of politics and social phenomena and to the current and historical record of the events and problems that shape people's lives. Whether a research question meets this criterion is essentially a societal judgment. The second criterion directs our attention to the scholarly literature of social science, to the intellectual puzzles not yet posed, to puzzles that remain to be solved, and to the scientific theories and methods available to solve them.

Political scientists have no difficulty finding subject matter that meets our first criterion. Ten major wars during the last four hundred years have killed almost thirty million people (Levy 1985:372); some "limited wars," such as those between the United States and North Vietnam and between Iran and Iraq, have each claimed over a million lives; and nuclear war, were it to occur, could kill billions of human beings. Political mismanagement, both domestic and international, has led to economic privation on a global basis—as in the 1930s—as well as to regional and local depression, as evidenced by the tragic experiences of much of Africa and Latin America during the 1980s. In general, cross-national variation in political institutions is associated with great variation in the conditions of ordinary human life, which are reflected in differences in life expectancy and infant mortality between countries with similar levels of economic development (Russett 1978:913–28). Within the United States, programs designed to alleviate poverty or social disorganization seem to have varied greatly in their efficacy. It cannot be doubted that research which contributes even marginally to an understanding of these issues is important.

While social scientists have an abundance of significant questions that can be investigated, the tools for understanding them are scarce and rather crude. Much has been written about war or social misery that adds little to the understanding of these issues because it fails either to describe these phenomena systematically or to make valid causal or descriptive inferences. Brilliant insights can contribute to understanding by yielding interesting new hypotheses, but brilliance is not a method of empirical research. All hypotheses need to be evaluated empirically before they can make a contribution to knowledge. This book offers no advice on becoming brilliant. What it can do, however, is to emphasize the importance of conducting research so that it constitutes a contribution to knowledge.

Our second criterion for choosing a research question, "making a contribution," means explicitly locating a research design within the framework of the existing social scientific literature. This ensures that the investigator understand the "state of the art" and minimizes the chance of duplicating what has already been done. It also guarantees that the work done will be important to others, thus improving the success of the community of scholars taken as a whole. Making an explicit contribution to the literature can be done in many different ways. We list a few of the possibilities here:

1. Choose a hypothesis seen as important by scholars in the literature but for which no one has completed a systematic study. If we find evidence in favor of or opposed to the favored hypothesis, we will be making a contribution.

2. Choose an accepted hypothesis in the literature that we suspect is false (or one we believe has not been adequately confirmed) and investigate whether it is indeed false or whether some other theory is correct.
3. Attempt to resolve or provide further evidence of one side of a controversy in the literature—perhaps demonstrate that the controversy was unfounded from the start.
4. Design research to illuminate or evaluate unquestioned assumptions in the literature.
5. Argue that an important topic has been overlooked in the literature and then proceed to contribute a systematic study to the area.
6. Show that theories or evidence designed for some purpose in one literature could be applied in another literature to solve an existing but apparently unrelated problem.

Focusing too much on making a contribution to a scholarly literature without some attention to topics that have real-world importance runs the risk of descending to politically insignificant questions. Conversely, attention to the current political agenda without regard to issues of the amenability of a subject to systematic study within the framework of a body of social science knowledge leads to careless work that adds little to our deeper understanding.

Our two criteria for choosing research questions are not necessarily in opposition to one another. In the long run, understanding real-world phenomena is enhanced by the generation and evaluation of explanatory hypotheses through the use of the scientific method. But in the short term, there may be a contradiction between practical usefulness and long-term scientific value. For instance, Mankiw (1990) points out that macroeconomic theory and applied macroeconomics diverged sharply during the 1970s and 1980s: models that had been shown to be theoretically incoherent were still used to forecast the direction of the U.S. economy, while the new theoretical models designed to correct these flaws remained speculative and were not sufficiently refined to make accurate predictions.

The criteria of practical applicability to the real world and contribution to scientific progress may seem opposed to one another when a researcher chooses a topic. Some researchers will begin with a real-world problem that is of great social significance: the threat of nuclear war, the income gap between men and women, the transition to democracy in Eastern Europe. Others may start with an intellectual problem generated by the social science literature: a contradiction between several experimental studies of decision-making under uncertainty or an inconsistency between theories of congressional voting and recent election outcomes. The distinction between the criteria is, of course, not hard and fast. Some research questions satisfy both criteria from the beginning, but in designing research, researchers often begin nearer one than the other.[4]

Wherever it begins, the process of designing research to answer a specific question should move toward the satisfaction of our two criteria. And obviously our direction of movement will depend on where we start. If we are motivated by a social scientific puzzle, we must ask how to make that research topic more relevant to real-world topics of significance—for instance, how might laboratory experiments better illuminate real-world strategic choices by political decision-makers or, what behavioral consequences might the theory have. If we begin with a real-world problem, we should ask how that problem can be studied with modern scientific methods so that it contributes to the stock of social science explanations. It may be that we will decide that moving too far from one criterion or the other is not the most fruitful approach. Laboratory experimenters may argue that the search for external referents is premature and that more progress will be made by refining theory and method in the more controlled environment of the laboratory. And in terms of a long-term research program, they may be right. Conversely, the scholar motivated by a real-world problem may argue that accurate description is needed before moving to explanation. And such a researcher may also be right. Accurate description is an important step in explanatory research programs.

In either case, a research program, and if possible a specific research project, should aim to satisfy our two criteria: it should deal with a significant real-world topic and be designed to contribute, directly or indirectly, to a specific scholarly literature. Since our main concern in this book is making qualitative research more scientific, we will primarily address the researcher who starts with the "real-world" perspective. But our analysis is relevant to both types of investigator.

If we begin with a significant real-world problem rather than with an established literature, it is essential to devise a workable plan for studying it. *A proposed topic that cannot be refined into a specific research project permitting valid descriptive or causal inference should be modified along the way or abandoned.* A proposed topic that will make no contri-

[4] The dilemma is not unlike that faced by natural scientists in deciding whether to conduct applied or basic research. For example, applied research in relation to a particular drug or disease may, in the short run, improve medical care without contributing as much to the general knowledge of the underlying biological mechanisms. Basic research may have the opposite consequence. Most researchers would argue, as we do for the social sciences, that the dichotomy is false and that basic research will ultimately lead to the powerful applied results. However, all agree that the best research design is one that somehow manages both to be directly relevant to solving real-world problems and to furthering the goals of a specific scientific literature.

bution to some scholarly literature should similarly be changed. Having tentatively chosen a topic, we enter a dialogue with the literature. What questions of interest to us have already been answered? How can we pose and refine our question so that it seems capable of being answered with the tools available? We may start with a burning issue, but we will have to come to grips both with the literature of social science and the problems of inference.

1.2.2 Improving Theory

A social science theory is a reasoned and precise speculation about the answer to a research question, including a statement about why the proposed answer is correct. Theories usually imply several more specific descriptive or causal hypotheses. A theory must be consistent with prior evidence about a research question. "A theory that ignores existing evidence is an oxymoron. If we had the equivalent of 'truth in advertising' legislation, such an oxymoron should not be called a theory" (Lieberson 1992:4; see also Woods and Walton 1982).

The development of a theory is often presented as the first step of research. It sometimes comes first in practice, but it need not. In fact, we cannot develop a theory without knowlege of prior work on the subject and the collection of some data, since even the research question would be unknown. Nevertheless, despite whatever amount of data has already been collected, there are some general ways to evaluate and improve the usefulness of a theory. We briefly introduce each of these here but save a more detailed discussion for later chapters.

First, choose theories that could be wrong. Indeed, vastly more is learned from theories that *are* wrong than from theories that are stated so broadly that they could not be wrong even in principle.[5] We need to be able to give a direct answer to the question: What evidence would convince us that we are wrong?[6] If there is no answer to this question, then we do not have a theory.

Second, to make sure a theory is falsifiable, choose one that is capable of generating as many *observable implications* as possible. This choice will allow more tests of the theory with more data and a greater variety of data, will put the theory at risk of being falsified more times, and will make it possible to collect data so as to build strong evidence for the theory.

[5] This is the principle of falsifiability (Popper 1968). It is an issue on which there are varied positions in the philosophy of science. However, very few of them disagree with the principle that theories should be stated clearly enough so that they could be wrong.

[6] This is probably the most commonly asked question at job interviews in our department and many others.

Third, in designing theories, be as concrete as possible. Vaguely stated theories and hypotheses serve no purpose but to obfuscate. Theories that are stated precisely and make specific predictions can be shown more easily to be wrong and are therefore better.

Some researchers recommend following the principle of "parsimony." Unfortunately, the word has been used in so many ways in casual conversation and scholarly writings that the principle has become obscured (see Sober [1988] for a complete discussion). The clearest definition of parsimony was given by Jeffreys (1961:47): "Simple theories have higher prior probabilities."[7] Parsimony is therefore a judgment, or even assumption, about the nature of the world: it is assumed to be simple. The principle of choosing theories that imply a simple world is a rule that clearly applies in situations where there is a high degree of certainty that the world is indeed simple. Scholars in physics seem to find parsimony appropriate, but those in biology often think of it as absurd. In the social sciences, some forcefully defend parsimony in their subfields (e.g., Zellner 1984), but we believe it is only occasionally appropriate. Given the precise definition of parsimony as an assumption about the world, we should never insist on parsimony as a general principle of designing theories, but it is useful in those situations where we have some knowledge of the simplicity of the world we are studying.

Our point is that we do not advise researchers to seek parsimony as an essential good, since there seems little reason to adopt it unless we already know a lot about a subject. We do not even need parsimony to avoid excessively complicated theories, since it is directly implied by the maxim that the theory should be just as complicated as all our evidence suggest. Situations with insufficient evidence relative to the complexity of the theory being investigated can lead to what we call "indeterminate research designs" (see section 4.1), but these are problems of research design and not assumptions about the world.

All our advice thus far applies if we have not yet collected our data and begun any analysis. However, if we have already gathered the data, we can certainly use these rules to modify our theory and gather new data, and thus generate new observable implications of the new theory. Of course, this process is expensive, time consuming, and probably wasteful of the data already collected. What then about the situation where our theory is in obvious need of improvement but we cannot afford to collect additional data? This situation—in which researchers often find themselves—demands great caution and self-

[7] This phrase has come to be known as the "Jeffreys-Wrinch Simplicity Postulate." The concept is similar to Occam's razor.

restraint. Any intelligent scholar can come up with a "plausible" theory for any set of data after the fact, yet to do so demonstrates nothing about the veracity of the theory. The theory will fit the data nicely and still may be wildly wrong—indeed, demonstrably wrong with most other data. Human beings are very good at recognizing patterns but not very good at recognizing nonpatterns. (Most of us even see patterns in random ink blots!) Ad hoc adjustments in a theory that does not fit existing data must be used rarely and with considerable discipline.[8]

There is still the problem of what to do when we have finished our data collection and analysis and wish to work on improving a theory. In this situation, we recommend following two rules: First, if our prediction is conditional on several variables and we are willing to drop one of the conditions, we may do so. For example, if we hypothesized originally that democratic countries with advanced social welfare systems do not fight each other, it would be permissible to extend that hypothesis to all modern democracies and thus evaluate our theory against more cases and increase its chances of being falsified. The general point is that after seeing the data, we may modify our theory in a way that makes it apply to a larger range of phenomena. Since such an alteration in our thesis exposes it more fully to falsification, modification in this direction should not lead to ad hoc explanations that merely appear to "save" an inadequate theory by restricting its range to phenomena that have already been observed to be in accord with it.

The opposite practice, however, is generally inappropriate. After observing the data, we should not just add a restrictive condition and then proceed as if our theory, with that qualification, has been shown to be correct. If our original theory was that modern democracies do not fight wars with one another due to their constitutional systems, it would be less permissible, having found exceptions to our "rule," to restrict the proposition to democracies with advanced social welfare systems *once it has been ascertained by inspection of the data that such a qualification would appear to make our proposition correct*. Or suppose that our original theory was that revolutions only occur under conditions of severe economic depression, but we find that this is not true in one of our case studies. In this situation it would not be reasonable merely to add general conditions such as, revolutions never occur during periods of prosperity except when the military is weak, the political leadership is repressive, the economy is based on a small number of prod-

[8] If we have chosen a topic of real-world importance and/or one which makes some contribution to a scholarly literature, the social nature of academia will correct this situation: someone will replicate our study with another set of data and demonstrate that we were wrong.

ucts, and the climate is warm. Such a formulation is merely a fancy (and misleading) way of saying "my theory is correct, except in country *x*." Since we have already discovered that our theory is incorrect for country *x*, it does not help to turn this falsification into a spurious generalization. Without efforts to collect new data, we will have no admissible evidence to support the new version of the theory.

So our basic rule with respect to altering our theory after observing the data is: *we can make the theory less restrictive (so that it covers a broader range of phenomena and is exposed to more opportunities for falsification), but we should not make it more restrictive without collecting new data to test the new version of the theory*. If we cannot collect additional data, then we are stuck; and we do not propose any magical way of getting unstuck. At some point, deciding that we are wrong is best; indeed, negative findings can be quite valuable for a scholarly literature. Who would not prefer one solid negative finding over any number of flimsy positive findings based on ad hoc theories?

Moreover, if we are wrong, we need not stop writing after admitting defeat. We may add a section to our article or a chapter to our book about future empirical research and current theoretical speculation. In this context, we have considerably more freedom. We may suggest additional conditions that might be plausibly attached to our theory, if we believe they might solve the problem, propose a modification of another existing theory or propose a range of entirely different theories. In this situation, we cannot conclude anything with a great deal of certainty (except perhaps that the theory we stated at the outset is wrong), but we do have the luxury of inventing new research designs or data-collection projects that could be used to decide whether our speculations are correct. These can be very valuable, especially in suggesting areas where future researchers can look.

Admittedly, as we discussed above, social science does not operate strictly according to rules: the need for creativity sometimes mandates that the textbook be discarded! And data can discipline thought. Hence researchers will sometimes, after confronting data, have inspirations about how they should have constructed the theory in the first place. Such a modification, even if restrictive, may be worthwhile if we can convince ourselves and others that modifying the theory in the way that we propose is something we could have done before we collected the data if we had thought of it. But until tested with *new* data, the status of such a theory will remain very uncertain, and it should be labeled as such.

