

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 non-replicable 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, adminis-

tering 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 meteorite'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 accor-

dance 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 knowledge 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.

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 data-gathering 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 non-

cooperative 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 cor-

rect. 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, Lieberman (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 trust-

worthy.) 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.