# Experimental and Quasi-Experimental Designs for Generalized Causal Inference

# Preface

THIS IS a book for those who have already decided that identifying a dependable relationship between a cause and its effects is a high priority and who wish to consider experimental methods for doing so. Such causal relationships are of great importance in human affairs. The rewards associated with being correct in identifying causal relationships can be high, and the costs of misidentification can be tremendous. To know whether increased schooling pays off in later life happiness or in increased lifetime earnings is a boon to individuals facing the decision about whether to spend more time in school, and it also helps policymakers determine how much financial support to give educational institutions. In health, from the earliest years of human existence, causation has helped to identify which strategies are effective in dealing with disease. In pharmacology, divinations and reflections on experience in the remote past sometimes led to the development of many useful treatments, but other judgments about effective plants and ways of placating gods were certainly more incorrect than correct and presumably contributed to many unnecessary deaths. The utility of finding such causal connections is so widely understood that much effort goes to locating them in both human affairs in general and in science in particular.

However, history also teaches us that it is rare for those causes to be so universally true that they hold under all conditions with all types of people and at all historical time periods. All causal statements are inevitably contingent. Thus, although threat from an out-group often causes in-group cohesion, this is not always the case. For instance, in 1492 the king of Granada had to watch as his Moorish subjects left the city to go to their ancestral homes in North Africa, being unwilling to fight against the numerically superior troops of the Catholic kings of Spain who were lodged in Santa Fe de la Frontera nearby. Here, the external threat from the out-group of Christian Spaniards led not to increased social cohesion among the Moslem Spaniards but rather to the latter's disintegration as a defensive force. Still, some causal hypotheses are more contingent than others. It is of obvious utility to learn as much as one can about those contingencies and to identify those relationships that hold more consistently. For instance, aspirin is such a wonderful drug because it reduces the symptoms associated with many different kinds of illness, including head colds, colon cancer, and cardiovascular disease; it works whether taken at low or high altitudes, in warm or cold climes, in

capsule or liquid form, by children or adults; and it is effective in people who suffer from many types of secondary infirmity other than stomach ulcers. However, other drugs are more limited in their range of application, able to alleviate only one type of cancer, say, and then only in patients with a certain degree of physical strength, only when the dose is strictly adhered to, or only if antibodies have not already developed to counteract the drug. Although the lay use of causal language is often general, it is important to identify the most important conditions that limit the applicability of causal connections.

This book has two major purposes that parallel this dual interest in identifying causal connections and in understanding their generality. The first is to describe ways in which testing causal propositions can be improved in specific research projects. To achieve this we prefer to use what we call structural design features from the theory of experimentation rather than to use statistical modeling procedures. Recent statistical developments concerning causal inference in observational data (e.g., Holland, 1986; Rosenbaum, 1995a; Rubin, 1986) have advanced understanding enormously. However, to judge from our experience in consulting to field experiments, those developments may have also created an unrealistic expectation among some readers that new statistics, such as propensity score matching or stratification, selection bias models, and hidden bias sensitivity analyses, can by themselves suffice to warrant valid causal inference. Although such adjustments are sometimes necessary and frequently useful *after* good experimental design features are in place, they may work poorly without such features. Too few economists and statisticians who are associated with these new developments have emphasized the importance of such design features—though we are pleased to see more recent emphasis on these design features in, for example, Heckman, Ichimura, and Todd (1997) on the use of common measurement frameworks and local controls, Winship and Morgan (1999) on the usefulness of multiple pretests and posttests, and Rosenbaum (1999b) on the importance of many design choices in observational data. We want this book to complement such statistical work by emphasizing that, in the interplay between design and statistics, design rules (Shadish & Cook, 1999)!

The second purpose of this book is to describe ways to improve generalizations about causal propositions. Although formal sampling procedures are the best warranted means of generalizing, they rarely apply to generalizing about causal relationships. So we turn instead to improving causal generalization through a grounded theory of causal generalization. This theory reflects the principles that scientists use in their daily work to make generalizations in such diverse areas as animal modeling of human disease, deciding if a specimen belongs to a more general category, identifying general trends in literature reviews, and deciding whether epidemiological studies support a general connection between secondhand smoke and cancer. The result is, we hope, a more practical theory of causal generalization than sampling theory but one that incorporates sampling theory as a special case.

In these dual purposes, this book is a successor to Campbell and Stanley (1963) and Cook and Campbell (1979). However, it differs from them in several important ways. The most obvious is the emphasis now placed on the generalization of causal connections. Though this past work clearly acknowledged the importance of such generalization and even coined the phrase "external validity" to refer to it, much more emphasis was placed on examining the plausibility of conclusions about whether a particular relationship was likely to be causal in the unique research context in which it was tested. In this book, methods for studying external validity now receive the extensive attention that our past work gave to internal validity.

A second difference is that we have had to grapple with recent philosophy of science that questions some of the most important pillars on which our received scientific logic stands, especially as concerns the possibility of objectivity and the fallibility of both induction *and* deduction as ways of acquiring certain knowledge. Also relevant are the implications of many descriptive findings from metascience (the systematic study of the history, sociology, psychology, and philosophy of science) that illustrate the high frequency with which scientific practice deviates from the preferred scientific logic of the day. Science is conducted by humans and is validated by a collectivity of scientists who have cognitive and economic interests to define, defend, and promote. So even more than its predecessors, this book assumes panfallibility, the total and inevitable absence of certain knowledge from the methods social scientists use. But we do not throw in the towel because of this belief, nor do we counsel that "anything goes." The fallible nature of knowledge need not entail either worthlessness (i.e., if it's not perfect, it's worthless) or strong methodological relativism (that no method ever has any privileged status over any other for any purpose). Rather, we defend the beliefs that some causal statements are better warranted than others and that logic and craft experience in science indicate that some practices are often (but not always) superior to others *for causal purposes,* though not necessarily for other purposes.

A third difference concerns the emphasis placed on design elements rather than on designs, especially when considering experimental studies without random assignment to treatment conditions. The scientific practice most often associated with causal research is the experiment, which in all its many forms is the main focus of this book. Today, experimentation refers to a systematic study designed to examine the consequences of deliberately varying a potential causal agent. Experiments require (1) variation in the treatment, (2) posttreatment measures of outcomes, (3) at least one unit on which observation is made, and (4) a mechanism for inferring what the outcome would have been without treatment—the so-called "counterfactual inference" against which we infer that the treatment produced an effect that otherwise would not have occurred. We shall see that there are many other structural features of experimentation, most of which serve the purpose of improving the quality of this counterfactual inference. But as popular as experiments are in the natural sciences, mathematical statistics, medicine, psychology,

education, and labor economics, they are not the only form of research that claims to justify causal conclusions. Many correlational studies in sociology, political science, developmental science, and certain branches of economics rely on causal ideas for theory development but do not knowingly use the structures or the formal language of experimentation. Yet we contend that all nonexperimental methods can be analyzed for the structural design elements that are or are not present in them, clarifying the likely strengths and weaknesses they have for inferring cause. In describing the structural elements that characterize experimentation and in showing how they can be combined to create experimental designs that have not been used before, we claim a general utility for thinking in terms of structural design elements rather than in terms of a finite series of designs. Such designs were the centerpiece of the predecessors to this book (Campbell & Stanley, 1963; Cook & Campbell, 1979). By focusing on design elements instead, we hope to help readers acquire a set of tools that is flexible enough so that some of them will be relevant for improving causal claims in almost any research context.

A fourth difference is that this book, unlike Cook and Campbell (1979), does not deal so much with the statistical analysis of data. Rather than full chapters of statistical detail, we present brief paragraphs or occasional chapter appendices about data analysis, couched more at a conceptual level, with infrequent equations, often placed in footnotes—just enough, we hope, to clarify some of the essential issues and to refer the reader to more detailed sources. In part, the reason for this change is practical. Twenty years ago, accessible descriptions of statistical procedures for such methods as time series or nonequivalent control group designs were so rare that extended treatment was warranted. Today, however, statistical treatments of these matters are widespread at an array of technical levels, so our space is better devoted to developments concerning design and generalization. However, our reduced attention to statistics also reflects our preference for design solutions over statistical solutions for causal inference, for all the reasons previously cited.

A fifth difference is that this book includes extended treatment of randomized experiments, with three chapters devoted to their logic and design and to practical problems and solutions in their implementation. Especially in the latter area, the past few decades have seen many new developments addressing a host of problems such as poor treatment implementation, preventing and analyzing attrition, ensuring the integrity of the assignment process, and conducting experiments that better address certain ethical and legal matters. These developments promise to improve the practicality of randomized experiments. As an added benefit, many of them will improve nonrandomized experiments as well.

A sixth difference is that this book introduces some emendations to the general conceptual scheme that has always been the central hallmark of Campbell's work over the years, the validity typology. The changes are minor in most respects, for we still retain the overall emphasis on four validity types (internal, statistical conclusion, construct, and external) and on the centrality of identifying plausible threats to validity in practical causal inference. But we have changed the scheme

in a number of ways. For example, with statistical conclusion validity, we have tried to display a greater sensitivity to the magnitude of an effect than to its statistical significance. Our thinking on generalization (both external and construct validity) now reflects the influence of Cronbach's (e.g., 1982) cogent writings on the problems of causal generalization. And we have made minor changes to lists of threats to validity. Although many of these changes may be of interest only to fellow theorists of experimental methodology, we hope that some of them (for example, the increased emphasis on magnitude of effects) will have an impact on the practice of experimentation as well.

Despite these changes, this book retains an overall emphasis on field experimentation, on human behavior in nonlaboratory settings (although much of the book will apply to laboratory experiments). In such settings as schools, businesses, clinics, hospitals, welfare agencies, and homes, researchers have far from perfect control, are typically guests and not royalty, have to negotiate and not command, and often must compromise rather than get everything they would like. Some compromises cause more worry than others. In particular, field experimenters are reluctant to give up all control over the measurement, selection, and treatment scheduling process and, especially, over treatment assignment, for causal inference is most difficult when individuals completely self-select themselves into groups that vary in treatment exposure. However, it is clear that such control is usually a subject for negotiation rather than unilateral decision.

As with all books, the authors owe debts to many people who have helped to shape its ideas. Colleagues who gave us raw data with which to create graphs and figures include Xavier Ballart (Figure 6.4), Dick Berk (7.5), Robert Gebotys (6.2), Janet Hankin (6.3), Lynn McClannahan (6.14), Dick McCleary (6.1, 6.10), Jack McKillip (6.13), Steve Mellor (7.3), Mel Mark (7.3), and Clara Riba (6.4). Others helped us by reading and criticizing parts or all of the book, by providing examples to use in it, or by stimulating us to think more about key problems, including Mary Battle, Joseph Cappelleri, Laura Dreuth, Peter Grant (and his students), John Hetherington, Paul Holland, Karen Kirkhart, Dan Lewis, Ken Lichstein, Sue Marcus, Mel Mark, Dick McCleary, Jack McKillip, David Murray, Jennifer Owens, Dave Rindskopf, Virgil Sheets, William Trochim, Alan Vaux, Steve West (and his students), and Chris Winship. We single out Laura Leviton, Scott Maxwell, and Chip Reichardt for providing exceptionally detailed and helpful reviews. However, because the book has been in the writing for a decade, memory for these contributions and influences undoubtedly fails us, and so we apologize to all those whose names we have inadvertently omitted.