One important consequence of these rules is that pilot projects are often very useful, especially in research where data must be gathered by interviewing or other particularly costly means. Preliminary datagathering may lead us to alter the research questions or modify the theory. Then new data can be gathered to test the new theory, and the problem of using the same data to generate and test a theory can be avoided.

1.2.3 Improving Data Quality

"Data" are systematically collected elements of information about the world. They can be qualitative or quantitative in style. Sometimes data are collected to evaluate a very specific theory, but not so infrequently, scholars collect data before knowing precisely what they are interested in finding out. Moreover, even if data are collected to evaluate a specific hypothesis, researchers may ultimately be interested in questions that had not occurred to them previously.

In either case—when data are gathered for a specific purpose or when data are used for some purpose not clearly in mind when they were gathered—certain rules will improve the quality of those data. In principle, we can think about these rules for improving data separately from the rules in section 1.2.2 for improving theory. In practice any data-collection effort requires some degree of theory, just as formulating any theory requires some data (see Coombs 1964).

Our first and most important guideline for improving data quality is: *record and report the process by which the data are generated*. Without this information we cannot determine whether using standard procedures in analyzing the data will produce biased inferences. Only by knowing the process by which the data were generated will we be able to produce valid descriptive or causal inferences. In a quantitative opinion poll, recording the data-generation process requires that we know the exact method by which the sample was drawn and the specific questions that were asked. In a qualitative comparative case study, reporting the precise rules by which we choose the small number of cases for analysis is critical. We give additional guidelines in chapter 6 for case selection in qualitative research, but even more important than choosing a good method is being careful to record and report whatever method was used and all the information necessary for someone else to apply it.[9]

In section 1.2.2 we argued for theories that are capable of generating

[9] We find that many graduate students are unnecessarily afraid of sharing data and the information necessary to replicate their results. They are afraid that someone will steal their hard work or even prove that they were wrong. These are all common fears, but they are almost always unwarranted. Publication (or at least sending copies of research papers to other scholars) and sharing data is the best way to guarantee credit for one's contributions. Moreover, sharing data will only help others follow along in the research you started. When their research is published, they will cite your effort and advance your visibility and reputation.

many observable implications. Our second guideline for improving data quality is *in order better to evaluate a theory, collect data on as many of its observable implications as possible*. This means collecting as much data in as many diverse contexts as possible. Each additional implication of our theory which we observe provides another context in which to evaluate its veracity. The more observable implications which are found to be consistent with the theory, the more powerful the explanation and the more certain the results.

When adding data on new observable implications of a theory, we can (a) collect more observations on the same dependent variable, or (b) record additional dependent variables. We can, for instance, disaggregate to shorter time periods or smaller geographic areas. We can also collect information on dependent variables of less direct interest; if the results are as the theory predicts, we will have more confidence in the theory.

For example, consider the rational deterrence theory: potential initiators of warfare calculate the costs and benefits of attacking other states, and these calculations can be influenced by credible threats of retaliation. The most direct test of this theory would be to assess whether, given threats of war, decisions to attack are associated with such factors as the balance of military forces between the potential attacker and the defender or the interests at stake for the defender (Huth 1988). However, even though using only cases in which threats are issued constitutes a set of observable implications of the theory, they are only part of the observations that could be gathered (and used alone may lead to selection bias), since situations in which threats themselves are deterred would be excluded from the data set. Hence it might be worthwhile also to collect data on an additional dependent variable (i.e., a different set of observable implications) based on a measurement of whether threats are made by states that have some incentives to do so.

Insofar as sufficient good data on deterrence in international politics is lacking, it could also be helpful to test a different theory, one with similar motivational assumptions, for a different dependent variable under different conditions but which is still an observable implication of the same theory. For instance, we could construct a laboratory experiment to see whether, under simulated conditions, "threats" are deterred rather than accentuated by military power and firm bargaining behavior. Or we could examine whether other actors in analogous situations, such as oligopolistic firms competing for market share or organized-crime families competing for turf, use deterrence strategies and how successful they are under varying conditions. Indeed, economists working in the field of industrial organization have used noncooperative game theory, on which deterrence theory also relies, to study such problems as entry into markets and pricing strategies (Fudenberg and Tirole 1989). Given the close similarity between the theories, empirical evidence supporting game theory's predictions about firm behavior would increase the plausibility of related hypotheses about state behavior in international politics. Uncertainty would remain about the applicability of conclusions from one domain to another, but the issue is important enough to warrant attempts to gain insight and evidence wherever they can be found.

Obviously, to collect data forever without doing any analysis would preclude rather than facilitate completion of useful research. In practice, limited time and resources will always constrain data-collection efforts. Although more information, additional cases, extra interviews, another variable, and other relevant forms of data collection will always improve the certainty of our inferences to some degree, promising, potential scholars can be ruined by too much information as easily as by too little. Insisting on reading yet another book or getting still one more data set without ever writing a word is a prescription for being unproductive.

Our third guideline is: *maximize the validity of our measurements*. Validity refers to measuring what we think we are measuring. The unemployment rate may be a good indicator of the state of the economy, but the two are not synonymous. In general, it is easiest to maximize validity by adhering to the data and not allowing unobserved or unmeasurable concepts get in the way. If an informant responds to our question by indicating ignorance, then we know he *said* that he was ignorant. Of that, we have a valid measurement. However, what he really *meant* is an altogether different concept—one that cannot be measured with a high degree of confidence. For example, in countries with repressive governments, expressing ignorance may be a way of making a critical political statement for some people; for others, it is a way of saying "I don't know."

Our fourth guideline is: *ensure that data-collection methods are reliable*. Reliability means that applying the same procedure in the same way will always produce the same measure. When a reliable procedure is applied at different times and nothing has happened in the meantime to change the "true" state of the object we are measuring, the same result will be observed.[10] Reliable measures also produce the same re-

[10] We can check reliability ourselves by measuring the same quantity twice and seeing whether the measures are the same. Sometimes this seems easy, such as literally asking the same question at different times during an interview. However, asking the question once may influence the respondent to respond in a consistent fashion the second time, so we need to be careful that the two measurements are indeed independent.

sults when applied by different researchers, and this outcome depends, of course, upon there being explicit procedures that can be followed.[11]

Our final guideline is: *all data and analyses should, insofar as possible, be replicable*. Replicability applies not only to data, so that we can see whether our measures are reliable, but to the entire reasoning process used in producing conclusions. On the basis of our research report, a new researcher should be able to duplicate our data and trace the logic by which we reached our conclusions. Replicability is important even if no one actually replicates our study. Only by reporting the study in sufficient detail so that it can be replicated is it possible to evaluate the procedures followed and methods used.

Replicability of data may be difficult or impossible in some kinds of research: interviewees may die or disappear, and direct observations of real-world events by witnesses or participants cannot be repeated. Replicability has also come to mean different things in different research traditions. In quantitative research, scholars focus on replicating the analysis after starting with the same data. As anyone who has ever tried to replicate the quantitative results of even prominent published works knows well, it is usually a lot harder than it should be and always more valuable than it seems at the outset (see Dewald et al. 1986 on replication in quantitative research).

The analogy in traditional qualitative research is provided by footnotes and bibliographic essays. Using these tools, succeeding scholars should be able to locate the sources used in published work and make their own evaluations of the inferences claimed from this information. For research based on direct observation, replication is more difficult. One scholar could borrow another's field notes or tape recorded interviews to see whether they support the conclusions made by the original investigator. Since so much of the data in field research involve conversations, impressions, and other unrecorded participatory information, this reanalysis of results using the same data is not often done. However, some important advances might be achieved if more scholars tried this type of replication, and it would probably also encourage others to keep more complete field notes. Occasionally, an entire research project, including data collection, has been replicated. Since we cannot go back in time, the replication cannot be perfect but can be quite valuable nonetheless. Perhaps the most extensive replication of

[11] An example is the use of more than one coder to extract systematic information from transcripts of in-depth interviews. If two people use the same coding rules, we can see how often they produce the same judgment. If they do not produce reliable measures, then we can make the coding rules more precise and try again. Eventually, a set of rules can often be generated so that the application of the same procedure by different coders will yield the same result.

a qualitative study is the sociological study of Middletown, Indiana, begun by Robert and Helen Lynd. Their first "Middletown" study was published in 1929 and was replicated in a book published in 1937. Over fifty years after the original study, a long series of books and articles are being published that replicate these original studies (see Caplow et al., 1983a, 1983b and the citations therein). All qualitative replication need not be this extensive, but this major research project should serve as an exemplar for what is possible.

All research should attempt to achieve as much replicability as possible: scholars should always record the exact methods, rules, and procedures used to gather information and draw inferences so that another researcher can do the same thing and draw (one hopes) the same conclusion. Replicability also means that scholars who use unpublished or private records should endeavor to ensure that future scholars will have access to the material on similar terms; taking advantage of privileged access without seeking access for others precludes replication and calls into question the scientific quality of the work. Usually our work will not be replicated, but we have the responsibility to act as if someone may wish to do so. Even if the work is not replicated, providing the materials for such replication will enable readers to understand and evaluate what we have done.

1.2.4 Improving the Use of Existing Data

Fixing data problems by collecting new and better data is almost always an improvement on trying to use existing, flawed data in better ways; however, the former approach is not always possible. Social scientists often find themselves with problematic data and little chance to acquire anything better; thus, they have to make the best of what they have.

Improving the use of previously collected data is the main topic taught in classes on statistical methods and is, indeed, the chief contribution of inferential statistics to the social sciences. The precepts on this topic that are so clear in the study of inferential statistics also apply to qualitative research. The remainder of this book deals with these precepts more fully. Here we provide merely a brief outline of the guidelines for improving the use of previously collected data.

First, whenever possible, we should use data to generate inferences that are "unbiased," that is, correct on average. To understand this very specific idea from statistical research, imagine applying the same methodology (in quantitative or qualitative research) for analyzing and drawing conclusions from data across many data sets. Because of small errors in the data or in the application of the procedure, a single application of this methodology would probably never be exactly correct. An "unbiased" procedure will be correct when taken as an average across many applications—even if no single application is correct. The procedure will not systematically tilt the outcome in one direction or another.

Achieving unbiased inferences depends, of course, both on the original collection of the data and its later use; and, as we pointed out before, it is always best to anticipate problems before data collection begins. However, we mention these issues briefly here because when using the data, we need to be particularly careful to analyze whether sources of bias were overlooked during data collection. One such source, which can lead to biased inferences, is that of selection bias: choosing observations in a manner that systematically distorts the population from which they were drawn. Although an obvious example is deliberately choosing only cases which support our theory, selection bias can occur in much more subtle ways. Another difficulty can result from omitted variable bias, which refers to the exclusion of some control variable that might influence a seeming causal connection between our explanatory variables and that which we want to explain. We discuss these and numerous other potential pitfalls in producing unbiased inferences in chapters 2–6.

The second guideline is based on the statistical concept of "efficiency": an efficient use of data involves maximizing the information used for descriptive or causal inference. Maximizing efficiency requires not only using all our data, but also using all the relevant information in the data to improve inferences. For example, if the data are disaggregated into small geographical units, we should use it that way, not just as a national aggregate. The smaller aggregates will have larger degrees of uncertainty associated with them, but if they are, at least in part, observable implications of the theory, they will contain some information which can be brought to bear on the inference problem.

1.3 Themes of This Volume

We conclude this overview chapter by highlighting the four important themes in developing research designs that we have discussed here and will elaborate throughout this book.

1.3.1 Using Observable Implications to Connect Theory and Data

In this chapter we have emphasized that every theory, to be worthwhile, must have implications about the observations we expect to find if the theory is correct. These *observable implications* of the theory must guide our data collection, and help distinguish relevant from irrelevant facts. In chapter 2.6 we discuss how theory affects data collection, as well as how data disciplines theoretical imagination. Here, we want to stress that theory and empirical research must be tightly connected. Any theory that does real work for us has implications for empirical investigation; no empirical investigation can be successful without theory to guide its choice of questions. Theory and data collection are both essential aspects of the process by which we seek to decide whether a theory should be provisionally viewed true or false, subject as it is in both cases to the uncertainty that characterizes all inference.

We should ask of any theory: What are its observable implications? We should ask about any empirical investigations: Are the observations relevant to the implications of our theory, and, if so, what do they enable us to infer about the correctness of the theory? In any social scientific study, the implications of the theory and the observation of facts need to mesh with one another: social science conclusions cannot be considered reliable if they are not based on theory and data in strong connection with one another and forged by formulating and examining the observable implications of a theory.

1.3.2 Maximizing Leverage

The scholar who searches for additional implications of a hypothesis is pursuing one of the most important achievements of all social science: *explaining as much as possible with as little as possible*. Good social science seeks to increase the significance of what is explained relative to the information used in the explanation. If we can accurately explain what at first appears to be a complicated effect with a single causal variable or a few variables, the *leverage* we have over a problem is very high. Conversely, if we can explain many effects on the basis of one or a few variables we also have high leverage. Leverage is low in the social sciences in general and even more so in particular subject areas. This may be because scholars do not yet know how to increase it or because nature happens not to be organized in a convenient fashion or for both of these reasons. Areas conventionally studied qualitatively are often those in which leverage is low. Explanation of anything seems to require a host of explanatory variables: we use a lot to explain a little. In such cases, our goal should be to design research with more leverage.

There are various ways in which we can increase our leverage over a research problem. The primary way is to increase the number of observable implications of our hypothesis and seek confirmation of those implications. As we have described above, this task can involve

(1) improving the theory so that it has more observable implications, (2) improving the data so more of these implications are indeed observed and used to evaluate the theory, and (3) improving the use of the data so that more of these implications are extracted from existing data. None of these, nor the general concept of maximizing leverage, are the same as the concept of parsimony, which, as we explained in section 1.2.2, is an assumption about the nature of the world rather than a rule for designing research.