We acknowledge several organizations for their support of the research and preparation of this book. William Shadish's contribution was partially supported by a sabbatical award from the Institute for Policy Research at Northwestern University, by a Supplemental Sabbatical Award from the James McKeen Cattell Foundation, by a Professional Development Assignment Award from the University of Memphis, and by both the Center for Applied Psychological Research and the psychology department at the University of Memphis. Thomas Cook's contribution

was partially funded by fellowships from the Institute for Advanced Study in the Behavioral Sciences at Stanford University and from the Max Planck Institute for Human Development in Berlin.

Finally, we want to acknowledge the contributions of the third author of this book, Donald Thomas Campbell, who passed away in May 1996 when this book was only half done. Acknowledging those contributions is no easy task. Clearly, they go far beyond the particular writing he did for this book, given how profoundly and broadly his ideas influenced both his colleagues and his students. He was the founder of the entire tradition of field experimentation and quasi-experimentation represented in this book, a tradition that is so closely associated with him that we and others often call it Campbellian. Many of the most important concepts in this book, such as internal and external validity, threats to validity and their logic, and quasi-experimentation, were originated and developed by him. Many others of his ideas—about the fallibility of knowledge constructions ("We are cousins to the amoeba, and have received no direct revelations not shared with it. How then, indeed, could we know for certain?"), about the fitful and haphazard nature of scientific progress ("The fish-scale model of collective omniscience"), and about the social nature of the scientific enterprise ("A tribal model of the social system vehicle carrying scientific knowledge")—are so much a part of our thinking that they appear implicitly throughout the book. Our debt to Campbell, both as his colleagues and as his students, is undoubtedly greater than we recognize.

Campbell (e.g., 1988) was fond of a metaphor often used by the philosopher and mathematician W. V. Quine, that scientists are like sailors who must repair a rotting ship at sea. They trust the great bulk of timbers while they replace a particularly weak plank. Each of the timbers that they now trust they may, in its turn, replace. The proportion of the planks they are replacing to those they treat as sound must always be small. Campbell used this metaphor to illustrate the pervasive role of trust in science, and the lack of truly firm foundations in science. In the spirit of this metaphor, then, the following four lines from Seamus Heaney's (1991) poem "The Settle Bed" are an apt summary not only of Campbell's love of Quine's metaphor, but also of Campbell's own contribution to one of the ships of science:

> And now this is 'an inheritance'—
> Upright, rudimentary, unshiftably planked
> In the long ago, yet willable forward
> Again and again and again.[1]

Alternatively, for those readers whose preferences are more folksy, we close with words that Woody Guthrie wrote in the song *Another Man's Done Gone,* written as he anticipated his own death: "I don't know, I may go down or up or anywhere,

[1]Excerpt from "The Settle Bed," from *Opened Ground: Selected Poems 1966–1998* by Seamus Heaney. Copyright © 1998 by Seamus Heaney. Reprinted by permission of Farrar, Straus and Giroux, LLC.

but I feel like this scribbling might stay." We hope this book helps keep Don's seminal contributions to field experimentation alive for generations to come.

William R. Shadish
Memphis, Tennessee

Thomas D. Cook
Evanston, Illinois

# 1

# Experiments and Generalized Causal Inference

**Ex·per·i·ment** (ĭk-spĕr′ə-mənt): [Middle English from Old French from Latin *experimentum,* from *experiri,* to try; see *per-* in Indo-European Roots.] n. Abbr. exp., expt. 1. a. A test under controlled conditions that is made to demonstrate a known truth, examine the validity of a hypothesis, or determine the efficacy of something previously untried. b. The process of conducting such a test; experimentation. 2. An innovative act or procedure: *"Democracy is only an experiment in government" (William Ralph Inge).*

**Cause** (kôz): [Middle English from Old French from Latin *causa,* reason, purpose.] n. 1. a. The producer of an effect, result, or consequence. b. The one, such as a person, an event, or a condition, that is responsible for an action or a result. v. 1. To be the cause of or reason for; result in. 2. To bring about or compel by authority or force.

TO MANY historians and philosophers, the increased emphasis on experimentation in the 16th and 17th centuries marked the emergence of modern science from its roots in natural philosophy (Hacking, 1983). Drake (1981) cites Galileo's 1612 treatise *Bodies That Stay Atop Water, or Move in It* as ushering in modern experimental science, but earlier claims can be made favoring William Gilbert's 1600 study *On the Loadstone and Magnetic Bodies,* Leonardo da Vinci's (1452–1519) many investigations, and perhaps even the 5th-century B.C. philosopher Empedocles, who used various empirical demonstrations to argue against Parmenides (Jones, 1969a, 1969b). In the everyday sense of the term, humans have been experimenting with different ways of doing things from the earliest moments of their history. Such experimenting is as natural a part of our life as trying a new recipe or a different way of starting campfires.

However, the scientific revolution of the 17th century departed in three ways from the common use of observation in natural philosophy at that time. First, it increasingly used observation to correct errors in theory. Throughout history, natural philosophers often used observation *in* their theories, usually to win philosophical arguments by finding observations that supported their theories. However, they still subordinated the use of observation to the practice of deriving theories from "first principles," starting points that humans know to be true by our nature or by divine revelation (e.g., the assumed properties of the four basic elements of fire, water, earth, and air in Aristotelian natural philosophy). According to some accounts, this subordination of evidence to theory degenerated in the 17th century: "The Aristotelian principle of appealing to experience had degenerated among philosophers into dependence on reasoning supported by casual examples and the refutation of opponents by pointing to apparent exceptions not carefully examined" (Drake, 1981, p. xxi). When some 17th-century scholars then began to use observation to *correct* apparent errors in theoretical and religious first principles, they came into conflict with religious or philosophical authorities, as in the case of the Inquisition's demands that Galileo recant his account of the earth revolving around the sun. Given such hazards, the fact that the new experimental science tipped the balance toward observation and away from dogma is remarkable. By the time Galileo died, the role of systematic observation was firmly entrenched as a central feature of science, and it has remained so ever since (Harré, 1981).

Second, before the 17th century, appeals to experience were usually based on passive observation of ongoing systems rather than on observation of what happens after a system is deliberately changed. After the scientific revolution in the 17th century, the word **experiment** (terms in **boldface** in this book are defined in the Glossary) came to connote taking a deliberate action followed by systematic observation of what occurred afterward. As Hacking (1983) noted of Francis Bacon: "He taught that not only must we observe nature in the raw, but that we must also 'twist the lion's tale', that is, manipulate our world in order to learn its secrets" (p. 149). Although passive observation reveals much about the world, active manipulation is required to discover some of the world's regularities and possibilities (Greenwood, 1989). As a mundane example, stainless steel does not occur naturally; humans must manipulate it into existence. Experimental science came to be concerned with observing the effects of such manipulations.

Third, early experimenters realized the desirability of controlling extraneous influences that might limit or bias observation. So telescopes were carried to higher points at which the air was clearer, the glass for microscopes was ground ever more accurately, and scientists constructed laboratories in which it was possible to use walls to keep out potentially biasing ether waves and to use (eventually sterilized) test tubes to keep out dust or bacteria. At first, these controls were developed for astronomy, chemistry, and physics, the natural sciences in which interest in science first bloomed. But when scientists started to use experiments in areas such as public health or education, in which extraneous influences are harder to control (e.g., Lind, 1753), they found that the controls used in natural

science in the laboratory worked poorly in these new applications. So they developed new methods of dealing with extraneous influence, such as **random assignment** (Fisher, 1925) or adding a nonrandomized **control group** (Coover & Angell, 1907). As theoretical and observational experience accumulated across these settings and topics, more sources of **bias** were identified and more methods were developed to cope with them (Dehue, 2000).

Today, the key feature common to all experiments is still to deliberately vary something so as to discover what happens to something else later—to discover the effects of presumed causes. As laypersons we do this, for example, to assess what happens to our blood pressure if we exercise more, to our weight if we diet less, or to our behavior if we read a self-help book. However, *scientific* experimentation has developed increasingly specialized substance, language, and tools, including the practice of field experimentation in the social sciences that is the primary focus of this book. This chapter begins to explore these matters by (1) discussing the nature of causation that experiments test, (2) explaining the specialized terminology (e.g., randomized experiments, quasi-experiments) that describes social experiments, (3) introducing the problem of how to generalize causal connections from individual experiments, and (4) briefly situating the experiment within a larger literature on the nature of science.

## EXPERIMENTS AND CAUSATION

A sensible discussion of experiments requires both a vocabulary for talking about causation and an understanding of key concepts that underlie that vocabulary.

### Defining Cause, Effect, and Causal Relationships

Most people intuitively recognize causal relationships in their daily lives. For instance, you may say that another automobile's hitting yours was a cause of the damage to your car; that the number of hours you spent studying was a cause of your test grades; or that the amount of food a friend eats was a cause of his weight. You may even point to more complicated causal relationships, noting that a low test grade was demoralizing, which reduced subsequent studying, which caused even lower grades. Here the same variable (low grade) can be both a cause and an effect, and there can be a **reciprocal relationship** between two variables (low grades and not studying) that cause each other.

Despite this intuitive familiarity with causal relationships, a precise definition of cause and effect has eluded philosophers for centuries.[1] Indeed, the definitions

1. Our analysis reflects the use of the word *causation* in ordinary language, not the more detailed discussions of cause by philosophers. Readers interested in such detail may consult a host of works that we reference in this chapter, including Cook and Campbell (1979).

of terms such as *cause* and *effect* depend partly on each other and on the causal relationship in which both are embedded. So the 17th-century philosopher John Locke said: "That which produces any simple or complex idea, we denote by the general name *cause,* and that which is produced, *effect*" (1975, p. 324) and also: "A *cause* is that which makes any other thing, either simple *idea,* substance, or mode, begin to be; and an *effect* is that, which had its beginning from some other thing" (p. 325). Since then, other philosophers and scientists have given us useful definitions of the three key ideas—cause, effect, and causal relationship—that are more specific and that better illuminate how experiments work. We would not defend any of these as the true or correct definition, given that the latter has eluded philosophers for millennia; but we do claim that these ideas help to clarify the scientific practice of probing causes.