Maximizing leverage is so important and so general that *we strongly recommend that researchers routinely list all possible observable implications of their hypothesis that might be observed in their data or in other data.* It may be possible to test some of these new implications in the original data set—as long as the implication does not "come out of" the data but is a hypothesis independently suggested by the theory or a different data set. But it is better still to turn to other data. Thus we should also consider implications that might appear in other data—such as data about other units, data about other aspects of the units under study, data from different levels of aggregation, and data from other time periods such as predictions about the near future—and evaluate the hypothesis in those settings. The more evidence we can find in varied contexts, the more powerful our explanation becomes, and the more confidence we and others should have in our conclusions.

At first thought, some researchers may object to the idea of collecting observable implications from any source or at any level of aggregation different from that for which the theory was designed. For example, Lieberson (1985) applies to qualitative research the statistical idea of "ecological fallacy"—incorrectly using aggregate data to make inferences about individuals—to warn against cross-level inference.[12] We certainly agree that we can use aggregate data to make incorrect inferences about individuals: if we are interested in individuals, then studying individuals is generally a better strategy if we can obtain these data. However, if the inference we seek to make is more than a very narrowly cast hypothesis, our theory may have implications at many levels of analysis, and we will often be able to use data from all these levels to provide some information about our theory. Thus, even if we are primarily interested in an aggregate level of analysis, we can

[12] The phrase "ecological fallacy" is confusing because the process of reasoning from aggregate- to individual-level processes is neither ecological nor a fallacy. "Ecological" is an unfortunate choice of word to describe the aggregate level of analysis. Although Robinson (1990) concluded in his original article about this topic that using aggregate analysis to reason about individuals is a fallacy, quantitative social scientists and statisticians now widely recognize that some information about individuals does exist at aggregate levels of analysis, and many methods of unbiased "ecological" inference have been developed.

often gain leverage about our theory's veracity by looking at the data from these other levels.

For example, if we develop a theory to explain revolutions, we should look for observable implications of that theory not only in overall outcomes but also such phenomena as the responses to in-depth interviews of revolutionaries, the reactions of people in small communities in minor parts of the country, and official statements by party leaders. We should be willing to take whatever information we can acquire so long as it helps us learn about the veracity of our theory. If we can test our theory by examining outcomes of revolutions, fine. But in most cases very little information exists at that level, perhaps just one or a few observations, and their values are rarely unambiguous or measured without error. Many different theories are consistent with the existence of a revolution. Only by delving deeper in the present case, or bringing in relevant information existing in other cases, is it possible to distinguish among previously indistinguishable theories.

The only issue in using information at other levels and from other sources to study a theory designed at an aggregate level is whether these new observations contain *some* information that is relevant to evaluating implications of our theory. If these new observations help to test our theory, they should be used even if they are not the implications of greatest interest. For example, we may not care at all about the views of revolutionaries, but if their answers to our questions are consistent with our theory of revolutions, then the theory itself will be more likely to be correct, and the collection of additional information will have been useful. In fact, an observation at the most aggregate level of data analysis—the occurrence of a predicted revolution, for example—is merely one observed implication of the theory, and because of the small amount of information in it, it should not be privileged over other observable implications. We need to collect information on as many observable implications of our theory as possible.

1.3.3 Reporting Uncertainty

All knowledge and all inference—in quantitative and in qualitative research—is uncertain. Qualitative measurement is error-prone, as is quantitative, but the sources of error may differ. The qualitative interviewer conducting a long, in-depth interview with a respondent whose background he has studied is less likely to mismeasure the subject's real political ideology than is a survey researcher conducting a structured interview with a randomly selected respondent about whom he knows nothing. (Although the opposite is also possible if, for instance, he relies too heavily on an informant who is not trustworthy.) However, the survey researcher is less likely to generalize inappropriately from the particular cases interviewed to the broader population than is the in-depth researcher. Neither is immune from the uncertainties of measurement or the underlying probabilistic nature of the social world.

All good social scientists—whether in the quantitative or qualitative traditions—report estimates of the uncertainty of their inferences. Perhaps the single most serious problem with qualitative research in political science is the pervasive failure to provide reasonable estimates of the uncertainty of the investigator's inferences (see King 1990). We can make a valid inference in almost any situation, no matter how limited the evidence, by following the rules in this book, but we should avoid forging sweeping conclusions from weak data. The point is not that reliable inferences are impossible in qualitative research, but rather that we should always report a reasonable estimate of the *degree of certainty* we have in each of our inferences. Neustadt and May (1986:274), dealing with areas in which precise quantitative estimates are difficult, propose a useful method of encouraging policymakers (who are often faced with the necessity of reaching conclusions about what policy to follow out of inadequate data) to judge the uncertainty of their conclusions. They ask "How much of your own money would you wager on it?" This makes sense as long as we also ask, "At what odds?"

1.3.4 Thinking like a Social Scientist: Skepticism and Rival Hypotheses

The uncertainty of causal inferences means that good social scientists do not easily accept them. When told A causes B, someone who "thinks like a social scientist" asks whether that connection is a true causal one. It is easy to ask such questions about the research of others, but it is more important to ask them about our own research. There are many reasons why we might be skeptical of a causal account, plausible though it may sound at first glance. We read in the newspaper that the Japanese eat less red meat and have fewer heart attacks than Americans. This observation alone is interesting. In addition, the explanation—too much steak leads to the high rate of heart disease in the United States—is plausible. The skeptical social scientist asks about the accuracy of the data (how do we know about eating habits? what sample was used? are heart attacks classified similarly in Japan and the United States so that we are comparing similar phenomena?). Assuming that the data are accurate, what else might explain the effects: Are there other variables (other dietary differences, genetic features, life-style characteristics) that might explain the result? Might we have inadvertently reversed cause and effect? It is hard to imagine how not having a heart attack might cause one to eat less red meat but it is possible. Perhaps people lose their appetite for hamburgers and steak late in life. If this were the case, those who did not have a heart attack (for whatever reason) would live longer and eat less meat. This fact would produce the same relationship that led the researchers to conclude that meat was the culprit in heart attacks.

It is not our purpose to call such medical studies into question. Rather we wish merely to illustrate how social scientists approach the issue of causal inference: with skepticism and a concern for alternative explanations that may have been overlooked. Causal inference thus becomes a *process* whereby each conclusion becomes the occasion for further research to refine and test it. Through successive approximations we try to come closer and closer to accurate causal inference.

CHAPTER 3

Causality and Causal Inference

We have discussed two stages of social science research: summarizing historical detail (section 2.5) and making descriptive inferences by partitioning the world into systematic and nonsystematic components (section 2.6). Many students of social and political phenomena would stop at this point, eschewing causal statements and asking their selected and well-ordered facts to "speak for themselves."

Like historians, social scientists need to summarize historical detail and to make descriptive inferences. For some social scientific purposes, however, analysis is incomplete without causal inference. That is, just as causal inference is impossible without good descriptive inference, descriptive inference alone is often unsatisfying and incomplete. To say this, however, is not to claim that all social scientists must, in all of their work, seek to devise causal explanations of the phenomena they study. Sometimes causal inference is too difficult; in many other situations, descriptive inference is the ultimate goal of the research endeavor.

Of course, we should always be explicit in clarifying whether the goal of a research project is description or explanation. Many social scientists are uncomfortable with causal inference. They are so wary of the warning that "correlation is not causation" that they will not state causal hypotheses or draw causal inferences, referring to their research as "studying association and not causation." Others make apparent causal statements with ease, labeling unevaluated hypotheses or speculations as "explanations" on the basis of indeterminate research designs.[1] We believe that each of these positions evades the problem of causal inference.

[1] In view of some social scientists' preference for explanation over "mere description," it is not surprising that students of complicated events seek to dress their work in the trappings of explanatory jargon; otherwise, they fear being regarded as doing inferior work. At its core, real explanation is always based on causal inferences. We regard arguments in the literature about "noncausal explanation" as confusing terminology; in virtually all cases, these arguments are really about causal explanation or are internally inconsistent. If social scientists' failures to explain are not due to poor research or lack of imagination, but rather to the nature of the difficult but significant problems that they are examining, such feelings of inferiority are unjustified. Good description of important events is better than bad explanation of anything.

Avoiding causal language when causality is the real subject of investigation either renders the research irrelevant or permits it to remain undisciplined by the rules of scientific inference. Our uncertainty about causal inferences will never be eliminated. But this uncertainty should not suggest that we avoid attempts at causal inference. Rather we should draw causal inferences where they seem appropriate but also provide the reader with the best and most honest estimate of the uncertainty of that inference. It is appropriate to be bold in drawing causal inferences as long as we are cautious in detailing the uncertainty of the inference. It is important, further, that causal hypotheses be disciplined, approximating as closely as possible the rules of causal inference. Our purpose in much of chapters 4–6 is to explicate the circumstances under which causal inference is appropriate and to make it possible for qualitative researchers to increase the probability that their research will provide reliable evidence about their causal hypotheses.

In section 3.1 we provide a rigorous definition of causality appropriate for qualitative and quantitative research, then in section 3.2 we clarify several alternative notions of causality in the literature and demonstrate that they do not conflict with our more fundamental definition. In section 3.3 we discuss the precise assumptions about the world and the hypotheses required to make reliable causal inferences. We then consider in section 3.4 how to apply to causal inference the criteria we developed for judging descriptive inference. In section 3.5 we conclude this chapter with more general advice on how to construct causal explanations, theories, and hypotheses.

3.1 Defining Causality

In this section, we define causality as a *theoretical* concept independent of the data used to learn about it. Subsequently, we consider causal *inference* from our data. (For discussions of specific problems of causal inference, see chapters 4–6.) In section 3.1.1 we give our definition of causality in full detail, along with a simple quantitative example, and in section 3.1.2 we revisit our definition along with a more sophisticated qualitative example.

3.1.1 The Definition and a Quantitative Example

Our theoretical definition of causality applies most simply and clearly to a single unit.[2] As defined in section 2.4, a unit is one of the many elements to be observed in a study, such as a person, country, year, or

[2] Our point of departure in this section is Holland's article (1986) on causality and

political organization. For precision and clarity, we have chosen a single running example from quantitative research: the causal effect of incumbency status for a Democratic candidate for the U.S. House of Representatives on the proportion of votes this candidate receives. (Using only a Democratic candidate simplifies the example.) Let the dependent variable be the Democratic proportion of the two-party vote for the House. The key causal explanatory variable is then dichotomous, either the Democrat is an incumbent or not. (For simplicity throughout this section, we only consider districts where the Republican candidate lost the last election.)

Causal language can be confusing and our choice here is hardly unique. The "dependent variable" is sometimes called the "outcome variable." "Explanatory variables" are often referred to as "independent variables." We divide the explanatory variables into the "key causal variable" (also called the "cause" or the "treatment variable") and the "control variables." Finally, the key causal variable always takes on two or more values, which are often denoted by "treatment group" and "control group."

Now consider only the Fourth Congressional District in New York, and imagine an election in 1998 with a Democratic incumbent and one Republican (nonincumbent) challenger. Suppose the Democratic candidate received y_4^I fraction of the vote in this election (the subscript 4 denotes the Fourth District in New York and the superscript *I* refers to the fact that the Democrat is an *I*ncumbent). y_4^I is then a value of the dependent variable. To *define* the causal effect (a *theoretical* quantity), imagine that we go back in time to the start of the election campaign and everything remains the same, except that the Democratic incumbent decides not to run for re-election and the Democratic Party nominates another candidate (presumably the winner of the primary election). We denote the fraction of the vote that the Democratic (nonincumbent) candidate would receive by y_4^N (where *N* denotes a Democratic candidate who is a *N*on-incumbent).[3]

This *counterfactual* condition is the essence behind this definition of causality, and the difference between the actual vote (y_4^I) and the likely

what he calls "Rubin's Model." Holland bases his ideas on the work of numerous scholars. Donald Rubin's (1974, 1978) work on the subject was most immediately relevant, but he also cites Aristotle, Locke, Hume, Mill, Suppes, Granger, Fisher, Neyman, and others. We extend Holland's definition of a causal effect by using some ideas expressed clearly by Suppes (1970) and others concerning "probabilistic causality." We found this extension necessary since no existing approach alone is capable of defining causality with respect to a single unit *and* still allowing one to partition causal effects into systematic and nonsystematic components.

[3] See Gelman and King (1990) for details of this example. More generally, *I* and *N* can stand for the "treatment" and "control" group or for any two treatments experimentally

vote in this counterfactual situation (y_4^N) is the causal effect, a concept we define more precisely below. We must be very careful in defining counterfactuals; although they are obviously counter to the facts, they must be reasonable and it should be possible for the counterfactual event to have occurred under precisely stated circumstances. A key part of defining the appropriate counterfactual condition is clarifying precisely what we are holding constant while we are changing the value of the treatment variable. In the present example, the key causal (or treatment) variable is incumbency status, and it changes from "incumbent" to "non-incumbent." During this hypothetical change, we hold everything constant up to the moment of the Democratic Party's nomination decision—the relative strength of the Democrats and Republicans in past elections in this district, the nature of the nomination process, the characteristics of the congressional district, and the economic and political climate at the time, etc. We do *not* control for qualities of the candidates, such as name recognition, visibility, and knowledge of the workings of Congress, or anything else that follows the party nomination. The reason is that these are partly *consequences* of our treatment variable, incumbency. That is, the advantages of incumbency include name recognition, visibility, and so forth. If we did hold these constant, we would be controlling for and hence disregarding some of the most important effects of incumbency and as a result, would misinterpret its overall effect on the vote total. In fact, controlling for enough of the consequences of incumbency could make one incorrectly believe that incumbency had no effect at all.[4]

More formally, the causal effect of incumbency in the Fourth District in New York—the proportion of the vote received by the Democratic Party candidate that is attributable to incumbency status—would be the difference between these two vote fractions: $(y_4^I - y_4^N)$. For reasons that will become clear shortly, we refer to this difference as the *realized*

administered in fact or in theory. Of course, the decision to call one value of an explanatory variable a treatment and the other a control is entirely arbitrary, if this language is used at all.

[4] Jon Elster (1983:34–36) has claimed "the meaning of causality can not be rendered by counterfactual statements" in many situations, such as those in which a third factor accounts for both the apparent explanatory and dependent variables. In our language, Elster is simply pointing to common problems of *inferences*, which are always uncertain to some extent. However, these difficulties of inference do not invalidate a *definition* of causality in terms of counterfactuals. Despite his objections, Elster acknowledges that counterfactual statements "have an important role in causal analysis" (Elster 1983:36). Hence Elster's argument is more cogent, we think, as a set of valuable warnings against careless use of counterfactuals than as a critique of their fundamental definitional importance in causal reasoning.

causal effect and write it in more general notation for unit *i* instead of only district 4:[5]

$$\text{(Realized Causal Effect for unit } i) = y_i^I - y_i^N \tag{3.1}$$

Of course, this effect is defined only in theory since in any one real election we might observe *either* y_4^I *or* y_4^N or neither, but never both. Thus, this simple definition of causality demonstrates that we can never hope to know a causal effect for certain. Holland (1986) refers to this problem as *the fundamental problem of causal inference*, and it is indeed a *fundamental* problem since no matter how perfect the research design, no matter how much data we collect, no matter how perceptive the observers, no matter how diligent the research assistants, and no matter how much experimental control we have, we will never know a causal inference for certain. Indeed, most of the empirical issues of research designs that we discuss in this book involve this fundamental problem, and most of our suggestions constitute partial attempts to avoid it.

Our working definition of causality differs from Holland's, since in section 2.6 we have argued that social science always needs to partition the world into systematic and nonsystematic components, and Holland's definition does not make this distinction clearly.[6] To see the importance of this partitioning, think about what would happen if we could rerun the 1998 election campaign in the Fourth District in New York, with a Democratic incumbent and a Republican challenger. A slightly different total vote would result, due to nonsystematic features of election campaigns—aspects of politics that do not persist from one campaign to the next, even if the campaigns begin on identical footing. Some of these nonsystematic features might include a verbal gaffe, a surprisingly popular speech or position on an issue, an unexpectedly bad performance in a debate, bad weather during one candidate's rally or on election day, or the results of some investigative journalism. We can therefore imagine a variable that would express the values of the Democratic vote across hypothetical replications of this same election.

[5] We can specialize for district 4 by substituting "4" for "*i*" in the following equation.

[6] The reason for this is probably that Holland is a statistician who comes very close to an extreme version of "Perspective 2" random variation, which is described in section 2.6. In his description of the "statistical solution" to the problem of causal inference, he most closely approximates our definition of a causal effect, but this definition is mostly about using different units to solve the Fundamental Problem instead of retaining the definition of causality in just one. In particular, his expected value operator averages over units, whereas ours (described below) averages over hypothetical replications of the same experiment for just a single unit (see Holland 1986:947).

As noted above (see section 2.6), this variable is called a "random variable" since it has nonsystematic features: it is affected by explanatory variables not encompassed in our theoretical analysis or contains fundamentally unexplainable variability.[7] We define the random variable representing the proportion of votes received by the incumbent Democratic candidate as Y_4^I (note the capital Y) and the proportion of votes that would be received in hypothetical replications by a Democratic nonincumbent as Y_4^N.

We now define the *random causal effect* for district 4 as the difference between these two random variables. Since we wish to retain some generality, we again switch notation from district 4 to unit *i*:

$$\text{(Random Causal Effect for unit } i) = (Y_i^I - Y_i^N) \tag{3.2}$$

(Just as in the definition of a random variable, a random causal effect is a causal effect that varies over hypothetical replications of the same experiment but also represents many interesting systematic features of elections.) *If* we could observe two separate vote proportions in district 4 at the same time, one from an election with and one without a Democratic incumbent running, then we could directly observe the realized causal effect in equation (3.1). Of course, because of the Fundamental Problem of Causal Inference, we cannot observe the realized causal effect. Thus, the realized causal effect in equation 3.1 is a single *unobserved* realization of the random causal effect in equation 3.2. In other words, across many hypothetical replications of the same election in district 4 with a Democratic incumbent, and across many hypothetical replications of the same election but with a Democratic nonincumbent, the (unobserved) realized causal effect becomes a random causal effect.

Describing causality as one of the systematic features of random variables may seem unduly complicated. But it has two virtues. First, it makes our definition of causality directly analogous to those systematic features (such as a mean or variance) of a phenomenon that serve

[7] As we explained in more detail in section 2.2, this phrasing can be confusing. A "random variable" contains some systematic component and thus is not always entirely unpredictable. Unfortunately, this language has a specific meaning in statistics and the concepts underlying it are important. The original reason for the terminology is that randomness does not mean "anything goes" or "anything could happen." Instead, it refers to one of many possible very well-specified probabilistic processes. For example, the random process governing which side of a coin lands upward when flipped in the air is a very different random process than the one governing the growth of the European Economic Community's bureaucracy or the uncertain political consequence of a change in Italy's electoral system. The key to our representation is that each of these "random" processes have systematic and probabilistic components.

as objects of descriptive inference: means and variances are also systematic features of random variables (as in section 2.2). Secondly, it enables us to partition a causal inference problem into systematic and nonsystematic components. Although many systematic features of a random variable might be of interest, the most relevant for our running example is the *mean causal effect* for unit *i*. To explain what we mean by this, we return to our New York election example.

Recall that the random variable refers to the vote fraction received by the Democrat (incumbent or nonincumbent) across a large number of hypothetical replications of the same election. We define the expected value of this random variable—the vote fraction averaged across these replications—for the nonincumbent as

$$E(Y_4^N) = \mu_4^N$$

and for the incumbent as

$$E(Y_4^I) = \mu_4^I.$$

Then, the mean causal effect of incumbency in unit *i* is a systematic feature of the random causal effect and is defined as the difference between these two expected values (again generalized to unit *i* instead of to district 4):

$$\begin{aligned}\text{Mean Causal Effect for unit } i \equiv \beta & \qquad (3.3)\\ &= E(\text{Random Causal Effect for unit } i)\\ &= E(Y_i^I - Y_i^N)\\ &= E(Y_i^I) - E(Y_i^N)\\ &= \mu_i^I - \mu_i^N\end{aligned}$$

where in the first line of this equation, β (beta) refers to this mean causal effect. In the second line, we indicate that the mean causal effect for unit *i* is just the mean (expected value) of the random causal effect, and in the third and fourth lines we show how to calculate the mean. The last line is another way of writing the difference in the means of the two sets of hypothetical elections. (The average of the difference between two random variables equals the difference of the averages.) To summarize in words: *the causal effect is the difference between the systematic component of observations made when the explanatory variable takes* *one value and the systematic component of comparable observations when the explanatory variable takes on another value.*

The last line of equation 3.3 is similar to equation 3.1, and as such, the Fundamental Problem of Causal Inference still exists in this formulation. Indeed, the problem expressed this way is even more formidable because even if we could get around the Fundamental Problem for a realized causal effect, we would still have all the usual problems of inference, including the problem of separating out systematic and nonsystematic components of the random causal effect. From here on, we use Holland's phrase, the Fundamental Problem of Causal Inference, to refer to the problem that he identified *as well as* to these standard problems of inference, which we have added to his formulation. In the box on page 95, we provide a more general notation for causal effects, which will prove useful throughout the rest of this book.

Many other systematic features of these random causal effects might be of interest in various circumstances. For example, we might wish to know the variance in the possible (realized) causal effects of incumbency status on Democratic vote in unit *i*, just as with the variance in the vote itself that we described in equation 2.3 in section 2.6. To calculate the variance of the causal effect, we apply the variance operation

$$\text{(variance of the causal effect in unit } i) = V(Y_i^I - Y_i^N)$$

in which we avoid introducing a new symbol for the result of the variance calculation, $V(Y_i^I - Y_i^N)$. Certainly new incumbents would wish to know the variation in the causal effect of incumbency so they can judge how closely their experience will be to that of previous incumbents and how much to rely on their estimated mean causal effect of incumbency from previous elections. It is especially important to understand that this variance in the causal effect is a fundamental part of the world and is not uncertainty due to estimation.

3.1.2 A Qualitative Example

We developed our precise definition of causality in section 3.1. Since some of the concepts in that section are subtle and quite sophisticated, we illustrated our points with a very simple running example from quantitative research. This example helped us communicate the concepts we wished to stress without also having to attend to the contextual detail and cultural sensitivity that characterize good qualitative research. In this section, we proceed through our definition of causality again, but this time via a qualitative example.

Political scientists would learn a lot if they could rerun history with everything constant save for one investigator-controlled explanatory variable. For example, one of the major questions that faces those involved with politics and government has to do with the consequences of a particular law or regulation. Congress passes a tax bill that is intended to have a particular consequence—lead to particular investments, increase revenue by a certain amount, and change consumption patterns. Does it have this effect? We can observe what happens after the tax is passed to see if the intended consequences appear; but even if they do, it is never certain that they *result* from the law. The change in investment policy might have happened anyway. If we could rerun history with and without the new regulation, then we would have much more leverage in estimating the causal effect of this law. Of course, we cannot do this. But the logic will help us design research to give us an approximate answer to our question.

Consider now the following extended example from comparative politics. In the wake of the collapse of the Soviet system, numerous governments in the ex-Soviet republics and in Eastern Europe have instituted new governmental forms. They are engaged—as they themselves realize—in a great political experiment: they are introducing new constitutions, constitutions that they hope will have the intended effect of creating stable democratic systems. One of the constitutional choices is between parliamentary and presidential forms of government. Which system is more likely to lead to a stable democracy is the subject of considerable debate among scholars in the field (Linz 1993; Horowitz 1993; Lijphart 1993). The debate is complex, not the least because of the numerous types of parliamentary and presidential systems and the variety of the other constitutional provisions that might accompany and interact with this choice (such as the nature of the electoral system). It is not our purpose to provide a thorough analysis of these choices but rather a greatly simplified version of the choice in order to define a causal effect in the context of this qualitative example. In so doing, we highlight the distinction between systematic and nonsystematic features of a causal effect.

The debate about presidential versus parliamentary systems involves varied features of the two systems. We will focus on two: the extent to which each system represents the varied interests of the citizenry and encourages strong and decisive leadership. The argument is that parliamentary systems do a better job of representing the full range of societal groups and interests in the government since there are many legislative seats to be filled, and they can be filled by representatives elected from various groups. In contrast, the all-or-nothing character of presidential systems means that some groups will feel left out of the government, be disaffected, and cause greater instability. On the other hand, parliamentary systems—especially if they adequately represent the full range of social groups and interests—are likely to be deadlocked and ineffective in providing decisive government. These characteristics, too, can lead to disaffection and instability.[8]

The key purpose of this section is to formulate a precise definition of a causal effect. To do so, imagine that we could institute a parliamentary system and, periodically over the next decade or so, measure the degree of democratic stability (perhaps by actual survival or demise of democracy, attempted coups, or other indicators of instability), and in the same country and at the same time, institute a presidential system, also measuring its stability over the same period with the same measures. The *realized causal effect* would be the difference between the degree of stability observed under a presidential system and that under a parliamentary system. The impossibility of measuring this causal effect directly is another example of the fundamental problem of causal inference.

As part of this definition, we also need to distinguish between systematic and nonsystematic effects of the form of government. To do this, we imagine running this hypothetical experiment many times. We define the *mean causal effect* to be the average of the realized causal effects across replications of these experiments. Taking the average in this way causes the nonsystematic features of this problem to cancel out and leaves the mean causal effect to include only systematic features. Systematic features include indecisiveness in a parliamentary system or disaffection among minorities in a presidential one. Nonsystematic features might include the sudden illness of a president that throws the government into chaos. The latter event would not be a persistent feature of a presidential system; it would appear in one trial of the experiment but not in others.[9]

Another interesting feature of this example is the variance of the causal effect. Any country thinking of choosing one of these political systems would be interested in its mean causal effect on democratic stability; however, this one country gets only one chance—only one replication of this experiment. Given this situation, political leaders may be interested in more than the average causal effect. They may wish to understand what the maximum and minimum causal effects, or at least the *variance* of the causal effects, might be. For example, it may be that presidentialism reduces democratic stability on average

[8] These distinctions are themselves debated. Some argue that a presidential system can do a better representational job. And others argue that parliamentary systems can be more decisive.

[9] The distinction between a systematic and nonsystematic feature is by no means always clear-cut. The sudden illness of a president appears to be a nonsystematic feature of the presidential system. On the other hand, the general vulnerability of presidential systems to the vagaries of the health and personality of a single individual is a systematic effect that raises the likelihood that *some* nonsystematic feature will appear.

but that the variability of this effect is enormous—sometimes increasing stability a lot, sometimes decreasing it substantially. This variance translates into risk for a polity. In this circumstance, it may be that citizens and political leaders would prefer to choose an option that produces only slightly less stability on average but has a lower variance in causal effect and thus minimizes the chance of a disastrous outcome.

3.2 Clarifying Alternative Definitions of Causality

In section 3.1, we defined causality in terms of a causal effect: the mean causal effect is the difference between the systematic component of a dependent variable when the causal variable takes on two different values. In this section, we use our definition of causality to clarify several alternative proposals and apparently complicating ideas. We show that the important points made by other authors about "causal mechanisms" (section 3.2.1), "multiple" causality (section 3.2.2), and "symmetric" versus "asymmetric" causality (section 3.2.3) do not conflict with our more basic definition of causality.

3.2.1 "Causal Mechanisms"

Some scholars argue that the central idea of causality is that of a set of "causal mechanisms" posited to exist between cause and effect (see Little 1991:15). This view makes intuitive sense: any coherent account of causality needs to specify how the effects are exerted. For example, suppose a researcher is interested in the effect of a new bilateral tax treaty on reducing the United States's current account deficit with Japan. According to our definition of causality, the causal effect here is the reduction in the expected current account deficit with the tax treaty in effect as compared to the same situation (at the same time and for the same countries) with the exception that the treaty was not in effect. The causal mechanism operating here would include, in turn, the signing and ratification of the tax treaty, newspaper reports of the event, meetings of the relevant actors within major multinational companies, compensatory actions to reduce their total international tax burden (such as changing its transfer pricing rules or moving manufacturing plants between countries), further actions by other companies and workers to take advantage of the movements of capital and labor between countries, and so on, until we reach the final effect on the balance of payments between the United States and Japan.