### Cause

Consider the cause of a forest fire. We know that fires start in different ways—a match tossed from a car, a lightning strike, or a smoldering campfire, for example. None of these causes is necessary because a forest fire can start even when, say, a match is not present. Also, none of them is sufficient to start the fire. After all, a match must stay "hot" long enough to start combustion; it must contact combustible material such as dry leaves; there must be oxygen for combustion to occur; and the weather must be dry enough so that the leaves are dry and the match is not doused by rain. So the match is part of a constellation of conditions without which a fire will not result, although some of these conditions can be usually taken for granted, such as the availability of oxygen. A lighted match is, therefore, what Mackie (1974) called an **inus condition**—"an *insufficient* but *nonredundant* part of an *unnecessary* but *sufficient* condition" (p. 62; italics in original). It is insufficient because a match cannot start a fire without the other conditions. It is nonredundant only if it adds something fire-promoting that is uniquely different from what the other factors in the constellation (e.g., oxygen, dry leaves) contribute to starting a fire; after all, it would be harder to say whether the match caused the fire if someone else simultaneously tried starting it with a cigarette lighter. It is part of a sufficient condition to start a fire in combination with the full constellation of factors. But that condition is not necessary because there are other sets of conditions that can also start fires.

A research example of an inus condition concerns a new potential treatment for cancer. In the late 1990s, a team of researchers in Boston headed by Dr. Judah Folkman reported that a new drug called Endostatin shrank tumors by limiting their blood supply (Folkman, 1996). Other respected researchers could not replicate the effect even when using drugs shipped to them from Folkman's lab. Scientists eventually replicated the results after they had traveled to Folkman's lab to learn how to properly manufacture, transport, store, and handle the drug and how to inject it in the right location at the right depth and angle. One observer labeled these contingencies the "in-our-hands" phenomenon, meaning "even we don't

know which details are important, so it might take you some time to work it out" (Rowe, 1999, p. 732). Endostatin was an inus condition. It was insufficient cause by itself, and its effectiveness required it to be embedded in a larger set of conditions that were not even fully understood by the original investigators.

Most causes are more accurately called inus conditions. Many factors are usually required for an effect to occur, but we rarely know all of them and how they relate to each other. This is one reason that the causal relationships we discuss in this book are not deterministic but only increase the probability that an effect will occur (Eells, 1991; Holland, 1994). It also explains why a given causal relationship will occur under some conditions but not universally across time, space, human populations, or other kinds of treatments and outcomes that are more or less related to those studied. To different degrees, all causal relationships are context dependent, so the generalization of experimental effects is always at issue. That is why we return to such generalizations throughout this book.

### *Effect*

We can better understand what an effect is through a **counterfactual** model that goes back at least to the 18th-century philosopher David Hume (Lewis, 1973, p. 556). A counterfactual is something that is contrary to fact. In an experiment, we observe what *did happen* when people received a treatment. The counterfactual is knowledge of what *would have happened* to those same people if they simultaneously had not received treatment. An **effect** is the difference between what did happen and what would have happened.

We cannot actually observe a counterfactual. Consider phenylketonuria (PKU), a genetically-based metabolic disease that causes mental retardation unless treated during the first few weeks of life. PKU is the absence of an enzyme that would otherwise prevent a buildup of phenylalanine, a substance toxic to the nervous system. When a restricted phenylalanine diet is begun early and maintained, retardation is prevented. In this example, the cause could be thought of as the underlying genetic defect, as the enzymatic disorder, or as the diet. Each implies a different counterfactual. For example, if we say that a restricted phenylalanine diet caused a decrease in PKU-based mental retardation in infants who are phenylketonuric at birth, the counterfactual is whatever would have happened had these same infants not received a restricted phenylalanine diet. The same logic applies to the genetic or enzymatic version of the cause. But it is impossible for these very same infants *simultaneously* to both have and not have the diet, the genetic disorder, or the enzyme deficiency.

So a central task for all cause-probing research is to create reasonable approximations to this physically impossible counterfactual. For instance, if it were ethical to do so, we might contrast phenylketonuric infants who were given the diet with other phenylketonuric infants who were not given the diet but who were similar in many ways to those who were (e.g., similar race, gender, age, socioeconomic status, health status). Or we might (if it were ethical) contrast infants who

were not on the diet for the first 3 months of their lives with those same infants after they were put on the diet starting in the 4th month. Neither of these approximations is a true counterfactual. In the first case, the individual infants in the treatment condition are different from those in the comparison condition; in the second case, the identities are the same, but time has passed and many changes other than the treatment have occurred to the infants (including permanent damage done by phenylalanine during the first 3 months of life). So two central tasks in experimental design are creating a high-quality but necessarily imperfect source of counterfactual inference and understanding how this source differs from the treatment condition.

This counterfactual reasoning is fundamentally qualitative because causal inference, even in experiments, is fundamentally qualitative (Campbell, 1975; Shadish, 1995a; Shadish & Cook, 1999). However, some of these points have been formalized by statisticians into a special case that is sometimes called Rubin's Causal Model (Holland, 1986; Rubin, 1974, 1977, 1978, 1986). This book is not about statistics, so we do not describe that model in detail (West, Biesanz, & Pitts [2000] do so and relate it to the Campbell tradition). A primary emphasis of Rubin's model is the analysis of cause in experiments, and its basic premises are consistent with those of this book.[2] Rubin's model has also been widely used to analyze causal inference in case-control studies in public health and medicine (Holland & Rubin, 1988), in path analysis in sociology (Holland, 1986), and in a paradox that Lord (1967) introduced into psychology (Holland & Rubin, 1983); and it has generated many statistical innovations that we cover later in this book. It is new enough that critiques of it are just now beginning to appear (e.g., Dawid, 2000; Pearl, 2000). What is clear, however, is that Rubin's is a very general model with obvious and subtle implications. Both it and the critiques of it are required material for advanced students and scholars of cause-probing methods.

### *Causal Relationship*

How do we know if cause and effect are related? In a classic analysis formalized by the 19th-century philosopher John Stuart Mill, a causal relationship exists if (1) the cause preceded the effect, (2) the cause was related to the effect, and (3) we can find no plausible alternative explanation for the effect other than the cause. These three characteristics mirror what happens in experiments in which (1) we manipulate the presumed cause and observe an outcome afterward; (2) we see whether variation in the cause is related to variation in the effect; and (3) we use various methods during the experiment to reduce the plausibility of other explanations for the effect, along with ancillary methods to explore the plausibility of those we cannot rule out (most of this book is about methods for doing this).

2. However, Rubin's model is not intended to say much about the matters of causal generalization that we address in this book.

Hence experiments are well-suited to studying causal relationships. No other scientific method regularly matches the characteristics of causal relationships so well. Mill's analysis also points to the weakness of other methods. In many correlational studies, for example, it is impossible to know which of two variables came first, so defending a causal relationship between them is precarious. Understanding this **logic of causal relationships** and how its key terms, such as cause and effect, are defined helps researchers to critique cause-probing studies.

## Causation, Correlation, and Confounds

A well-known maxim in research is: *Correlation does not prove causation.* This is so because we may not know which variable came first nor whether alternative explanations for the presumed effect exist. For example, suppose income and education are correlated. Do you have to have a high income before you can afford to pay for education, or do you first have to get a good education before you can get a better paying job? Each possibility may be true, and so both need investigation. But until those investigations are completed and evaluated by the scholarly community, a simple **correlation** does not indicate which variable came first. Correlations also do little to rule out alternative explanations for a relationship between two variables such as education and income. That relationship may not be causal at all but rather due to a third variable (often called a **confound**), such as intelligence or family socioeconomic status, that causes both high education and high income. For example, if high intelligence causes success in education and on the job, then intelligent people would have correlated education and incomes, not because education causes income (or vice versa) but because both would be caused by intelligence. Thus a central task in the study of experiments is identifying the different kinds of confounds that can operate in a particular research area and understanding the strengths and weaknesses associated with various ways of dealing with them.

## Manipulable and Nonmanipulable Causes

In the intuitive understanding of experimentation that most people have, it makes sense to say, "Let's see what happens if we require welfare recipients to work"; but it makes no sense to say, "Let's see what happens if I change this adult male into a three-year-old girl." And so it is also in scientific experiments. Experiments explore the effects of things that can be *manipulated,* such as the dose of a medicine, the amount of a welfare check, the kind or amount of psychotherapy, or the number of children in a classroom. Nonmanipulable events (e.g., the explosion of a supernova) or attributes (e.g., people's ages, their raw genetic material, or their biological sex) cannot be causes in experiments because we cannot deliberately vary them to see what then happens. Consequently, most scientists and philosophers agree that it is much harder to discover the effects of nonmanipulable causes.

To be clear, we are not arguing that *all* causes must be manipulable—only that *experimental* causes must be so. Many variables that we correctly think of as causes are not directly manipulable. Thus it is well established that a genetic defect causes PKU even though that defect is not directly manipulable. We can investigate such causes indirectly in nonexperimental studies or even in experiments by manipulating biological processes that prevent the gene from exerting its influence, as through the use of diet to inhibit the gene's biological consequences. Both the nonmanipulable gene and the manipulable diet can be viewed as causes—both covary with PKU-based retardation, both precede the retardation, and it is possible to explore other explanations for the gene's and the diet's effects on cognitive functioning. However, investigating the manipulable diet as a cause has two important advantages over considering the nonmanipulable genetic problem as a cause. First, only the diet provides a direct action to solve the problem; and second, we will see that studying manipulable agents allows a higher quality source of counterfactual inference through such methods as random assignment. When individuals with the nonmanipulable genetic problem are compared with persons without it, the latter are likely to be different from the former in many ways other than the genetic defect. So the counterfactual inference about what would have happened to those with the PKU genetic defect is much more difficult to make.

Nonetheless, nonmanipulable causes should be studied using whatever means are available and seem useful. This is true because such causes eventually help us to find manipulable agents that can then be used to ameliorate the problem at hand. The PKU example illustrates this. Medical researchers did not discover how to treat PKU effectively by first trying different diets with retarded children. They first discovered the nonmanipulable biological features of retarded children affected with PKU, finding abnormally high levels of phenylalanine and its associated metabolic and genetic problems in those children. Those findings pointed in certain ameliorative directions and away from others, leading scientists to experiment with treatments they thought might be effective and practical. Thus the new diet resulted from a sequence of studies with different immediate purposes, with different forms, and with varying degrees of uncertainty reduction. Some were experimental, but others were not.

Further, **analogue experiments** can sometimes be done on nonmanipulable causes, that is, experiments that manipulate an agent that is similar to the cause of interest. Thus we cannot change a person's race, but we can chemically induce skin pigmentation changes in volunteer individuals—though such analogues do not match the reality of being Black every day and everywhere for an entire life. Similarly, past events, which are normally nonmanipulable, sometimes constitute a **natural experiment** that may even have been randomized, as when the 1970 Vietnam-era draft lottery was used to investigate a variety of outcomes (e.g., Angrist, Imbens, & Rubin, 1996a; Notz, Staw, & Cook, 1971).

Although experimenting on manipulable causes makes the job of discovering their effects easier, experiments are far from perfect means of investigating causes.

Sometimes experiments modify the conditions in which testing occurs in a way that reduces the fit between those conditions and the situation to which the results are to be generalized. Also, knowledge of the effects of manipulable causes tells nothing about how and why those effects occur. Nor do experiments answer many other questions relevant to the real world—for example, which questions are worth asking, how strong the need for treatment is, how a cause is distributed through society, whether the treatment is implemented with theoretical fidelity, and what value should be attached to the experimental results.

In addition, in experiments, we first manipulate a treatment and only then observe its effects; but in some other studies we first observe an effect, such as AIDS, and then search for its cause, whether manipulable or not. Experiments cannot help us with that search. Scriven (1976) likens such searches to detective work in which a crime has been committed (e.g., a robbery), the detectives observe a particular pattern of evidence surrounding the crime (e.g., the robber wore a baseball cap and a distinct jacket and used a certain kind of gun), and then the detectives search for criminals whose known method of operating (their **modus operandi** or **m.o.**) includes this pattern. A criminal whose m.o. fits that pattern of evidence then becomes a suspect to be investigated further. Epidemiologists use a similar method, the case-control design (Ahlbom & Norell, 1990), in which they observe a particular health outcome (e.g., an increase in brain tumors) that is not seen in another group and then attempt to identify associated causes (e.g., increased cell phone use). Experiments do not aspire to answer all the kinds of questions, not even all the types of causal questions, that social scientists ask.

## Causal Description and Causal Explanation

The unique strength of experimentation is in describing the consequences attributable to deliberately varying a treatment. We call this **causal description.** In contrast, experiments do less well in clarifying the mechanisms through which and the conditions under which that causal relationship holds—what we call **causal explanation.** For example, most children very quickly learn the descriptive causal relationship between flicking a light switch and obtaining illumination in a room. However, few children (or even adults) can fully explain *why* that light goes on. To do so, they would have to decompose the treatment (the act of flicking a light switch) into its causally efficacious features (e.g., closing an insulated circuit) and its nonessential features (e.g., whether the switch is thrown by hand or a motion detector). They would have to do the same for the effect (either incandescent or fluorescent light can be produced, but light will still be produced whether the light fixture is recessed or not). For full explanation, they would then have to show how the causally efficacious parts of the treatment influence the causally affected parts of the outcome through identified mediating processes (e.g., the

passage of electricity through the circuit, the excitation of photons).[3] Clearly, the cause of the light going on is a complex cluster of many factors. For those philosophers who equate cause with identifying that constellation of variables that necessarily, inevitably, and infallibly results in the effect (Beauchamp, 1974), talk of cause is not warranted until everything of relevance is known. For them, there is no causal description without causal explanation. Whatever the philosophic merits of their position, though, it is not practical to expect much current social science to achieve such complete explanation.

The practical importance of causal explanation is brought home when the switch fails to make the light go on and when replacing the light bulb (another easily learned manipulation) fails to solve the problem. Explanatory knowledge then offers clues about how to fix the problem—for example, by detecting and repairing a short circuit. Or if we wanted to create illumination in a place without lights and we had explanatory knowledge, we would know exactly which features of the cause-and-effect relationship are essential to create light and which are irrelevant. Our explanation might tell us that there must be a source of electricity but that that source could take several different molar forms, such as a battery, a generator, a windmill, or a solar array. There must also be a switch mechanism to close a circuit, but this could also take many forms, including the touching of two bare wires or even a motion detector that trips the switch when someone enters the room. So causal explanation is an important route to the generalization of causal descriptions because it tells us which features of the causal relationship are essential to transfer to other situations.

This benefit of causal explanation helps elucidate its priority and prestige in all sciences and helps explain why, once a novel and important causal relationship is discovered, the bulk of basic scientific effort turns toward explaining why and how it happens. Usually, this involves decomposing the cause into its causally effective parts, decomposing the effects into its causally affected parts, and identifying the processes through which the effective causal parts influence the causally affected outcome parts.

These examples also show the close parallel between descriptive and explanatory causation and **molar** and **molecular** causation.[4] Descriptive causation usually concerns simple bivariate relationships between molar treatments and molar outcomes, molar here referring to a package that consists of many different parts. For instance, we may find that psychotherapy decreases depression, a simple descriptive causal relationship between a molar treatment package and a molar outcome. However, psychotherapy consists of such parts as verbal interactions, **placebo-**

3. However, the full explanation a physicist would offer might be quite different from this electrician's explanation, perhaps invoking the behavior of subparticles. This difference indicates just how complicated is the notion of explanation and how it can quickly become quite complex once one shifts levels of analysis.

4. By *molar*, we mean something taken as a whole rather than in parts. An analogy is to physics, in which molar might refer to the properties or motions of masses, as distinguished from those of molecules or atoms that make up those masses.

generating procedures, setting characteristics, time constraints, and payment for services. Similarly, many depression measures consist of items pertaining to the physiological, cognitive, and affective aspects of depression. Explanatory causation breaks these molar causes and effects into their molecular parts so as to learn, say, that the verbal interactions and the placebo features of therapy both cause changes in the cognitive symptoms of depression, but that payment for services does not do so even though it is part of the molar treatment package.

If experiments are less able to provide this highly-prized explanatory causal knowledge, why are experiments so central to science, especially to basic social science, in which theory and explanation are often the coin of the realm? The answer is that the dichotomy between descriptive and explanatory causation is less clear in scientific practice than in abstract discussions about causation. First, many causal explanations consist of chains of descriptive causal links in which one event causes the next. Experiments help to test the links in each chain. Second, experiments help distinguish between the validity of competing explanatory theories, for example, by testing competing mediating links proposed by those theories. Third, some experiments test whether a descriptive causal relationship varies in strength or direction under Condition A versus Condition B (then the condition is a **moderator** variable that explains the conditions under which the effect holds). Fourth, some experiments add quantitative or qualitative observations of the links in the explanatory chain (**mediator** variables) to generate and study explanations for the descriptive causal effect.

Experiments are also prized in applied areas of social science, in which the identification of practical solutions to social problems has as great or even greater priority than explanations of those solutions. After all, explanation is not always required for identifying practical solutions. Lewontin (1997) makes this point about the Human Genome Project, a coordinated multibillion-dollar research program to map the human genome that it is hoped eventually will clarify the genetic causes of diseases. Lewontin is skeptical about aspects of this search:

> What is involved here is the difference between explanation and intervention. Many disorders can be *explained* by the failure of the organism to make a normal protein, a failure that is the consequence of a gene mutation. But *intervention* requires that the normal protein be provided at the right place in the right cells, at the right time and in the right amount, or else that an alternative way be found to provide normal cellular function. What is worse, it might even be necessary to keep the abnormal protein away from the cells at critical moments. None of these objectives is served by knowing the DNA sequence of the defective gene. (Lewontin, 1997, p. 29)

Practical applications are not immediately revealed by theoretical advance. Instead, to reveal them may take decades of follow-up work, including tests of simple descriptive causal relationships. The same point is illustrated by the cancer drug Endostatin, discussed earlier. Scientists knew the action of the drug occurred through cutting off tumor blood supplies; but to successfully use the drug to treat cancers in mice required administering it at the right place, angle, and depth, and those details were not part of the usual scientific explanation of the drug's effects.

In the end, then, causal descriptions and causal explanations are in delicate balance in experiments. What experiments do best is to improve causal descriptions; they do less well at explaining causal relationships. But most experiments can be designed to provide better explanations than is typically the case today. Further, in focusing on causal descriptions, experiments often investigate molar events that may be less strongly related to outcomes than are more molecular mediating processes, especially those processes that are closer to the outcome in the explanatory chain. However, many causal descriptions are still dependable and strong enough to be useful, to be worth making the building blocks around which important policies and theories are created. Just consider the dependability of such causal statements as that school desegregation causes white flight, or that outgroup threat causes ingroup cohesion, or that psychotherapy improves mental health, or that diet reduces the retardation due to PKU. Such dependable causal relationships are useful to policymakers, practitioners, and scientists alike.

## MODERN DESCRIPTIONS OF EXPERIMENTS

Some of the terms used in describing modern experimentation (see Table 1.1) are unique, clearly defined, and consistently used; others are blurred and inconsistently used. The common attribute in all experiments is control of treatment (though control can take many different forms). So Mosteller (1990, p. 225) writes, "In an experiment the investigator controls the application of the treatment"; and Yaremko, Harari, Harrison, and Lynn (1986, p. 72) write, "one or more independent variables are manipulated to observe their effects on one or more dependent variables." However, over time many different experimental subtypes have developed in response to the needs and histories of different sciences (Winston, 1990; Winston & Blais, 1996).

**TABLE 1.1 The Vocabulary of Experiments**

| |
|---|
| *Experiment:* A study in which an intervention is deliberately introduced to observe its effects. |
| *Randomized Experiment:* An experiment in which units are assigned to receive the treatment or an alternative condition by a random process such as the toss of a coin or a table of random numbers. |
| *Quasi-Experiment:* An experiment in which units are not assigned to conditions randomly. |
| *Natural Experiment:* Not really an experiment because the cause usually cannot be manipulated; a study that contrasts a naturally occurring event such as an earthquake with a comparison condition. |
| *Correlational Study:* Usually synonymous with nonexperimental or observational study; a study that simply observes the size and direction of a relationship among variables. |

## Randomized Experiment

The most clearly described variant is the **randomized experiment,** widely credited to Sir Ronald Fisher (1925, 1926). It was first used in agriculture but later spread to other topic areas because it promised control over extraneous sources of variation without requiring the physical isolation of the laboratory. Its distinguishing feature is clear and important—that the various treatments being contrasted (including no treatment at all) are assigned to experimental **units**[5] by chance, for example, by coin toss or use of a table of random numbers. If implemented correctly, random assignment creates two or more groups of units that are probabilistically similar to each other on the average.[6] Hence, any outcome differences that are observed between those groups at the end of a study are likely to be due to treatment, not to differences between the groups that already existed at the start of the study. Further, when certain assumptions are met, the randomized experiment yields an estimate of the size of a treatment effect that has desirable statistical properties, along with estimates of the probability that the true effect falls within a defined confidence interval. These features of experiments are so highly prized that in a research area such as medicine the randomized experiment is often referred to as the gold standard for treatment outcome research.[7]

Closely related to the randomized experiment is a more ambiguous and inconsistently used term, **true experiment.** Some authors use it synonymously with randomized experiment (Rosenthal & Rosnow, 1991). Others use it more generally to refer to any study in which an **independent variable** is deliberately manipulated (Yaremko et al., 1986) and a **dependent variable** is assessed. We shall not use the term at all given its ambiguity and given that the modifier *true* seems to imply restricted claims to a single correct experimental method.