From the standpoint of processes through which causality operates, an emphasis on causal mechanisms makes intuitive sense: any coherent account of causality needs to specify how its effects are exerted. Identifying causal mechanisms is a popular way of doing empirical analyses. It has been called, in slightly different forms, "process tracing" (which we discuss in section 6.3.3), "historical analysis," and "detailed case studies." Many of the details of well-done case studies involve identifying these causal mechanisms.

However, identifying the causal mechanisms requires causal inference, using the methods discussed below. That is, to demonstrate the causal status of each potential linkage in such a posited mechanism, the investigator would have to define and then estimate the causal effect underlying it. To portray an internally consistent causal mechanism requires using our more fundamental definition of causality offered in section 3.1 for each link in the chain of causal events.

Hence our definition of causality is logically prior to the identification of causal mechanisms. Furthermore, there always exists in the social sciences an infinity of causal steps between any two links in the chain of causal mechanisms. If we posit that an explanatory variable causes a dependent variable, a "causal mechanisms" approach would require us to identify a list of causal links between the two variables. This definition would also require us to identify a series of causal linkages, to define causality for each pair of consecutive variables in the sequence, and to identify the linkages between any two of these variables and the connections between each pair of variables. This approach quickly leads to infinite regress, and at no time does it alone give a precise definition of causality for any one cause and one effect.

In our example of the effect of a presidential versus parliamentary system on democratic stability (section 3.1.2), the hypothesized causal mechanisms include greater minority disaffection under a presidential regime and lesser governmental decisiveness under a parliamentary regime. These intervening effects—caused by the constitutional system and, in turn, affecting political stability—can be directly observed. We could monitor the attitudes or behaviors of minorities to see how they differ under the two experimental conditions or study the decisiveness of the governments under each system. Yet even if the causal effect of presidential versus parliamentary systems could operate in different ways, our definition of the causal effect would remain valid. We can define a causal effect without understanding all the causal mechanisms involved, but we cannot identify causal mechanisms without defining the concept of causal effect.

In our view, identifying the mechanisms by which a cause has its effect often builds support for a theory and is a very useful operational procedure. Identifying causal mechanisms can sometimes give us more leverage over a theory by making observations at a different level of analysis into implications of the theory. The concept can also create new causal hypotheses to investigate. However, we should not confuse a definition of causality with the nondefinitional, albeit often useful, operational procedure of identifying causal mechanisms.

3.2.2 "Multiple Causality"

Charles Ragin, in a recent work (1987:34–52), argues for a methodology with many explanatory variables and few observations in order that one can take into account what he calls "multiple causation." That is, "The phenomenon under investigation has alternative determinants—what Mill (1843) referred to as the problem of 'plurality of causes.'" This is the problem referred to as "equifinality" in general systems theory (George 1982:11). In situations of multiple causation, these authors argue that the same outcome can be caused by combinations of different independent variables.[10]

Under conditions in which different explanatory variables can account for the same outcome on a dependent variable, according to Ragin, some statistical methods will falsely reject the hypothesis that these variables have causal status. Ragin is correct that some statistical models (or relevant qualitative research designs) could fail to alert an investigator to the existence of "multiple causality," but appropriate statistical models can easily handle situations like these (some of which Ragin discusses).

Moreover, the fundamental features of "multiple causality" are compatible with our definition of causality. They are also no different for quantitative than qualitative research. The idea contains no new features or theoretical requirements. For example, consider the hypothesis that a person's level of income depends *both* on high educational attainment *and* highly educated parents. Having one but not both is insufficient. In this case, we need to compare categories of our causal variable: respondents who have high educational attainment and highly educated parents, the two groups who have one but not the other, and the group with neither. Thus, the concept of "multiple causation" puts greater demands on our data since we now have four cat-

[10] This idea is often explained in terms of no explanatory variable being either necessary or sufficient for a particular value of a dependent variable to occur. However, this is misleading terminology because the distinction between necessary and sufficient conditions largely disappears when we allow for the possibility that causes are probabilistic. As Little (1991:27) explains, "Consider the claim that poor communication among superpowers during crisis increases the likelihood of war. This is a probabilistic claim; it identifies a causal variable (poor communication) and asserts that this variable increases the probability of a given outcome (war). It cannot be translated into a claim about the necessary and sufficient conditions for war, however; it is irreducibly probabilistic."

egories of our causal variables, but it does not require a modification of our definition of causality. For our definition, we would need to measure the expected income for the same person, at the same time, experiencing each of the four conditions.

But what happens if different causal explanations generate the same values of the dependent variable? For example, suppose we consider whether or not one graduated from college as our (dichotomous) causal variable in a population of factory workers. In this situation, both groups could quite reasonably earn the same income (our dependent variable). One reason might be that this explanatory variable (college attendance) has no causal effect on income among factory workers, perhaps because a college education does not help one perform better. Alternatively, different explanations might lead to the same level of income for those educated and those not educated. College graduates might earn a particular level of income because of their education, whereas those who had no college education might earn the same level of income because of their four years of additional seniority on the job. In this situation wouldn't we be led to conclude that "college education" has no causal effect on income levels for those who will become factory workers?

Fortunately, our definition of causality requires that we more carefully specify the counterfactual condition. In the present example, the values of the key causal variable to be varied are (1) college education, as compared to (2) no college education but four additional years of job seniority. The dependent variable is starting annual income. Our causal effect is then defined as follows: we record the income of a person graduating from college who goes to work in a factory. Then, we go back in time four years, put this same person to work in the same factory instead of in college and, at the end of four years, measure his or her income "again." The expected difference between these two levels of income for this one individual is our definition of the mean causal effect. In the present situation, we have imagined that this causal effect is zero. But this does not mean that "college education has no effect on income," only that the average difference between treatment groups (1) and (2) is zero. In fact, there is no logically unique definition of "the causal effect of college education" since one cannot define a causal effect without at least two conditions. The conditions need not be the two listed here, but they must be very clearly identified.

An alternative pair of causal conditions is to compare a college graduate with someone without a college degree but with the same level of job seniority as the college graduate. In one sense, this is unrealistic, since the non-college graduate would have to do something for the four years while not attending college, but perhaps we would be willing to imagine that this person had a different, irrelevant job for those four years. Put differently, this alternative counterfactual is the effect of a college education compared to that of none, with job seniority held constant. Failure to hold seniority constant in the two causal conditions would cause any research design to yield estimates of our first counterfactual instead of this revised one. If the latter were the goal, but no controls were introduced, our empirical analysis would be flawed due to "omitted variable bias" (which we introduce in section 5.2).

Thus, the issues addressed under the label "multiple causation" do not confound our definition of causality although they may make greater demands in our subsequent analyses. The fact that some dependent variables, and perhaps all interesting social science–dependent variables, are influenced by many causal factors does not make our definition of causality problematic. The key to understanding these very common situations is to define the counterfactual conditions making up each causal effect very precisely. We demonstrate in chapter 5 that researchers need not identify "all" causal effects on a dependent variable to provide estimates of the one causal effect of interest (even if that were possible). A researcher can focus on only the one effect of interest, establish firm conclusions, and then move on to others that may be of interest (see sections 5.2 and 5.3).[11]

3.2.3 "Symmetric" and "Asymmetric" Causality

Stanley Lieberson (1985:63–64) distinguishes between what he refers to as "symmetrical" and "asymmetrical" forms of causality. He is interested in causal effects which differ when an explanatory variable is increased as compared to when it is decreased. In his words,

> In examining the causal influence of X_1 [an explanatory variable] on Y [a dependent variable], for example, one has also to consider whether shifts to a given value of X_1 from either direction have the same consequences for Y. . . . If the causal relationship between X_1 [an explanatory variable] and Y

[11] Our emphasis on distinguishing systematic from nonsystematic components of observations subject to causal inference reflects our general view that the world, at least as we know it, is probabilistic rather than deterministic. Hence, we also disagree with Ragin's premise (1987:15) that "explanations which result from applications of the comparative method are not conceived in probabilistic terms because every instance of a phenomenon is examined and accounted for if possible." Even if it were possible to collect a census of information on every instance of a phenomenon and every permutation and combination of values of the explanatory variables, the world still would have produced these data according to some probabilistic process (as defined in section 2.6). This

> [a dependent variable] is symmetrical or truly reversible, then the effect on Y of an increase in X_1 will disappear if X_1 shifts back to its earlier level (assuming that all other conditions are constant).

As an example of Lieberson's point, imagine that the Fourth Congressional District in New York had no incumbent in 1998 and that the Democratic candidate received 55 percent of the vote. Lieberson would define the causal effect of incumbency as the increase in the vote if the winning Democrat in 1998 runs as an incumbent in the next election in the year 2000. This effect would be "symmetric" if the absence of an incumbent in the subsequent election (in year 2002) caused the vote to return to 55 percent. The effect might be "asymmetric" if, for example, the incumbent Democrat raised money and improved the Democratic party's campaign organization; as a result, if no incumbent were running in 2002, the Democratic candidate might receive more than 55 percent of the vote.

Lieberson's argument is clever and very important. However, in our view, his argument does not constitute a *definition* of causality, but applies only to some causal *inferences*—the process of learning about a causal effect from existing observations. In section 3.1, we defined causality for a single unit. In the present example, a causal effect can be defined theoretically on the basis of hypothetical events occurring only in the 1998 election in the Fourth District in New York. Our definition is the difference in the systematic component of the vote in this district with an incumbent in this election and without an incumbent in the same election, time, and district.

In contrast, Lieberson's example involves no hypothetical quantities and therefore cannot be a causal definition. This example involves only what would actually occur if the explanatory variable changed in two real elections from nonincumbent to incumbent, versus incumbent to nonincumbent in two other elections. Any empirical analysis of this example would involve numerous problems of inference. We discuss many of these problems of causal inference in chapters 4–6. In the present example, we might ask whether the estimated effect seemed larger only because we failed to account for a large number of recently registered citizens in the Fourth District. Or, did the surge in support for the Democrat in the election in which she or he was an incumbent

seems to invalidate Ragin's "Boolean Algebra" approach as a general way of designing theoretical explanations or making inferences; to learn from data requires the same logic of scientific inference that we discuss in this book. However, his approach can still be valuable as a form of formal theory (see section 3.5.2): it enables the investigator to specify a theory and its implications in a way that might be much more difficult without it.

seem smaller than it should because we necessarily discarded districts where the Democrat lost the first election?

Thus, Lieberson's concepts of "symmetrical" and "asymmetrical" causality are important to consider in the context of causal inference. However, they should not be confused with a theoretical definition of causality, which we give in section 3.1.

3.3 Assumptions Required for Estimating Causal Effects

How do we avoid the Fundamental Problem of Causal Inference and also the problem of separating systematic from nonsystematic components? The full answer to this question will consume chapters 4–6, but we provide an overview here of what is required in terms of the two possible assumptions that enable us to get around the fundamental problem. These are *unit homogeneity* (which we discuss in section 3.3.1) and *conditional independence* (section 3.3.2). These assumptions, like any other attempt to circumvent the Fundamental Problem of Causal Inference, always involve some untestable assumptions. It is the responsibility of all researchers to make the substantive implications of this weak spot in their research designs extremely clear and visible to readers. Causal inferences should not appear like magic. The assumptions can and should be justified with whatever side information or prior research can be mustered, but it always must be explicitly recognized.

3.3.1 Unit Homogeneity

If we cannot rerun history at the same time and the same place with different values of our explanatory variable each time—as a true solution to the Fundamental Problem of Causal Inference would require—we can attempt to make a second-best assumption: we can rerun our experiment in two different units that are "homogeneous." *Two units are homogeneous when the expected values of the dependent variables from each unit are the same when our explanatory variable takes on a particular value.* (That is, $\mu_1^N = \mu_2^N$ and $\mu_1^I = \mu_2^I$.) For example, if we observe $X = 1$ (an incumbent) in district 1 and $X = 0$ (no incumbent) in district 2, an assumption of unit homogeneity means that we can use the observed proportions of the vote in two separate districts for inference about the causal effect β, which we assume is the same in both districts. For a data set with n observations, unit homogeneity is the assumption that all units with the same value of the explanatory variables have the same expected value of the dependent variable. Of course, this is only an assumption and it can be wrong: the two districts might differ in some unknown way that would bias our causal inference. Indeed, any two real districts *will* differ in some ways; application of this assumption requires that these districts must be the same on average over many hypothetical replications of the election campaign. For example, patterns of rain (which might inhibit voter turnout in some areas) would not differ across districts on average unless there were systematic climatic differences between the two areas.

In the following quotation, Holland (1986:947) provides a clear example of the unit homogeneity assumption (defined from his perspective of a realized causal effect instead of the mean causal effect). Since very little randomness exists in the experiment in the following example, his definition and ours are close. (Indeed, as we show in section 4.2, with a small number of units, the assumption of unit homogeneity is most useful when the amount of randomness is fairly low.)

> If [the unit] is a room in a house, *t* [for '*t*reatment'] means that I flick the light switch in that room, *c* [for '*c*ontrol'] means that I do not, and [the dependent variable] indicates whether the light is on or not a short time after applying either *t* or *c*, then I might be inclined to *believe* that I can *know* the values of [the dependent variable for both *t* and *c*] by simply flicking the switch. It is clear, however, that it is only because of the plausibility of certain assumptions about the situation that this *belief* of mine can be shared by anyone else. If, for example, the light has been flicking off and on for no apparent reason while I am contemplating beginning this experiment, I might doubt that I would know the values of [the dependent variable for both *t* and *c*] after flicking the switch—at least until I was clever enough to figure out a new experiment!