## Quasi-Experiment

Much of this book focuses on a class of designs that Campbell and Stanley (1963) popularized as **quasi-experiments.**[8] Quasi-experiments share with all other

5. Units can be people, animals, time periods, institutions, or almost anything else. Typically in field experimentation they are people or some aggregate of people, such as classrooms or work sites. In addition, a little thought shows that random assignment of units to treatments is the same as assignment of treatments to units, so these phrases are frequently used interchangeably.

6. The word *probabilistically* is crucial, as is explained in more detail in Chapter 8.

7. Although the term *randomized experiment* is used this way consistently across many fields and in this book, statisticians sometimes use the closely related term *random experiment* in a different way to indicate experiments for which the outcome cannot be predicted with certainty (e.g., Hogg & Tanis, 1988).

8. Campbell (1957) first called these compromise designs but changed terminology very quickly; Rosenbaum (1995a) and Cochran (1965) refer to these as **observational studies,** a term we avoid because many people use it to refer to correlational or nonexperimental studies, as well. Greenberg and Shroder (1997) use *quasi-experiment* to refer to studies that randomly assign groups (e.g., communities) to conditions, but we would consider these group-randomized experiments (Murray, 1998).

experiments a similar purpose—to test descriptive causal hypotheses about manipulable causes—as well as many structural details, such as the frequent presence of control groups and pretest measures, to support a counterfactual inference about what would have happened in the absence of treatment. But, by definition, quasi-experiments lack random assignment. Assignment to conditions is by means of **self-selection**, by which units choose treatment for themselves, or by means of **administrator selection**, by which teachers, bureaucrats, legislators, therapists, physicians, or others decide which persons should get which treatment. However, researchers who use quasi-experiments may still have considerable control over selecting and scheduling measures, over how nonrandom assignment is executed, over the kinds of **comparison groups** with which treatment groups are compared, and over some aspects of how treatment is scheduled. As Campbell and Stanley note:

> There are many natural social settings in which the research person can introduce something like experimental design into his scheduling of data collection procedures (e.g., the *when* and *to whom* of measurement), even though he lacks the full control over the scheduling of experimental stimuli (the *when* and *to whom* of exposure and the ability to randomize exposures) which makes a true experiment possible. Collectively, such situations can be regarded as quasi-experimental designs. (Campbell & Stanley, 1963, p. 34)

In quasi-experiments, the cause is manipulable and occurs before the effect is measured. However, quasi-experimental design features usually create less compelling support for counterfactual inferences. For example, quasi-experimental control groups may differ from the treatment condition in many systematic (nonrandom) ways other than the presence of the treatment. Many of these ways could be alternative explanations for the observed effect, and so researchers have to worry about ruling them out in order to get a more valid estimate of the treatment effect. By contrast, with random assignment the researcher does not have to think *as much* about all these alternative explanations. If correctly done, random assignment makes most of the alternatives less likely as causes of the observed treatment effect at the start of the study.

In quasi-experiments, the researcher has to enumerate alternative explanations one by one, decide which are plausible, and then use logic, design, and measurement to assess whether each one is operating in a way that might explain any observed effect. The difficulties are that these alternative explanations are never completely enumerable in advance, that some of them are particular to the context being studied, and that the methods needed to eliminate them from contention will vary from alternative to alternative and from study to study. For example, suppose two nonrandomly formed groups of children are studied, a volunteer **treatment group** that gets a new reading program and a control group of nonvolunteers who do not get it. If the treatment group does better, is it because of treatment or because the cognitive development of the volunteers was increasing more rapidly even before treatment began? (In a randomized experiment, maturation rates would

have been probabilistically equal in both groups.) To assess this alternative, the researcher might add multiple pretests to reveal maturational trend before the treatment, and then compare that trend with the trend after treatment.

Another alternative explanation might be that the nonrandom control group included more disadvantaged children who had less access to books in their homes or who had parents who read to them less often. (In a randomized experiment, both groups would have had similar proportions of such children.) To assess this alternative, the experimenter may measure the number of books at home, parental time spent reading to children, and perhaps trips to libraries. Then the researcher would see if these variables differed across treatment and control groups in the hypothesized direction that could explain the observed treatment effect. Obviously, as the number of plausible alternative explanations increases, the design of the quasi-experiment becomes more intellectually demanding and complex—especially because we are never certain we have identified all the alternative explanations. The efforts of the quasi-experimenter start to look like attempts to bandage a wound that would have been less severe if random assignment had been used initially.

The ruling out of alternative hypotheses is closely related to a falsificationist logic popularized by Popper (1959). Popper noted how hard it is to be sure that a general conclusion (e.g., all swans are white) is correct based on a limited set of observations (e.g., all the swans I've seen were white). After all, future observations may change (e.g., someday I may see a black swan). So **confirmation** is logically difficult. By contrast, observing a disconfirming instance (e.g., a black swan) is sufficient, in Popper's view, to falsify the general conclusion that all swans are white. Accordingly, Popper urged scientists to try deliberately to falsify the conclusions they wish to draw rather than only to seek information corroborating them. Conclusions that withstand **falsification** are retained in scientific books or journals and treated as plausible until better evidence comes along. Quasi-experimentation is falsificationist in that it requires experimenters to identify a causal claim and then to generate and examine plausible alternative explanations that might falsify the claim.

However, such falsification can never be as definitive as Popper hoped. Kuhn (1962) pointed out that falsification depends on two assumptions that can never be fully tested. The first is that the causal claim is perfectly specified. But that is never the case. So many features of both the claim and the test of the claim are debatable—for example, which outcome is of interest, how it is measured, the conditions of treatment, who needs treatment, and all the many other decisions that researchers must make in testing causal relationships. As a result, disconfirmation often leads theorists to respecify part of their causal theories. For example, they might now specify novel conditions that must hold for their theory to be true and that were derived from the apparently disconfirming observations. Second, falsification requires measures that are perfectly valid reflections of the theory being tested. However, most philosophers maintain that all observation is theory-laden. It is laden both with intellectual nuances specific to the partially

unique scientific understandings of the theory held by the individual or group devising the test and also with the experimenters' extrascientific wishes, hopes, aspirations, and broadly shared cultural assumptions and understandings. If measures are not independent of theories, how can they provide independent theory tests, including tests of causal theories? If the possibility of theory-neutral observations is denied, with them disappears the possibility of definitive knowledge both of what seems to confirm a causal claim and of what seems to disconfirm it.

Nonetheless, a fallibilist version of falsification is possible. It argues that studies of causal hypotheses can still usefully improve understanding of general trends despite ignorance of all the contingencies that might pertain to those trends. It argues that causal studies are useful even if we have to respecify the initial hypothesis repeatedly to accommodate new contingencies and new understandings. After all, those respecifications are usually minor in scope; they rarely involve wholesale overthrowing of general trends in favor of completely opposite trends. Fallibilist falsification also assumes that theory-neutral observation is impossible but that observations can approach a more factlike status when they have been repeatedly made across different theoretical conceptions of a construct, across multiple kinds of measurements, and at multiple times. It also assumes that observations are imbued with multiple theories, not just one, and that different operational procedures do not share the same multiple theories. As a result, observations that repeatedly occur despite different theories being built into them have a special factlike status even if they can never be fully justified as completely theory-neutral facts. In summary, then, fallible falsification is more than just seeing whether observations disconfirm a prediction. It involves discovering and judging the worth of ancillary assumptions about the restricted specificity of the causal hypothesis under test and also about the heterogeneity of theories, viewpoints, settings, and times built into the measures of the cause and effect and of any contingencies modifying their relationship.

It is neither feasible nor desirable to rule out all *possible* alternative interpretations of a causal relationship. Instead, only *plausible* alternatives constitute the major focus. This serves partly to keep matters tractable because the number of possible alternatives is endless. It also recognizes that many alternatives have no serious empirical or experiential support and so do not warrant special attention. However, the lack of support can sometimes be deceiving. For example, the cause of stomach ulcers was long thought to be a combination of lifestyle (e.g., stress) and excess acid production. Few scientists seriously thought that ulcers were caused by a pathogen (e.g., virus, germ, bacteria) because it was assumed that an acid-filled stomach would destroy all living organisms. However, in 1982 Australian researchers Barry Marshall and Robin Warren discovered spiral-shaped bacteria, later named *Helicobacter pylori* (*H. pylori*), in ulcer patients' stomachs. With this discovery, the previously possible but implausible became plausible. By 1994, a U.S. National Institutes of Health Consensus Development Conference concluded that *H. pylori* was the major cause of most peptic ulcers. So labeling ri-

val hypotheses as plausible depends not just on what is logically possible but on social consensus, shared experience and, empirical data.

Because such factors are often context specific, different substantive areas develop their own lore about which alternatives are important enough to need to be controlled, even developing their own methods for doing so. In early psychology, for example, a control group with pretest observations was invented to control for the plausible alternative explanation that, by giving practice in answering test content, pretests would produce gains in performance even in the absence of a treatment effect (Coover & Angell, 1907). Thus the focus on plausibility is a two-edged sword: it reduces the range of alternatives to be considered in quasi-experimental work, yet it also leaves the resulting causal inference vulnerable to the discovery that an implausible-seeming alternative may later emerge as a likely causal agent.

## Natural Experiment

The term *natural experiment* describes a naturally-occurring contrast between a treatment and a comparison condition (Fagan, 1990; Meyer, 1995; Zeisel, 1973). Often the treatments are not even potentially manipulable, as when researchers retrospectively examined whether earthquakes in California caused drops in property values (Brunette, 1995; Murdoch, Singh, & Thayer, 1993). Yet plausible causal inferences about the effects of earthquakes are easy to construct and defend. After all, the earthquakes occurred before the observations on property values, and it is easy to see whether earthquakes are related to property values. A useful source of counterfactual inference can be constructed by examining property values in the same locale before the earthquake or by studying similar locales that did not experience an earthquake during the same time. If property values dropped right after the earthquake in the earthquake condition but not in the comparison condition, it is difficult to find an alternative explanation for that drop.

Natural experiments have recently gained a high profile in economics. Before the 1990s economists had great faith in their ability to produce valid causal inferences through statistical adjustments for initial nonequivalence between treatment and control groups. But two studies on the effects of job training programs showed that those adjustments produced estimates that were not close to those generated from a randomized experiment and were unstable across tests of the model's sensitivity (Fraker & Maynard, 1987; LaLonde, 1986). Hence, in their search for alternative methods, many economists came to do natural experiments, such as the economic study of the effects that occurred in the Miami job market when many prisoners were released from Cuban jails and allowed to come to the United States (Card, 1990). They assume that the release of prisoners (or the timing of an earthquake) is independent of the ongoing processes that usually affect unemployment rates (or housing values). Later we explore the validity of this assumption—of its desirability there can be little question.

### Nonexperimental Designs

The terms **correlational design**, **passive observational design**, and **nonexperimental design** refer to situations in which a presumed cause and effect are identified and measured but in which other structural features of experiments are missing. Random assignment is not part of the design, nor are such **design elements** as pretests and control groups from which researchers might construct a useful counterfactual inference. Instead, reliance is placed on measuring alternative explanations individually and then statistically controlling for them. In cross-sectional studies in which all the data are gathered on the respondents at one time, the researcher may not even know if the cause precedes the effect. When these studies are used for causal purposes, the missing design features can be problematic unless much is already known about which alternative interpretations are plausible, unless those that are plausible can be validly measured, and unless the substantive model used for statistical adjustment is well-specified. These are difficult conditions to meet in the real world of research practice, and therefore many commentators doubt the potential of such designs to support strong causal inferences in most cases.

## EXPERIMENTS AND THE GENERALIZATION OF CAUSAL CONNECTIONS

The strength of experimentation is its ability to illuminate causal inference. The weakness of experimentation is doubt about the extent to which that causal relationship generalizes. We hope that an innovative feature of this book is its focus on generalization. Here we introduce the general issues that are expanded in later chapters.

### Most Experiments Are Highly Local But Have General Aspirations

Most experiments are highly localized and particularistic. They are almost always conducted in a restricted range of settings, often just one, with a particular version of one type of treatment rather than, say, a sample of all possible versions. Usually, they have several measures—each with theoretical assumptions that are different from those present in other measures—but far from a complete set of all possible measures. Each experiment nearly always uses a convenient sample of people rather than one that reflects a well-described population; and it will inevitably be conducted at a particular point in time that rapidly becomes history.

Yet readers of experimental results are rarely concerned with what happened in that particular, past, local study. Rather, they usually aim to learn either about theoretical constructs of interest or about a larger policy. Theorists often want to

connect experimental results to theories with broad conceptual applicability, which requires generalization at the linguistic level of constructs rather than at the level of the operations used to represent these constructs in a given experiment. They nearly always want to generalize to more people and settings than are represented in a single experiment. Indeed, the value assigned to a substantive theory usually depends on how broad a range of phenomena the theory covers. Similarly, policymakers may be interested in whether a causal relationship would hold (probabilistically) across the many sites at which it would be implemented as a policy, an inference that requires generalization beyond the original experimental study context. Indeed, all human beings probably value the perceptual and cognitive stability that is fostered by generalizations. Otherwise, the world might appear as a buzzing cacophony of isolated instances requiring constant cognitive processing that would overwhelm our limited capacities.

In defining generalization as a problem, we do not assume that more broadly applicable results are always more desirable (Greenwood, 1989). For example, physicists who use particle accelerators to discover new elements may not expect that it would be desirable to introduce such elements into the world. Similarly, social scientists sometimes aim to demonstrate that an effect is possible and to understand its mechanisms without expecting that the effect can be produced more generally. For instance, when a "sleeper effect" occurs in an attitude change study involving persuasive communications, the implication is that change is manifest after a time delay but not immediately so. The circumstances under which this effect occurs turn out to be quite limited and unlikely to be of any general interest other than to show that the theory predicting it (and many other ancillary theories) may not be wrong (Cook, Gruder, Hennigan & Flay, 1979). Experiments that demonstrate limited generalization may be just as valuable as those that demonstrate broad generalization.

Nonetheless, a conflict seems to exist between the localized nature of the causal knowledge that individual experiments provide and the more generalized causal goals that research aspires to attain. Cronbach and his colleagues (Cronbach et al., 1980; Cronbach, 1982) have made this argument most forcefully, and their works have contributed much to our thinking about causal generalization. Cronbach noted that each experiment consists of *units* that receive the experiences being contrasted, of the *treatments* themselves, of *observations* made on the units, and of the *settings* in which the study is conducted. Taking the first letter from each of these four words, he defined the acronym *utos* to refer to the "instances on which data are collected" (Cronbach, 1982, p. 78)—to the actual people, treatments, measures, and settings that were sampled in the experiment. He then defined two problems of generalization: (1) generalizing to the "domain about which [the] question is asked" (p. 79), which he called *UTOS;* and (2) generalizing to "units, treatments, variables, and settings not directly observed" (p. 83), which he called **UTOS.*[9]

9. We oversimplify Cronbach's presentation here for pedagogical reasons. For example, Cronbach only used capital *S*, not small *s*, so that his system referred only to *utoS*, not *utos*. He offered diverse and not always consistent definitions of *UTOS* and **UTOS*, in particular. And he does not use the word *generalization* in the same broad way we do here.

Our theory of causal generalization, outlined below and presented in more detail in Chapters 11 through 13, melds Cronbach's thinking with our own ideas about generalization from previous works (Cook, 1990, 1991; Cook & Campbell, 1979), creating a theory that is different in modest ways from both of these predecessors. Our theory is influenced by Cronbach's work in two ways. First, we follow him by describing experiments consistently throughout this book as consisting of the elements of units, treatments, observations, and settings,[10] though we frequently substitute *persons* for *units* given that most field experimentation is conducted with humans as participants. We also often substitute *outcome* for *observations* given the centrality of observations about outcome when examining causal relationships. Second, we acknowledge that researchers are often interested in two kinds of generalization about each of these five elements, and that these two types are inspired by, but not identical to, the two kinds of generalization that Cronbach defined. We call these **construct validity** generalizations (inferences about the constructs that research operations represent) and **external validity** generalizations (inferences about whether the causal relationship holds over variation in persons, settings, treatment, and measurement variables).

## Construct Validity: Causal Generalization as Representation

The first causal generalization problem concerns how to go from the particular units, treatments, observations, and settings on which data are collected to the higher order constructs these instances represent. These constructs are almost always couched in terms that are more abstract than the particular instances sampled in an experiment. The labels may pertain to the individual elements of the experiment (e.g., is the outcome measured by a given test best described as intelligence or as achievement?). Or the labels may pertain to the nature of relationships among elements, including causal relationships, as when cancer treatments are classified as cytotoxic or cytostatic depending on whether they kill tumor cells directly or delay tumor growth by modulating their environment. Consider a randomized experiment by Fortin and Kirouac (1976). The treatment was a brief educational course administered by several nurses, who gave a tour of their hospital and covered some basic facts about surgery with individuals who were to have elective abdominal or thoracic surgery 15 to 20 days later in a single Montreal hospital. Ten specific outcome measures were used after the surgery, such as an activities of daily living scale and a count of the analgesics used to control pain. Now compare this study with its likely target constructs—whether

10. We occasionally refer to time as a separate feature of experiments, following Campbell (1957) and Cook and Campbell (1979), because time can cut across the other factors independently. Cronbach did not include time in his notational system, instead incorporating time into treatment (e.g., the scheduling of treatment), observations (e.g., when measures are administered), or setting (e.g., the historical context of the experiment).

patient education (the target cause) promotes physical recovery (the target effect) among surgical patients (the target population of units) in hospitals (the target universe of settings). Another example occurs in basic research, in which the question frequently arises as to whether the actual manipulations and measures used in an experiment really tap into the specific cause and effect constructs specified by the theory. One way to dismiss an empirical challenge to a theory is simply to make the case that the data do not really represent the concepts as they are specified in the theory.

Empirical results often force researchers to change their initial understanding of what the domain under study is. Sometimes the reconceptualization leads to a more restricted inference about what has been studied. Thus the planned causal agent in the Fortin and Kirouac (1976) study—*patient education*—might need to be respecified as *informational patient education* if the information component of the treatment proved to be causally related to recovery from surgery but the tour of the hospital did not. Conversely, data can sometimes lead researchers to think in terms of target constructs and categories that are more general than those with which they began a research program. Thus the creative analyst of patient education studies might surmise that the treatment is a subclass of interventions that function by increasing "perceived control" or that recovery from surgery can be treated as a subclass of "personal coping." Subsequent readers of the study can even add their own interpretations, perhaps claiming that perceived control is really just a special case of the even more general self-efficacy construct. There is a subtle interplay over time among the original categories the researcher intended to represent, the study as it was actually conducted, the study results, and subsequent interpretations. This interplay can change the researcher's thinking about what the study particulars actually achieved at a more conceptual level, as can feedback from readers. But whatever reconceptualizations occur, the first problem of causal generalization is always the same: How can we generalize from a sample of instances and the data patterns associated with them to the particular target constructs they represent?

## External Validity: Causal Generalization as Extrapolation

The second problem of generalization is to infer whether a causal relationship holds over variations in persons, settings, treatments, and outcomes. For example, someone reading the results of an experiment on the effects of a kindergarten Head Start program on the subsequent grammar school reading test scores of poor African American children in Memphis during the 1980s may want to know if a program with partially overlapping cognitive and social development goals would be as effective in improving the mathematics test scores of poor Hispanic children in Dallas if this program were to be implemented tomorrow.

This example again reminds us that generalization is not a synonym for *broader* application. Here, generalization is from one city to another city and

from one kind of clientele to another kind, but there is no presumption that Dallas is somehow broader than Memphis or that Hispanic children constitute a broader population than African American children. Of course, some generalizations are from narrow to broad. For example, a researcher who randomly samples experimental participants from a national population may generalize (probabilistically) from the sample to all the other unstudied members of that same population. Indeed, that is the rationale for choosing random selection in the first place. Similarly, when policymakers consider whether Head Start should be continued on a national basis, they are not so interested in what happened in Memphis. They are more interested in what would happen on the average across the United States, as its many local programs still differ from each other despite efforts in the 1990s to standardize much of what happens to Head Start children and parents. But generalization can also go from the broad to the narrow. Cronbach (1982) gives the example of an experiment that studied differences between the performances of groups of students attending private and public schools. In this case, the concern of individual parents is to know which type of school is better for their particular child, not for the whole group. Whether from narrow to broad, broad to narrow, or across units at about the same level of aggregation, all these examples of external validity questions share the same need—to infer the extent to which the effect holds over variations in persons, settings, treatments, or outcomes.

## Approaches to Making Causal Generalizations

Whichever way the causal generalization issue is framed, experiments do not seem at first glance to be very useful. Almost invariably, a given experiment uses a limited set of operations to represent units, treatments, outcomes, and settings. This high degree of localization is not unique to the experiment; it also characterizes case studies, performance monitoring systems, and opportunistically-administered marketing questionnaires given to, say, a haphazard sample of respondents at local shopping centers (Shadish, 1995b). Even when questionnaires are administered to nationally representative samples, they are ideal for representing that particular population of persons but have little relevance to citizens outside of that nation. Moreover, responses may also vary by the setting in which the interview took place (a doorstep, a living room, or a work site), by the time of day at which it was administered, by how each question was framed, or by the particular race, age, and gender combination of interviewers. But the fact that the experiment is not alone in its vulnerability to generalization issues does not make it any less a problem. So what is it that justifies any belief that an experiment can achieve a better fit between the sampling particulars of a study and more general inferences to constructs or over variations in persons, settings, treatments, and outcomes?