In this example, the unit homogeneity assumption is that if we had flicked the switch (in Holland's notation, applied *t*) in both periods, the expected value (of whether the light will be on) would be the same. Unit homogeneity also assumes that if we had not flicked the switch (applied *c*) in both periods, the expected value would be the same, although not necessarily the same as when *t* is applied. Note that we would have to reset the switch to the off position after the first experiment to assure this, but we would also have to make the untestable assumption that flipping the switch on in the first period does not effect the two hypothetical expected values in the next period (such as if a fuse were blown after the first flip). In general, the unit homogeneity assumption is untestable for a single unit (although, in this case, we might be able to generate several new hypotheses about the causal mechanism by ripping the wall apart and inspecting the wiring).

A weaker, but also fully acceptable, version of unit homogeneity is the *constant effect* assumption. Instead of assuming that the expected value of the dependent variable is the same for different units with the same value of the explanatory variable, we need only to assume that the causal effect is constant. This is a weaker version of the unit homogeneity assumption, since the causal effect is only the difference between the two expected values. If the two expected values for units with the same value of the explanatory variable vary in the same way, the unit homogeneity assumption would be violated, but the constant effect assumption would still be valid. For example, two congressional districts could vary in the expected proportion of the vote for Democratic nonincumbents (say 45 percent vs. 65 percent), but incumbency could still add an additional ten percent to the vote of a Democratic candidate of either district.

The notion of unit homogeneity (or the less demanding assumption of constant causal effects) lies at the base of all scientific research. It is, for instance, the assumption underlying the method of comparative case studies. We compare several units that have varying values on our explanatory variables and observe the values of the dependent variables. We believe that the differences we observe in the values of the dependent variables are the result of the differences in the values of the explanatory variables that apply to the observations. What we have shown here is that our "belief" in this case necessarily relies upon an assumption of unit homogeneity or constant effects.

Note that we may seek homogeneous units across time or across space. We can compare the vote for the Democratic candidate when there is a Democratic incumbent running with the vote when there is no Democratic incumbent in the same district at different times or across different districts at the same time (or some combination of the two). Since a causal effect can only be estimated instead of known, we should not be surprised that the unit homogeneity assumption is generally untestable. But it is important that the nature of the assumption is made explicit. Across what range of units do we expect our assumption of a uniform incumbency effect to hold? All races for Congress? Congressional but not Senate races? Races in the North only? Races in the past two decades only?

Notice how the unit homogeneity assumption relates to our discussion in section 1.1.3 on complexity and "uniqueness." There we argued that social science generalization depends on our ability to simplify reality coherently. At the limit, simplifying reality for the purpose of making causal inferences implies meeting the standards for unit homogeneity: the observations being analyzed become, for the purposes of analysis, identical in relevant respects. Attaining unit homogeneity is often impossible; congressional elections, not to speak of revolutions, are hardly close analogies to light switches. But understanding the degree of heterogeneity in our units of analysis will help us to estimate the degree of uncertainty or likely biases to be attributed to our inferences.

3.3.2 Conditional Independence

Conditional independence is the assumption that values are assigned to explanatory variables independently of the values taken by the dependent variables. (The term is sometimes used in statistics, but it does not have the same definition as it commonly does in probability theory.) That is, after taking into account the explanatory variables (or controlling for them), the process of assigning values to the explanatory variable is independent of both (or, in general two or more) dependent variables, Y_i^N and Y_i^I. We use the term "assigning values" to the explanatory variables to describe the process by which these variables obtain the particular values they have. In experimental work, the researcher actually *assigns* values to the explanatory variables; some subjects are assigned to the treatment group and others to the control group. In nonexperimental work, the values that explanatory variables take may be "assigned" by nature or the environment. What is crucial in these cases is that the values of the explanatory variables are not caused by the dependent variables. The problem of "endogeneity" that exists when the explanatory variables are caused, at least in part, by the dependent variables is described in section 5.4.

Large-*n* analyses that involve the procedures of random selection and assignment constitute the most reliable way to assure conditional independence and do not require the unit homogeneity assumption. Random selection and assignment help us to make causal inferences because they *automatically* satisfy three assumptions that underlie the concept of conditional independence: (1) that the process of assigning values to the explanatory variables is independent of the dependent variables (that is, there is no endogeneity problem); (2) that selection bias, which we discuss in section 4.3, is absent; and (3) that omitted variable bias (section 5.2) is also absent. Thus, if we are able to meet these conditions in any way, either through random selection and assignment (as discussed in section 4.2) or through some other procedure, we can avoid the Fundamental Problem of Causal Inference.

Fortunately, random selection and assignment are *not* required to meet the conditional independence assumption. If the process by which the values of the explanatory variables are "assigned" is not independent of the dependent variables, we can still meet the conditional independence assumption if we learn about this process and include a measure of it among our control variables. For example, suppose we are interested in estimating the effect of the degree of residential segregation on the extent of conflict between Israelis and Palestinians in communities on the Israeli-occupied West Bank. Our conditional independence assumption would be severely violated if we looked only at the association between these two variables to find the causal effect. The reason is that the Israelis and Palestinians who choose to live in segregated neighborhoods may do so out of an ideological belief about who ultimately has rights to the West Bank. Ideological extremism (on both sides) may therefore lead to conflict. A measure that we believe to be residential segregation might really be a surrogate for ideology. The difference between the two explanations may be quite important, since a new housing policy might help remedy the conflict if residential segregation were the real cause, whereas this policy would be ineffective or even counterproductive if ideology were really the driving force. We might correct for the problem here by also measuring the ideology of the residents explicitly and controlling for it. For example, we could learn how popular extremist political parties are among the Israelis and PLO affiliation is among the Palestinians. We could then control for the possibly confounding effects of ideology by comparing communities with the same level of ideological extremism but differing levels of residential segregation.

When random selection and assignment are infeasible and we cannot control for the process of assignment and selection, we have to resort to some version of the unit homogeneity assumption in order to make valid causal inferences. Since that assumption will be only imperfectly met in social science research, we will have to be especially careful to specify our degree of uncertainty about causal inferences. This assumption will be particularly apparent when we discuss the procedures used in "matching" observations in section 5.6.

Notation for a Formal Model of a Causal Effect. We now generalize our notation for the convenience of later sections. In general, we will have n realizations of a random variable Y_i. In our running quantitative example, n is the number of congressional districts (435), and the realization y_i of the random variable Y_i is the observed Democratic proportion of the two-party vote in district i (such as 0.56). The expected nonincumbent Democratic proportion of the two-party vote (the average over all hypothetical replications) in district i is μ_i^N. We define the explanatory variable as X_i, which is coded in the present example as zero when district i has no Democratic incumbent and as one when district i has a Democratic incumbent. Then, we can denote the mean causal effect in unit i as

$$\beta = E(Y_i|X_i = 1) - E(Y_i|X_i = 0) = \mu_i^I - \mu_i^N \tag{3.4}$$

and incorporate it into the following simple formal model:

$$\begin{aligned} E(Y_\mathrm{i}) &= \mu_i^N + X_i(\mu_i^I - \mu_i^N) \\ &= \mu_i^N + X_i\beta \end{aligned} \tag{3.5}$$

Thus, when district i has no incumbent, and $X_i = 0$, the expected value is determined by substituting zero into equation (3.5) for X_i, and the answer is as before:

$$\begin{aligned} E(Y_i|X = 0) &= \mu_i^N + (0)\beta \\ &= \mu_i^N \end{aligned}$$

Similarly, when a Democratic incumbent is running in district i, the expected value is μ_i^I:

$$\begin{aligned} E(Y_i|X = 1) &= \mu_i^N + (1)\beta \\ &= \mu_i^N + \beta \\ &= \mu_i^N + (\mu_i^I - \mu_i^N) \\ &= \mu_i^I \end{aligned}$$

Thus, equation (3.5) provides a useful model of causal inference, and β—the difference between the two theoretical proportions—is our causal effect. Finally, for future reference, we simplify equation (3.5) one last time. If we assume that Y_i has a zero mean (or is written as a deviation from its mean, which does not limit the applicability of the model in any way), then we can drop the intercept from this equation, and write it more simply as

$$E(Y_i) = X_i\beta \tag{3.6}$$

The parameter β is still the theoretical value of the mean causal effect, a systematic feature of the random variables, and one of our goals in causal inference. This model is a special case of "regression analysis," which is common in quantitative research, but regression coefficients are only sometimes coincident with estimates of causal effects.

3.4 Criteria for Judging Causal Inferences

Recall that by defining causality in terms of random variables, we were able to draw a strict analogy between it and other systematic features of phenomena, such as a mean or a variance, on which we focus in making descriptive inferences. This analogy enables us to use precisely the same criteria to judge causal inferences as we used to judge descriptive inferences in section 2.7: *unbiasedness* and *efficiency*. Hence, most of what we said on this subject in Chapter 2 applies equally well to the causal inference problems we deal with here. In this section, we briefly formalize the relatively few differences between these two situations.

In section 2.7 the object of our inference was a mean (the expected value of a random variable), which we designate as μ. We conceptualize μ as a fixed, but unknown, number. An estimator of μ is said to be unbiased if it equals μ on average over many hypothetical replications of the same experiment.

As above, we continue to conceptualize the expected value of a random causal effect, denoted as β, as a fixed, but unknown, number. The unbiasedness is then defined analogously: an estimator of β is unbiased if it equals β on average over many hypothetical replications of the same experiment. Efficiency is also defined analogously as the variability across these hypothetical replications. These are very important concepts that will serve as the basis for our studies of many of the problems of causal inference in chapters 4–6. The two boxes that follow provide formal definitions.

A Formal Analysis of Unbiasedness of Causal Estimates. In this box, we demonstrate the unbiasedness of the estimator of the causal effect parameter from section 3.1. The notation and logic of these ideas closely parallel those from the formal definition of unbiasedness in the context of descriptive inference in section 2.7. The simple linear model with one explanatory and one dependent variable is as follows:[12]

[12] In order to avoid using a constant term, we assume that all variables have zero mean. This simplifies the presentation but does not limit our conclusions in any way.

$$E(Y_i) = \beta X_i$$

Our estimate of β is simply the least squares regression estimate:

$$b = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2} \tag{3.7}$$

To determine whether b is an unbiased estimator of β, we need to take the expected value, averaging over hypothetical replications:

$$\begin{aligned} E(b) &= E\left(\frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}\right) \\ &= \frac{\sum_{i=1}^{n} X_i E(Y_i)}{\sum_{i=1}^{n} X_i^2} \\ &= \frac{\sum_{i=1}^{n} X_i^2 \beta}{\sum_{i=1}^{n} X_i^2} \\ &= \beta \end{aligned} \tag{3.8}$$

which proves that b is an unbiased estimator of β.

A Formal Analysis of Efficiency. Here, we assess the efficiency of the standard estimator of the causal effect parameter β from section 3.1. We proved in equation (3.8) that this estimator is unbiased and now calculate its variance:

$$\begin{aligned} V(b) &= V\left(\frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}\right) \\ &= \frac{1}{\left(\sum_{i=1}^{n} X_i^2\right)^2} \sum_{i=1}^{n} X_i^2 V(Y_i) \\ &= \frac{V(Y_i)}{\sum_{i=1}^{n} X_i^2} \end{aligned} \tag{3.9}$$

$$= \frac{\sigma^2}{\sum_{i=1}^{n} X_i^2}$$

Thus, the variance of this estimator is a function of two components. First, the more random *each* unit in our data (the larger is σ^2) is, the more variable will be our estimator b; this should be no surprise. In addition, the larger the observed variance in the explanatory variable ($\sum_{i=1}^{n} X_i^2$), the less variable will be our estimate of b. In the extreme case of no variability in X, nothing can help us estimate the effect of changes in the explanatory variable on the dependent variable, and the formula predicts an infinite variance (complete uncertainty) in this instance. More generally, this component indicates that efficiency is greatest when we have evidence from a larger range of values of the explanatory variable. In general, then, it is best to evaluate our causal hypotheses in as many diverse situations as possible. One way to think of this latter point is to think about drawing a line with a ruler, two dots on a page, and a shaky hand. If the two dots are very close together (small variance of X), errors in the placement of the ruler will be much larger than if the dots are farther apart (the situation of a large variance in X).

3.5 Rules for Constructing Causal Theories

Much sensible advice about improving qualitative research is precise, specific, and detailed; it involves a manageable and therefore narrow aspect of qualitative research. However, even in the midst of solving a host of individual problems, we must keep the big picture firmly in mind: each specific solution must help in solving whatever is the general causal inference problem one aims to solve. Thus far in this chapter, we have provided a precise theoretical definition of a causal effect and discussed some of the issues involved in making causal inferences. We take a step back now and provide a broader overview of some rules regarding theory construction. As we discuss (and have discussed in section 1.2), improving theory does not end when data collection begins.

Causal theories are designed to show the causes of a phenomenon or set of phenomena. Whether originally conceived as deductive or inductive, any theory includes an interrelated set of causal hypotheses. Each hypothesis specifies a posited relationship between variables that creates observable implications: if the specified explanatory variables take on certain values, other specified values are predicted for the dependent variables. Testing or evaluating any causal hypothesis requires causal inference. The overall theory, of which the hypotheses are parts should be *internally consistent*, or else hypotheses can be generated that contradict one another.

Theories and hypotheses that fit these definitions have an enormous range. In this section, we provide five rules that will help in formulating good theories, and we provide a discussion of each with examples.

3.5.1 Rule 1: Construct Falsifiable Theories

By this first rule, we do not only mean that a "theory" incapable of being wrong is not a theory. We also mean that we should design theories so that they can be shown to be wrong as easily and quickly as possible. Obviously, we should not actually try to be wrong, but even an incorrect theory is better than a statement that is neither wrong nor right. The emphasis on falsifiable theories forces us to keep the right perspective on uncertainty and guarantees that we treat theories as tentative and not let them become dogma. We should always be prepared to reject theories in the face of sufficient scientific evidence against them. One question that should be asked about any theory (or of any hypothesis derived from the theory) is simply: what evidence would falsify it? The question should be asked of all theories and hypotheses but, above all, the researcher who poses the theory in the first place should ask it of his or her own.

Karl Popper is most closely identified with the idea of falsifiability (Popper 1968). In Popper's view, a fundamental asymmetry exists between confirming a theory (verification) and disconfirming it (falsification). The former is almost irrelevant, whereas the latter is the key to science. Popper believes that a theory once stated immediately becomes part of the body of accepted scientific knowlege. Since theories are general, and hypotheses specific, theories technically imply an infinite number of hypotheses. However, empirical tests can only be conducted on a finite number of hypotheses. In that sense, "theories are not verifiable" because we can never test all observable implications of a theory (Popper 1968:252). Each hypothesis tested may be shown to be consistent with the theory, but any number of consistent empirical results will not change our opinions since the theory remains accepted scientific knowledge. On the other hand, if even a single hypothesis is shown to be wrong, and thus inconsistent with the theory, the theory is falsified, and it is removed from our collection of scientific knowledge. "The passing of tests therefore makes not a jot of difference to the status of any hypothesis, though the failing of just one test may make a great deal of difference" (Miller 1988:22). Popper did not mean falsification to be a deterministic concept. He recognized that any empirical inference is to some extent uncertain (Popper 1982). In his discussion of disconfirmation, he wrote, "even if the asymmetry [between falsification and verification] is admitted, it is still impossible, for various reasons, that any theoretical system should ever be conclusively falsified" (Popper 1968:42).

In our view, Popper's ideas are fundamental for *formulating* theories. We should always design theories that are vulnerable to falsification. We should also learn from Popper's emphasis on the tentative nature of any theory. However, for *evaluating* existing social scientific theories, the asymmetry between verification and falsification is not as significant. Either one adds to our scientific knowledge. The question is less whether, in some general sense, a theory is false or not—virtually every interesting social science theory has at least one observable implication that appears wrong—than *how much of the world the theory can help us explain*. By Popper's rule, theories based on the assumption of rational choice would have been rejected long ago since they have been falsified in many specific instances. However, social scientists often choose to retain the assumption, suitably modified, because it provides considerable power in many kinds of research problems (see Cook and Levi 1990). The same point applies to virtually every other social science theory of interest. The process of trying to falsify theories in the social sciences is really one of searching for their bounds of applicability. If some observable implication indicates that the theory does not apply, we learn something; similarly, if the theory works, we learn something too.

For scientists (and especially for social scientists) evaluating properly formulated theories, Popper's fundamental asymmetry seems largely irrelevant. O'Hear (1989:43) made a similar point about the application of Popper's ideas to the physical sciences:

> Popper always tends to speak in terms of *explanations* of *universal* theories. But once again, we have to insist that proposing and testing universal theories is only part of the aim of science. There may be no true universal theories, owing to conditions differing markedly through time and space; this is a possibility we cannot overlook. But even if this were so, science could still fulfil [sic] many of its aims in giving us knowledge and true predictions about conditions in and around our spatio-temporal niche.

Surely this same point applies even more strongly to the social sciences.

Furthermore, Popper's evaluation of theories does not fundamentally distinguish between a newly formulated theory and one that has withstood numerous empirical tests. When we are testing for the deterministic distinction between the truth or fiction of a universal theory (of which there exists no interesting examples), Popper's view is appropriate, but from our perspective of searching for the bounds of a theory's applicability, his view is less useful. As we have indicated many times in this book, we require all inferences about specific hypotheses to be made by stating a best guess (an estimate) and a measure of the uncertainty of this guess. Whether we discover that the inference is consistent with our theory or inconsistent, our conclusion will have as much effect on our belief in the theory. Both consistency and inconsistency provide information about the truth of the theory and should affect the certainty of our beliefs.[13]

Consider the hypothesis that Democratic and Republican campaign strategies during American presidential elections have a small net effect on the election outcome. Numerous more specific hypotheses are implied by this one, such as that television commercials, radio commercials, and debates all have little effect on voters. Any test of the theory must really be a test of one of these hypotheses. One test of the theory has shown that forecasts of the outcome can be made very accurately with variables available only at the time of the conventions—and thus before the campaign (Gelman and King 1993). This test is consistent with the theory (if we can predict the election before the campaign, the campaign can hardly be said to have much of an impact), but it does not absolutely verify it. Some aspect of the campaign could have some small effect that accounts for some of the forecasting errors (and few researchers doubt that this is true). Moreover, the prediction could have been luck, or the campaign could have not included any innovative (and hence unpredictable) tactics during the years for which data were collected.

We could conduct numerous other tests by including variables in the forecasting model that measure aspects of the campaign, such as relative amounts of TV and radio time, speaking ability of the candidates, and judgements as to the outcomes of the debates. If all of these hypotheses show no effect, then Popper would say that our opinion is not changed in any interesting way: the theory that presidential campaigns have no effect is still standing. Indeed, if we did a thousand

[13] Some might call us (or accuse us of being!) "justificationists" or even "probabilistic justificationists" (see Lakatos 1970), but if we must be labeled, we prefer the more coherent, philosophical Bayesian label (see Leamer 1978; Zellner 1971; and Barnett 1982). In fact, our main difference with Popper is our goals. Given his precise goal, we agree with his procedure; given our goal, perhaps he might agree with ours. However, we believe that our goals are closer to those in use in the social sciences and are also closer to the ones likely to be successful.

similar tests and all were consistent with the theory, the theory could still be wrong since we have not tried every one of the infinite number of possible variables measuring the campaign. So even with a lot of results consistent with the theory, it still *might* be true that presidential campaigns influence voter behavior.

However, if a single campaign event—such as substantial accusations of immoral behavior—is shown to have some effect on voters, the theory would be falsified. According to Popper, even though this theory was not conclusively falsified (which he recognized as impossible), we learn more from it than the thousand tests consistent with the theory.

To us, this is not the way social science is or should be conducted. After a thousand tests in favor and one against, even if the negative test seemed valid with a high degree of certainty, we would not drop the theory that campaigns have no effect. Instead, we might modify it to say perhaps that normal campaigns have no effect except when there is considerable evidence of immoral behavior by one of the candidates—but since this modification would make our theory more restrictive, we would need to evaluate it with a new set of data before being confident of its validity. The theory would still be very powerful, and we would know somewhat more about the bounds to which the theory applied with each passing empirical evaluation. Each test of a theory affects both the estimate of its validity and the uncertainty of that estimate; and it may also affect to what extent we wish the theory to apply.

In the previous discussion, we suggested an important approach to theory, as well as issued a caution. The approach we recommended is one of sensitivity to the contingent nature of theories and hypotheses. Below, we argue for seeking broad application for our theories and hypotheses. This is a useful research strategy, but we ought always to remember that theories in the social sciences are unlikely to be universal in their applicability. Those theories that are put forward as applying to everything, everywhere—some versions of Marxism and rational choice theory are examples of theories that have been put forward with claims of such universality—are either presented in a tautological manner (in which case they are neither true nor false) or in a way that allows empirical disconfirmation (in which case we will find that they make incorrect predictions). Most useful social science theories are valid under particular conditions (in election campaigns without strong evidence of immoral behavior by a candidate) or in particular settings (in industrialized but not less industrialized nations, in House but not Senate campaigns). We should always try to specify the bounds of applicability of the theory or hypothesis. The next step is to raise the question: Why do these bounds exist? What is it about Senate races that invalidates generalizations that are true for House races? What is it about industrialization that changes the causal effects? What variable is missing from our analysis which could produce a more generally applicable theory? By asking such questions, we move beyond the boundaries of our theory or hypothesis to show what factors need to be considered to expand its scope.

But a note of caution must be added. We have suggested that the process of evaluating theories and hypotheses is a flexible one: particular empirical tests neither confirm nor disconfirm them once and for all. When an empirical test is inconsistent with our theoretically based expectations, we do not immediately throw out the theory. We may do various things: We may conclude that the evidence may have been poor due to chance alone; we may adjust what we consider to be the range of applicability of a theory or hypothesis even if it does not hold in a particular case and, through that adjustment, maintain our acceptance of the theory or hypothesis. Science proceeeds by such adjustments; but they can be dangerous. If we take them too far we make our theories and hypotheses invulnerable to disconfirmation. The lesson is that we must be very careful in adapting theories to be consistent with new evidence. We must avoid stretching the theory beyond all plausibility by adding numerous exceptions and special cases.

If our study disconfirms some aspect of a theory, we may choose to retain the theory but add an exception. Such a procedure is acceptable as long as we recognize the fact that we are reducing the claims we make for the theory. The theory, though, is less valuable since it explains less; in our terminology, we have less *leverage* over the problem we seek to understand.[14] Furthermore, such an approach may yield a "theory" that is merely a useless hodgepodge of various exceptions and exclusions. At some point we must be willing to discard theories and hypotheses entirely. Too many exceptions, and the theory should be rejected. Thus, by itself, *parsimony, the normative preference for theories with fewer parts, is not generally applicable*. All we need is our more general notion of maximizing leverage, from which the idea of parsimony can be fully derived when it is useful. The idea that science is largely a process of explaining many phenomena with just a few makes clear that theories with fewer parts are not better or worse. To maximize leverage, we should attempt to formulate theories that explain as much as possible with as little as possible. Sometimes this formulation is achieved via parsimony, but sometimes not. We can con-

[14] As always, when we do modify a theory to be consistent with evidence we have collected, then the theory (or that part of it on which our evidence bears) should be evaluated in a different context or new data set.

ceive of examples by which a slightly more complicated theory will explain vastly more of the world. In such a situation, we would surely use the nonparsimonious theory, since it maximizes leverage more than the more parsimonious theory.[15]

3.5.2 Rule 2: Build Theories That Are Internally Consistent

A theory which is internally inconsistent is not only falsifiable—it is false. Indeed, this is the only situation where the veracity of a theory is known without any empirical evidence: if two or more parts of a theory generate hypotheses that contradict one another, then no evidence from the empirical world can uphold the theory. Ensuring that theories are internally consistent should be entirely uncontroversial, but consistency is frequently difficult to achieve. One method of producing internally consistent theories is with formal, mathematical modeling. *Formal modeling* is a practice most developed in economics but increasingly common in sociology, psychology, political science, anthropology, and elsewhere (see Ordeshook 1986). In political science, scholars have built numerous substantive theories from mathematical models in rational choice, social choice, spatial models of elections, public economics, and game theory. This research has produced many important results, and large numbers of plausible hypotheses. One of the most important contributions of formal modeling is revealing the internal inconsistency in verbally stated theories.

However, as with other hypotheses, formal models do not constitute verified explanations without empirical evaluation of their predic-

[15] Another formulation of Popper's view is that "you can't prove a negative." You cannot, he argues, because a result consistent with the hypothesis might just mean that you did the wrong test. Those who try to prove the negative will always run into this problem. Indeed, their troubles will be not only theoretical but professional as well since journals are more likely to publish positive results rather than negative ones.

This has led to what is called *the file drawer problem*, which is clearest in the quantitative literature. Suppose no patterns exist in the world. Then five of every one hundred tests of any pattern will fall outside the 95 percent confidence interval and thus produce incorrect inferences. If we were to assume that journals publish positive rather than negative results, they will publish only those 5 percent that are "significant"; that is, they will publish only the papers that come to the wrong conclusions, and our file drawers will be filled with all the papers that come to the right conclusions! (See Iyengar and Greenhouse (1988) for a review of the statistical literature on this problem.) In fact, these incentives are well known by researchers, and it probably affects their behaviors as well. Even though the acceptance rate at many major social science journals is roughly 5 percent, the situation is not quite this bad, but it is still a serious problem. In our view, the file drawer problem could be solved if everyone adopted our alternative position. *A negative result is as useful as a positive one; both can provide just as much information about the world*. So long as we present our estimates and a measure of our uncertainty, we will be on safe ground.

tions. Formality does help us reason more clearly, and it certainly ensures that our ideas are internally consistent, but it does not resolve issues of empirical evaluation of social science theories. An assumption in a formal model in the social sciences is generally a convenience for mathematical simplicity or for ensuring that an equilibrium can be found. Few believe that the political world is mathematical in the same way that some physicists believe the physical world is. Thus, formal models are merely models—abstractions that should be distinguished from the world we study. Indeed, some formal theories make predictions that depend on assumptions that are vastly oversimplified, and these theories are sometimes not of much empirical value. They are only more precise in the abstract than are informal social science theories: they do not make more specific predictions about the real world, since the conditions they specify do not correspond, even approximately, to actual conditions.

Simplifications are essential in formal modeling, as they are in all research, but we need to be cautious about the inferences we can draw about reality from the models. For example, assuming that all omitted variables have no effect on the results can be very useful in modeling. In many of the formal models of qualitative research that we present throughout this book, we do precisely this. Assumptions like this are not usually justified as a feature of the world; they are only offered as a convenient feature of our model of the world. The results, then, apply exactly to the situation in which these omitted variables are irrelevant and may or may not be similar to results in the real world. We do not have to check the assumption to work out the model and its implications, but it is *essential* that we check the assumption during empirical evaluation. The assumption need not be correct for the formal model to be useful. But we cannot take untested or unjustified theoretical assumptions and use them in constructing empirical research designs. Instead, we must generally supplement a formal theory with additional features to make it useful for empirical study.

A good formal model should be abstract so that the key features of the problem can be apparent and mathematical reasoning can be easily applied. Consider, then, a formal model of the effect of proportional representation on political party systems, which implies that proportional representation fragments party systems. The key causal variable is the type of electoral system—whether it is a proportional representation system with seats allocated to parties on the basis of their proportion of the vote or a single-member district system in which a single winner is elected in each district. The dependent variable is the number of political parties, often referred to as the degree of party-system fragmentation. The leading hypothesis is that electoral systems based on proportional representation generate more political parties than do district-based electoral systems. For the sake of simplicity, such a model might well include only variables measuring some essential features of the electoral system and the degree of party-system fragmentation. Such a model would generate only a *hypothesis*, not a conclusion, about the relationship between proportional representation and party-system fragmentation in the real world. Such a hypothesis would have to be tested through the use of qualitative or quantitative empirical methods.