### *Sampling and Causal Generalization*

The method most often recommended for achieving this close fit is the use of formal probability sampling of instances of units, treatments, observations, or settings (Rossi, Wright, & Anderson, 1983). This presupposes that we have clearly delineated populations of each and that we can sample with known probability from within each of these populations. In effect, this entails the random selection of instances, to be carefully distinguished from random assignment discussed earlier in this chapter. Random selection involves selecting cases by chance to represent that population, whereas random assignment involves assigning cases to multiple conditions.

In cause-probing research that is *not* experimental, random samples of individuals are often used. Large-scale longitudinal surveys such as the Panel Study of Income Dynamics or the National Longitudinal Survey are used to represent the population of the United States—or certain age brackets within it—and measures of potential causes and effects are then related to each other using time lags in measurement and statistical controls for group nonequivalence. All this is done in hopes of approximating what a randomized experiment achieves. However, cases of random selection from a broad population followed by random assignment from within this population are much rarer (see Chapter 12 for examples). Also rare are studies of random selection followed by a quality quasi-experiment. Such experiments require a high level of resources and a degree of logistical control that is rarely feasible, so many researchers prefer to rely on an implicit set of nonstatistical heuristics for generalization that we hope to make more explicit and systematic in this book.

Random selection occurs even more rarely with treatments, outcomes, and settings than with people. Consider the outcomes observed in an experiment. How often are they randomly sampled? We grant that the domain sampling model of classical test theory (Nunnally & Bernstein, 1994) assumes that the items used to measure a construct have been randomly sampled from a domain of all possible items. However, in actual experimental practice few researchers ever randomly sample items when constructing measures. Nor do they do so when choosing manipulations or settings. For instance, many settings will not agree to be sampled, and some of the settings that agree to be randomly sampled will almost certainly not agree to be randomly assigned to conditions. For treatments, no definitive list of possible treatments usually exists, as is most obvious in areas in which treatments are being discovered and developed rapidly, such as in AIDS research. In general, then, random sampling is always desirable, but it is only rarely and contingently feasible.

However, formal sampling methods are not the only option. Two informal, purposive sampling methods are sometimes useful—purposive sampling of heterogeneous instances and purposive sampling of typical instances. In the former case, the aim is to include instances chosen deliberately to reflect diversity on presumptively important dimensions, even though the sample is not formally random. In the latter

case, the aim is to explicate the kinds of units, treatments, observations, and settings to which one most wants to generalize and then to select at least one instance of each class that is impressionistically similar to the class mode. Although these purposive sampling methods are more practical than formal probability sampling, they are not backed by a statistical logic that justifies formal generalizations. Nonetheless, they are probably the most commonly used of all sampling methods for facilitating generalizations. A task we set ourselves in this book is to explicate such methods and to describe how they can be used more often than is the case today.

However, sampling methods of any kind are insufficient to solve either problem of generalization. Formal probability sampling requires specifying a target population from which sampling then takes place, but defining such populations is difficult for some targets of generalization such as treatments. Purposive sampling of heterogeneous instances is differentially feasible for different elements in a study; it is often more feasible to make measures diverse than it is to obtain diverse settings, for example. Purposive sampling of typical instances is often feasible when target modes, medians, or means are known, but it leaves questions about generalizations to a wider range than is typical. Besides, as Cronbach points out, most challenges to the causal generalization of an experiment typically emerge *after* a study is done. In such cases, sampling is relevant only if the instances in the original study were sampled diversely enough to promote responsible reanalyses of the data to see if a treatment effect holds across most or all of the targets about which generalization has been challenged. But packing so many sources of variation into a single experimental study is rarely practical and will almost certainly conflict with other goals of the experiment. Formal sampling methods usually offer only a limited solution to causal generalization problems. A theory of generalized causal inference needs additional tools.

### A Grounded Theory of Causal Generalization

Practicing scientists routinely make causal generalizations in their research, and they almost never use formal probability sampling when they do. In this book, we present a theory of causal generalization that is grounded in the actual practice of science (Matt, Cook, & Shadish, 2000). Although this theory was originally developed from ideas that were grounded in the construct and external validity literatures (Cook, 1990, 1991), we have since found that these ideas are common in a diverse literature about scientific generalizations (e.g., Abelson, 1995; Campbell & Fiske, 1959; Cronbach & Meehl, 1955; Davis, 1994; Locke, 1986; Medin, 1989; Messick, 1989, 1995; Rubins, 1994; Willner, 1991; Wilson, Hayward, Tunis, Bass, & Guyatt, 1995). We provide more details about this grounded theory in Chapters 11 through 13, but in brief it suggests that scientists make causal generalizations in their work by using five closely related principles:

1. *Surface Similarity.* They assess the apparent similarities between study operations and the prototypical characteristics of the target of generalization.

2. *Ruling Out Irrelevancies.* They identify those things that are irrelevant because they do not change a generalization.
3. *Making Discriminations.* They clarify key discriminations that limit generalization.
4. *Interpolation and Extrapolation.* They make interpolations to unsampled values within the range of the sampled instances and, much more difficult, they explore extrapolations beyond the sampled range.
5. *Causal Explanation.* They develop and test explanatory theories about the pattern of effects, causes, and mediational processes that are essential to the transfer of a causal relationship.

In this book, we want to show how scientists can and do use these five principles to draw generalized conclusions about a causal connection. Sometimes the conclusion is about the higher order constructs to use in describing an obtained connection at the sample level. In this sense, these five principles have analogues or parallels both in the construct validity literature (e.g., with construct content, with convergent and discriminant validity, and with the need for theoretical rationales for constructs) and in the cognitive science and philosophy literatures that study how people decide whether instances fall into a category (e.g., concerning the roles that prototypical characteristics and surface versus deep similarity play in determining category membership). But at other times, the conclusion about generalization refers to whether a connection holds broadly or narrowly over variations in persons, settings, treatments, or outcomes. Here, too, the principles have analogues or parallels that we can recognize from scientific theory and practice, as in the study of dose-response relationships (a form of interpolation-extrapolation) or the appeal to explanatory mechanisms in generalizing from animals to humans (a form of causal explanation).

Scientists use these five principles almost constantly during all phases of research. For example, when they read a published study and wonder if some variation on the study's particulars would work in their lab, they think about similarities of the published study to what they propose to do. When they conceptualize the new study, they anticipate how the instances they plan to study will match the prototypical features of the constructs about which they are curious. They may design their study on the assumption that certain variations will be irrelevant to it but that others will point to key discriminations over which the causal relationship does not hold or the very character of the constructs changes. They may include measures of key theoretical mechanisms to clarify how the intervention works. During data analysis, they test all these hypotheses and adjust their construct descriptions to match better what the data suggest happened in the study. The introduction section of their articles tries to convince the reader that the study bears on specific constructs, and the discussion sometimes speculates about how results might extrapolate to different units, treatments, outcomes, and settings.

Further, practicing scientists do all this not just with single studies that they read or conduct but also with multiple studies. They nearly always think about

how their own studies fit into a larger literature about both the constructs being measured and the variables that may or may not bound or explain a causal connection, often documenting this fit in the introduction to their study. And they apply all five principles when they conduct reviews of the literature, in which they make inferences about the kinds of generalizations that a body of research can support.

Throughout this book, and especially in Chapters 11 to 13, we provide more details about this grounded theory of causal generalization and about the scientific practices that it suggests. Adopting this grounded theory of generalization does not imply a rejection of formal probability sampling. Indeed, we recommend such sampling unambiguously when it is feasible, along with purposive sampling schemes to aid generalization when formal random selection methods cannot be implemented. But we also show that sampling is just one method that practicing scientists use to make causal generalizations, along with practical logic, application of diverse statistical methods, and use of features of design other than sampling.

## EXPERIMENTS AND METASCIENCE

Extensive philosophical debate sometimes surrounds experimentation. Here we briefly summarize some key features of these debates, and then we discuss some implications of these debates for experimentation. However, there is a sense in which all this philosophical debate is incidental to the practice of experimentation. Experimentation is as old as humanity itself, so it preceded humanity's philosophical efforts to understand causation and generalization by thousands of years. Even over just the past 400 years of scientific experimentation, we can see some constancy of experimental concept and method, whereas diverse philosophical conceptions of the experiment have come and gone. As Hacking (1983) said, "Experimentation has a life of its own" (p. 150). It has been one of science's most powerful methods for discovering descriptive causal relationships, and it has done so well in so many ways that its place in science is probably assured forever. To justify its practice today, a scientist need not resort to sophisticated philosophical reasoning about experimentation.

Nonetheless, it does help scientists to understand these philosophical debates. For example, previous distinctions in this chapter between molar and molecular causation, descriptive and explanatory cause, or probabilistic and deterministic causal inferences all help both philosophers and scientists to understand better both the purpose and the results of experiments (e.g., Bunge, 1959; Eells, 1991; Hart & Honore, 1985; Humphreys, 1989; Mackie, 1974; Salmon, 1984, 1989; Sobel, 1993; P. A. White, 1990). Here we focus on a different and broader set of critiques of science itself, not only from philosophy but also from the history, sociology, and psychology of science (see useful general reviews by Bechtel, 1988; H. I. Brown, 1977; Oldroyd, 1986). Some of these works have been explicitly about the nature of experimentation, seeking to create a justified role for it (e.g.,

Bhaskar, 1975; Campbell, 1982, 1988; Danziger, 1990; S. Drake, 1981; Gergen, 1973; Gholson, Shadish, Neimeyer, & Houts, 1989; Gooding, Pinch, & Schaffer, 1989b; Greenwood, 1989; Hacking, 1983; Latour, 1987; Latour & Woolgar, 1979; Morawski, 1988; Orne, 1962; R. Rosenthal, 1966; Shadish & Fuller, 1994; Shapin, 1994). These critiques help scientists to see some limits of experimentation in both science and society.

## The Kuhnian Critique

Kuhn (1962) described scientific revolutions as different and partly incommensurable paradigms that abruptly succeeded each other in time and in which the gradual accumulation of scientific knowledge was a chimera. Hanson (1958), Polanyi (1958), Popper (1959), Toulmin (1961), Feyerabend (1975), and Quine (1951, 1969) contributed to the critical momentum, in part by exposing the gross mistakes in logical positivism's attempt to build a philosophy of science based on reconstructing a successful science such as physics. All these critiques denied any firm foundations for scientific knowledge (so, by extension, experiments do not provide firm causal knowledge). The logical positivists hoped to achieve foundations on which to build knowledge by tying all theory tightly to theory-free observation through predicate logic. But this left out important scientific concepts that could not be tied tightly to observation; and it failed to recognize that all observations are impregnated with substantive and methodological theory, making it impossible to conduct theory-free tests.[11]

The impossibility of theory-neutral observation (often referred to as the Quine-Duhem thesis) implies that the results of any single test (and so any single experiment) are inevitably ambiguous. They could be disputed, for example, on grounds that the theoretical assumptions built into the outcome measure were wrong or that the study made a faulty assumption about how high a treatment dose was required to be effective. Some of these assumptions are small, easily detected, and correctable, such as when a voltmeter gives the wrong reading because the impedance of the voltage source was much higher than that of the meter (Wilson, 1952). But other assumptions are more paradigmlike, impregnating a theory so completely that other parts of the theory make no sense without them (e.g., the assumption that the earth is the center of the universe in pre-Galilean astronomy). Because the number of assumptions involved in any scientific test is very large, researchers can easily find some assumptions to fault or can even posit new

11. However, Holton (1986) reminds us not to overstate the reliance of positivists on empirical data: "Even the father of positivism, Auguste Comte, had written . . . that without a theory of some sort by which to link phenomena to some principles 'it would not only be impossible to combine the isolated observations and draw any useful conclusions, we would not even be able to remember them, and, for the most part, the fact would not be noticed by our eyes'" (p. 32). Similarly, Uebel (1992) provides a more detailed historical analysis of the protocol sentence debate in logical positivism, showing some surprisingly nonstereotypical positions held by key players such as Carnap.

assumptions (Mitroff & Fitzgerald, 1977). In this way, substantive theories are less testable than their authors originally conceived. How can a theory be tested if it is made of clay rather than granite?

For reasons we clarify later, this critique is more true of single studies and less true of programs of research. But even in the latter case, undetected constant biases can result in flawed inferences about cause and its generalization. As a result, no experiment is ever fully certain, and extrascientific beliefs and preferences always have room to influence the many discretionary judgments involved in all scientific belief.

## Modern Social Psychological Critiques

Sociologists working within traditions variously called social constructivism, epistemological relativism, and the strong program (e.g., Barnes, 1974; Bloor, 1976; Collins, 1981; Knorr-Cetina, 1981; Latour & Woolgar, 1979; Mulkay, 1979) have shown those extrascientific processes at work in science. Their empirical studies show that scientists often fail to adhere to norms commonly proposed as part of good science (e.g., objectivity, neutrality, sharing of information). They have also shown how that which comes to be reported as scientific knowledge is partly determined by social and psychological forces and partly by issues of economic and political power both within science and in the larger society—issues that are rarely mentioned in published research reports. The most extreme among these sociologists attributes *all* scientific knowledge to such extrascientific processes, claiming that "the natural world has a small or nonexistent role in the construction of scientific knowledge" (Collins, 1981, p. 3).

Collins does not deny *ontological realism,* that real entities exist in the world. Rather, he denies *epistemological (scientific) realism,* that whatever external reality may exist can constrain our scientific theories. For example, if atoms really exist, do they affect our scientific theories at all? If our theory postulates an atom, is it describing a real entity that exists roughly as we describe it? *Epistemological relativists* such as Collins respond negatively to both questions, believing that the most important influences in science are social, psychological, economic, and political, and that these might even be the only influences on scientific theories. This view is not widely endorsed outside a small group of sociologists, but it is a useful counterweight to naïve assumptions that scientific studies somehow directly reveal nature to us (an assumption we call *naïve realism*). The results of all studies, including experiments, are profoundly subject to these extrascientific influences, from their conception to reports of their results.

## Science and Trust

A standard image of the scientist is as a skeptic, a person who only trusts results that have been personally verified. Indeed, the scientific revolution of the 17th century

claimed that trust, particularly trust in authority and dogma, was antithetical to good science. Every authoritative assertion, every dogma, was to be open to question, and the job of science was to do that questioning.

That image is partly wrong. Any single scientific study is an exercise in trust (Pinch, 1986; Shapin, 1994). Studies trust the vast majority of already developed methods, findings, and concepts that they use when they test a new hypothesis. For example, statistical theories and methods are usually taken on faith rather than personally verified, as are measurement instruments. The ratio of trust to skepticism in any given study is more like 99% trust to 1% skepticism than the opposite. Even in lifelong programs of research, the single scientist trusts much more than he or she ever doubts. Indeed, thoroughgoing skepticism is probably impossible for the individual scientist, to judge from what we know of the psychology of science (Gholson et al., 1989; Shadish & Fuller, 1994). Finally, skepticism is not even an accurate characterization of past scientific revolutions; Shapin (1994) shows that the role of "gentlemanly trust" in 17th-century England was central to the establishment of experimental science. Trust pervades science, despite its rhetoric of skepticism.

## Implications for Experiments

The net result of these criticisms is a greater appreciation for the equivocality of all scientific knowledge. The experiment is not a clear window that reveals nature directly to us. To the contrary, experiments yield hypothetical and fallible knowledge that is often dependent on context and imbued with many unstated theoretical assumptions. Consequently, experimental results are partly relative to those assumptions and contexts and might well change with new assumptions or contexts. In this sense, all scientists are epistemological constructivists and relativists. The difference is whether they are strong or weak relativists. Strong relativists share Collins's position that only extrascientific factors influence our theories. Weak relativists believe that both the ontological world and the worlds of ideology, interests, values, hopes, and wishes play a role in the construction of scientific knowledge. Most practicing scientists, including ourselves, would probably describe themselves as ontological realists but weak epistemological relativists.[12] To the extent that experiments reveal nature to us, it is through a very clouded windowpane (Campbell, 1988).

Such counterweights to naïve views of experiments were badly needed. As recently as 30 years ago, the central role of the experiment in science was probably

12. If space permitted, we could extend this discussion to a host of other philosophical issues that have been raised about the experiment, such as its role in discovery versus confirmation, incorrect assertions that the experiment is tied to some specific philosophy such as logical positivism or pragmatism, and the various mistakes that are frequently made in such discussions (e.g., Campbell, 1982, 1988; Cook, 1991; Cook & Campbell, 1986; Shadish, 1995a).

taken more for granted than is the case today. For example, Campbell and Stanley (1963) described themselves as:

> committed to the experiment: as the only means for settling disputes regarding educational practice, as the only way of verifying educational improvements, and as the only way of establishing a cumulative tradition in which improvements can be introduced without the danger of a faddish discard of old wisdom in favor of inferior novelties. (p. 2)

Indeed, Hacking (1983) points out that "'experimental method' used to be just another name for scientific method" (p. 149); and experimentation was then a more fertile ground for examples illustrating basic philosophical issues than it was a source of contention itself.

Not so today. We now understand better that the experiment is a profoundly human endeavor, affected by all the same human foibles as any other human endeavor, though with well-developed procedures for partial control of some of the limitations that have been identified to date. Some of these limitations are common to all science, of course. For example, scientists tend to notice evidence that confirms their preferred hypotheses and to overlook contradictory evidence. They make routine cognitive errors of judgment and have limited capacity to process large amounts of information. They react to peer pressures to agree with accepted dogma and to social role pressures in their relationships to students, participants, and other scientists. They are partly motivated by sociological and economic rewards for their work (sadly, sometimes to the point of fraud), and they display all-too-human psychological needs and irrationalities about their work. Other limitations have unique relevance to experimentation. For example, if causal results are ambiguous, as in many weaker quasi-experiments, experimenters may attribute causation or causal generalization based on study features that have little to do with orthodox logic or method. They may fail to pursue all the alternative causal explanations because of a lack of energy, a need to achieve closure, or a bias toward accepting evidence that confirms their preferred hypothesis. Each experiment is also a social situation, full of social roles (e.g., participant, experimenter, assistant) and social expectations (e.g., that people should provide true information) but with a uniqueness (e.g., that the experimenter does not always tell the truth) that can lead to problems when social cues are misread or deliberately thwarted by either party. Fortunately, these limits are not insurmountable, as formal training can help overcome some of them (Lehman, Lempert, & Nisbett, 1988). Still, the relationship between scientific results and the world that science studies is neither simple nor fully trustworthy.

These social and psychological analyses have taken some of the luster from the experiment as a centerpiece of science. The experiment may have a life of its own, but it is no longer life on a pedestal. Among scientists, belief in the experiment as the *only* means to settle disputes about causation is gone, though it is still the preferred method in many circumstances. Gone, too, is the belief that the power experimental methods often displayed in the laboratory would transfer easily to applications in field settings. As a result of highly publicized science-related

events such as the tragic results of the Chernobyl nuclear disaster, the disputes over certainty levels of DNA testing in the O.J. Simpson trials, and the failure to find a cure for most cancers after decades of highly publicized and funded effort, the general public now better understands the limits of science.

Yet we should not take these critiques too far. Those who argue against theory-free tests often seem to suggest that every experiment will come out just as the experimenter wishes. This expectation is totally contrary to the experience of researchers, who find instead that experimentation is often frustrating and disappointing for the theories they loved so much. Laboratory results may not speak for themselves, but they certainly do not speak only for one's hopes and wishes. We find much to value in the laboratory scientist's belief in "stubborn facts" with a life span that is greater than the fluctuating theories with which one tries to explain them. Thus many basic results about gravity are the same, whether they are contained within a framework developed by Newton or by Einstein; and no successor theory to Einstein's would be plausible unless it could account for most of the stubborn factlike findings about falling bodies. There may not be pure facts, but some observations are clearly worth treating as if they were facts.

Some theorists of science—Hanson, Polanyi, Kuhn, and Feyerabend included—have so exaggerated the role of theory in science as to make experimental evidence seem almost irrelevant. But exploratory experiments that were unguided by formal theory and unexpected experimental discoveries tangential to the initial research motivations have repeatedly been the source of great scientific advances. Experiments have provided many stubborn, dependable, replicable results that then become the subject of theory. Experimental physicists feel that their laboratory data help keep their more speculative theoretical counterparts honest, giving experiments an indispensable role in science. Of course, these stubborn facts often involve both commonsense presumptions and trust in many well-established theories that make up the shared core of belief of the science in question. And of course, these stubborn facts sometimes prove to be undependable, are reinterpreted as experimental artifacts, or are so laden with a dominant focal theory that they disappear once that theory is replaced. But this is not the case with the great bulk of the factual base, which remains reasonably dependable over relatively long periods of time.

## A WORLD WITHOUT EXPERIMENTS OR CAUSES?

To borrow a thought experiment from MacIntyre (1981), imagine that the slates of science and philosophy were wiped clean and that we had to construct our understanding of the world anew. As part of that reconstruction, would we reinvent the notion of a manipulable cause? We think so, largely because of the practical utility that dependable manipulanda have for our ability to survive and prosper. Would we reinvent the experiment as a method for investigating such causes?

Again yes, because humans will always be trying to better know how well these manipulable causes work. Over time, they will refine how they conduct those experiments and so will again be drawn to problems of counterfactual inference, of cause preceding effect, of alternative explanations, and of all of the other features of causation that we have discussed in this chapter. In the end, we would probably end up with the experiment or something very much like it. This book is one more step in that ongoing process of refining experiments. It is about improving the yield from experiments that take place in complex field settings, both the quality of causal inferences they yield and our ability to generalize these inferences to constructs and over variations in persons, settings, treatments, and outcomes.