However, even though an implication of this model is that proportional representation fragments political parties, and even though no other variables were used in the model, using only two variables in an empirical analysis would be foolish. A study that indicates that countries with proportional representation have more fragmented party systems would ignore the problem of endogeneity (section 5.4), since countries which establish electoral systems based on a proportional allocation of seats to the parties may well have done so because of their already existent fragmented party systems. Omitted variable bias would also be a problem since countries with deep racial, ethnic, or religious divisions are probably also likely to have fragmented party systems, and countries with divisions of these kinds are more likely to have proportional representation.

Thus, both of the requirements for omitted variable bias (section 5.2) seem to be met: the omitted variable is correlated both with the explanatory and the dependent variable, and any analysis ignoring the variable of social division would therefore produce biased inferences.

The point should be clear: formal models are extremely useful for clarifying our thinking and developing internally consistent theories. For many theories, especially complex, verbally stated theories, it may be that only a formal model is capable of revealing and correcting internal inconsistencies. At the same time, formal models are unlikely to provide the correct empirical model for empirical testing. They certainly do not enable us to avoid any of the empirical problems of scientific inference.

3.5.3 Rule 3: Select Dependent Variables Carefully

Of course, we should do everything in research carefully, but choosing variables, especially dependent variables, is a particularly important decision. We offer the following three suggestions (based on mistakes that occur all too frequently in the quantitative and qualitative literatures):

First, *dependent variables should be dependent*. A very common mistake is to choose a dependent variable which in fact causes changes in our explanatory variables. We analyze the specific consequences of endogeneity and some ways to circumvent the problem in section 5.4, but we emphasize it here because the easiest way to avoid it is to choose explanatory variables that are clearly exogenous and dependent variables that are endogenous.

Second, *do not select observations based on the dependent variable so that the dependent variable is constant*. This, too, may seem a bit obvious, but scholars often choose observations in which the dependent variable does not vary at all (such as in the example discussed in section 4.3.1). Even if we do not deliberately design research so that the dependent variable is constant, it may turn out that way. But, as long as we have not predetermined that fact by our selection criteria, there is no problem. For example, suppose we select observations in two categories of an explanatory variable, and the dependent variable turns out to be constant across the two groups. This is merely a case where the estimated causal effect is zero.

Finally we should *choose a dependent variable that represents the variation we wish to explain*. Although this point seems obvious, it is actually quite subtle, as illustrated by Stanley Lieberson (1985:100):

> A simple gravitational exhibit at the Ontario Science Centre in Toronto inspires a heuristic example. In the exhibit, a coin and a feather are both released from the top of a vacuum tube and reach the bottom at virtually the same time. Since the vacuum is not a total one, presumably the coin reaches the bottom slightly ahead of the feather. At any rate, suppose we visualize a study in which a variety of objects is dropped without the benefit of such a strong control as a vacuum—just as would occur in nonexperimental social research. If social researchers find that the objects differ in the time that they take to reach the ground, typically they will want to know what characteristics determine these differences. Probably such characteristics of the objects as their density and shape will affect speed of the fall in a nonvacuum situation. If the social researcher is fortunate, such factors together will fully account for all of the differences among the objects in the velocity of their fall. If so, the social researcher will be very happy because all of the variation between objects will be accounted for. The investigator, applying standard social research-thinking will conclude that there is a complete understanding of the phenomenon *because all differences among the objects under study have been accounted for*. Surely there must be something faulty with our procedures if we can approach such a problem without even considering gravity itself.

The investigator's procedures in this example would be faulty only if the variable of interest were gravity. If gravity were the explanatory variable we cared about, our experiment does not vary it (since the experiment takes place in only one location) and therefore tells us nothing about it. However, the experiment Lieberson describes would be of great interest if we sought to understand variations in the time it will take for different types of objects to hit the ground when they are dropped from the same height under different conditions of air pressure. Indeed, even if we knew all about gravity, this experiment would still yield valuable information. But if, as Lieberson assumes, we were really interested in an inference about the causal effect of gravity, we would need a dependent variable which varied over observations with differing degrees of gravitational attraction. Likewise, in social science, we must be careful to ensure that we are really interested in understanding our dependent variable, rather than the background factors that our research design holds constant.

Thus, we need the entire range of variation in the dependent variable to be a possible outcome of the experiment in order to obtain an unbiased estimate of the impact of the explanatory variables. Artificial limits on the range or values of the dependent variable produce what we define (in section 4.3) as selection bias. For instance, if we are interested in the conditions under which armed conflict breaks out, we cannot choose as observations only those instances where the result is armed conflict. Such a study might tell us a great deal about variations among observations of armed conflict (as the gravity experiment tells us about variations in speed of fall of various objects) but will not enable us to explore the sources of armed conflict. A better design if we want to understand the sources of armed conflict would be one that selected observations according to our explanatory variables and allowed the dependent variable the *possibility* of covering the full range from there being little or no threat of a conflict through threat situations to actual conflict.

3.5.4 Rule 4: Maximize Concreteness

Our fourth rule, which follows from our emphasis on falsifiability, consistency, and variation in the dependent variable is to maximize concreteness. We should choose observable, rather than unobservable, concepts wherever possible. Abstract, unobserved concepts such as utility, culture, intentions, motivations, identification, intelligence, or the national interest are often used in social science theories. They can play a useful role in theory *formulation*; but they can be a hindrance to empirical *evaluation* of theories and hypotheses unless they can be defined in a way such that they, or at least their implications, can be observed and measured. Explanations involving concepts such as culture or national interest or utility or motivation are suspect unless we can measure the concept independently of the dependent variable that we are explaining. When such terms are used in explanations, it is too easy to use them in ways that are tautological or have no differentiating, observable implications. An act of an individual or a nation may be explained as resulting from a desire to maximize utility, to fulfill intentions, or to achieve the national interest. But the evidence that the act maximized utility or fulfilled intentions or achieved the national interest is the fact that the actor or the nation engaged in it. It is incumbent upon the researcher formulating the theory to specify clearly and precisely what observable implications of the theory would indicate its veracity and distinguish it from logical alternatives.

In no way do we mean to imply by this rule that concepts like intentions and motivations are unimportant. We only wish to recognize that the standard for explanation in any *empirical* science like ours must be *empirical* verification or falsification. Attempting to find empirical evidence of abstract, unmeasurable, and unobservable concepts will necessarily prove more difficult and less successful than for many imperfectly conceived specific and concrete concepts. The more abstract our concepts, the less clear will be the observable consequences and the less amenable the theory will be to falsification.

Researchers often use the following strategy. They begin with an abstract concept of the sort listed above. They agree that it cannot be measured directly; therefore, they suggest specific indicators of the abstract concept that can be measured and use them in their explanations. The choice of the specific indicator of the more abstract concept is justified on the grounds that it is observable. Sometimes it is the *only thing* that is observable (for instance, it is the only phenomenon for which data are available or the only type of historical event for which records have been kept). This is a perfectly respectable, indeed usually necessary, aspect of empirical investigation.

Sometimes, however, it has an unfortunate side. Often the specific indicator is far from the original concept and has only an indirect and uncertain relationship to it. It may not be a valid indicator of the abstract concept at all. But, after a quick apology for the gap between the abstract concept and the specific indicator, the researcher labels the indicator with the abstract concept and proceeds onward as if he were measuring that concept directly. Unfortunately, such reification is common in social science work, perhaps more frequently in quantitative than in qualitative research, but all too common in both. For example, the researcher has figures on mail, trade, tourism and student exchanges and uses these to compile an index of "societal integration" in Europe. Or the researcher asks some survey questions as to whether respondents are more concerned with the environment or making money and labels different respondents as "materialists" and "post-materialists." Or the researcher observes that federal agencies differ in the average length of employment of their workers and converts this into a measure of the "institutionalization" of the agencies.

We should be clear about what we mean here. The gap between concept and indicator is inevitable in much social science work. And we use general terms rather than specific ones for good reasons: they allow us to expand our frame of reference and the applicability of our theories. Thus we may talk of legislatures rather than of more narrowly defined legislative categories such as parliaments or specific institutions such as the German Bundestag. Or we may talk of "decision-making bodies" rather than legislatures when we want our theory to apply to an even wider range of institutions. (In the next section we, in fact, recommend this.) Science depends on such abstract classifications—or else we revert to summarizing historical detail. But our abstract and general terms must be connected to specific measureable concepts at some point to allow empirical testing. The fact of that connection—and the distance that must be traversed to make it—must always be kept in mind and made explicit. Furthermore, the choice of a high level of abstraction must have a real justification in terms of the theoretical problem at hand. It must help make the connection between the specific research at hand—in which the particular indicator is the main actor—and the more general problem. And it puts a burden on us to see that additional research using other specific indicators is carried on to bolster the assumption that our specific indicators really relate to some broader concept. The abstract terms used in the examples above—"societal integration," "post-materialism," and "institutionalization"—may be measured reasonably by the specific indicators cited. We do not deny that the leap from specific indicator to general abstract concept must be made—we have to make such a leap to carry on social science research. The leap must, however, be made with care, with justification, and with a constant "memory" of where the leap began.

Thus, we do not argue against abstractions. But we do argue for a language of social research that is as concrete and precise as possible. If we have no alternative to using unobservable constructs, as is usually the case in the social sciences, then we should at least *choose ideas with observable consequences*. For example, "intelligence" has never been directly observed but it is nevertheless a very useful concept. We have numerous tests and other ways to evaluate the implications of intelligence. On the other hand, if we have the choice between "the institutionalization of the presidency" and "size of the White House staff," it is usually better to choose the latter. We may argue that the size of the White House staff is related to the general concept of the institutionalization of the presidency, but we ought not to reify the narrower concept as identical to the broader. And, if size of staff means institutionalization, we should be able to find other measures of institutionalization that respond to the same explanatory variables as does size of staff. Below, we shall discuss "maximizing leverage" by expanding our dependent variables.

Our call for concreteness extends, in general, to the words we use to describe our theory. If a reader has to spend a lot of time extracting the precise meanings of the theory, the theory is of less use. There should be as little controversy as possible over what we mean when we describe a theory. To help in this goal of specificity, even if we are not conducting empirical research ourselves, we should spend time explicitly considering the observable implications of the theory and even possible research projects we could conduct. The vaguer our language, the less chance we will be wrong—but the less chance our work will be at all useful. It is better to be wrong than vague.

In our view, eloquent writing—a scarce commodity in social science—should be encouraged (and savored) in presenting the rationale for a research project, arguing for its significance, and providing rich descriptions of events. Tedium never advanced any science. However, as soon as the subject becomes causal or descriptive inference, where we are interested in observations and generalizations that are expected to persist, we require concreteness and specificity in language and thought.[16]

[16] The rules governing the best questions to ask in interviews are almost the same as those used in designing explanations: Be as concrete as possible. We should not ask conservative, white Americans, "Are you racist?", rather, "Would you mind if your daughter married a black man?" We should not ask someone if he or she is knowledgeable about politics; we should ask for the names of the Secretary of State and Speaker of the House. In general and wherever possible, *we must not ask an interviewee to do our work for us*. It is best not to ask for estimates of causal effects; we must ask for measures of the explanatory and dependent variables, and estimate the causal effect ourselves. We must not ask for motivations, but rather for facts.

This rule is not meant to imply that we should never ask people why they did something. Indeed, asking about motivations is often a productive means of generating hypotheses. Self-reported motivations may also be a useful set of observable implications. However, the answer given must be interpreted as the interviewee's response to the researcher's question, not necessarily as the correct answer. If questions such as these are to be of use, we should design research so that a particular answer given (with whatever justifications, embellishments, lies, or selective memories we may encounter) is an observable implication.

3.5.5 Rule 5: State Theories in as Encompassing Ways as Feasible

Within the constraints of guaranteeing that the theory will be falsifiable and that we maximize concreteness, the theory should be formulated so that it explains as much of the world as possible. We realize that there is some tension between this fifth rule and our earlier injunction to be concrete. We can only say that both goals are important, though in many cases they may conflict, and we need to be sensitive to both in order to draw a balance.

For example, we must not present our theory as if it only applies to the German Bundestag when there is reason to believe that it might apply to all independent legislatures. We need not provide evidence for all implications of the theory in order to state it, so long as we provide a reasonable estimate of uncertainty that goes along with it. It may be that we have provided strong evidence in favor of the theory in the German Bundestag. Although we have no evidence that it works elsewhere, we have no evidence against it either. The broader reference is useful if we remain aware of the need to evaluate its applicability. Indeed, expressing it as a hypothetically broader reference may force us to think about the structural features of the theory that would make it apply or not to other independent legislatures. For example, would it apply to the U.S. Senate, where terms are staggered, to the New Hampshire Assembly, which is much larger relative to the number of constituents, or to the British House of Commons, in which party voting is much stronger? An important exercise is stating what we think are systematic features of the theory that make it applicable in different areas. We may learn that we were wrong, but that is considerably better than not having stated the theory with sufficient precision in the first place.

This rule might seem to conflict with Robert Merton's ([1949] 1968) preference for "theories of the middle-range," but even a cursory reading of Merton should indicate that this is not so. Merton was reacting to a tradition in sociology where "theories" such as Parson's "theory of action" were stated so broadly that they could not be falsified. In political science, Easton's "systems theory" (1965) is in this same tradition (see Eckstein 1975:90). As one example of the sort of criticism he was fond of making, Merton ([1949] 1968: 43) wrote, "So far as one can tell, the theory of role-sets is not inconsistent with such broad theoretical orientations as Marxist theory, functional analysis, social behaviorism, Sorokin's integral sociology, or Parson's theory of action." Merton is not critical of the theory of role-sets, which he called a middle-range theory, rather he is arguing against those "broad theoretical orientations," with which almost any more specific theory or empirical observation is consistent. Merton favors "middle-range" theories but we believe he would agree that theories should be stated as broadly as possible as long as they remain falsifiable and concrete. Stating theories as broadly as possible is, to return to a notion raised earlier, a way of maximizing leverage. If the theory is testable—and the danger of very broad theories is, of course, that they may be phrased in ways that are not testable—then the broader the better; that is, the broader, the greater the leverage.