EXPERIMENTAL AND QUASI-EXPERIMENTAL DESIGNS FOR GENERALIZED CAUSAL INFERENCE

William R. Shadish
THE UNIVERSITY OF MEMPHIS

Thomas D. Cook
NORTHWESTERN UNIVERSITY

Donald T. Campbell

HOUGHTON MIFFLIN COMPANY Boston New York

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 rival 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.

14

A Critical Assessment of Our Assumptions

As·sump·tion (ə-sŭmp′shən): [Middle English *assumpcion,* from Latin *assumpti, assumptin-* adoption, from *assumptus,* past participle of *assmere,* to adopt; see assume.] n. 1. The act of taking to or upon oneself: *assumption of an obligation.* 2. The act of taking over: *assumption of command.* 3. The act of taking for granted: *assumption of a false theory.* 4. Something taken for granted or accepted as true without proof; a supposition: *a valid assumption.* 5. Presumption; arrogance. 6. Logic. A minor premise.

THIS BOOK covers five central topics across its 13 chapters. The first topic (Chapter 1) deals with our general understanding of descriptive causation and experimentation. The second (Chapters 2 and 3) deals with the types of validity and the specific validity threats associated with this understanding. The third (Chapters 4 through 7) deals with quasi-experiments and illustrates how combining design features can facilitate better causal inference. The fourth (Chapters 8 through 10) concerns randomized experiments and stresses the factors that impede and promote their implementation. The fifth (Chapters 11 through 13) deals with causal generalization, both theoretically and as concerns the conduct of individual studies and programs of research. The purpose of this last chapter is to critically assess some of the assumptions that have gone into these five topics, especially the assumptions that critics have found objectionable or that we anticipate they will find objectionable. We organize the discussion around each of the five topics and then briefly justify why we did not deal more extensively with nonexperimental methods for assessing causation.

We do not delude ourselves that we can be the best explicators of our own assumptions. Our critics can do that task better. But we want to be as comprehensive and as explicit as we can. This is in part because we are convinced of the advantages of falsification as a major component of any **epistemology** for the social sciences, and forcing out one's assumptions and confronting them is one part of falsification. But it is also because we would like to stimulate critical debate about these assumptions so that we can learn from those who would challenge our thinking. If there were to be a future book that carried even further forward the tradition emanating from Campbell and Stanley via Cook and Campbell to this book, then that future book would probably be all the better for building upon all the justified criticisms coming from those who do not agree with us, either on particulars or on the whole approach we have taken to the analysis of descriptive causation and its generalization. We would like this chapter not only to model the attempt to be critical about the assumptions all scholars must inevitably make but also to encourage others to think about these assumptions and how they might be addressed in future empirical or theoretical work.

CAUSATION AND EXPERIMENTATION

Causal Arrows and Pretzels

Experiments test the influence of one or at most a small subset of descriptive causes. If statistical interactions are involved, they tend to be among very few treatments or between a single treatment and a limited set of moderator variables. Many researchers believe that the causal knowledge that results from this typical experimental structure fails to map the many causal forces that simultaneously affect any given outcome in complex and nonlinear ways (e.g., Cronbach et al., 1980; Magnusson, 2000). These critics assert that experiments prioritize on arrows connecting A to B when they should instead seek to describe an explanatory pretzel or set of intersecting pretzels, as it were. They also believe that most causal relationships vary across units, settings, and times, and so they doubt whether there are any constant bivariate causal relationships (e.g., Cronbach & Snow, 1977). Those that do appear to be dependable in the data may simply reflect statistically underpowered tests of moderators or mediators that failed to reveal the true underlying complex causal relationships. True variation in effect sizes might also be obscured because the relevant substantive theory is underspecified, or the outcome measures are partially invalid, or the treatment contrast is attenuated, or causally implicated variables are truncated in how they are sampled (McClelland & Judd, 1993).

As valid as these objections are, they do not invalidate the case for experiments. The purpose of experiments is not to completely explain some phenomenon; it is to identify whether a particular variable or small set of variables makes a marginal difference in some outcome over and above all the other forces affecting that outcome. Moreover, **ontological** doubts such as the preceding have not stopped believers in more complex causal theories from acting as though many causal relationships can be usefully characterized as dependable main effects or as very simple nonlinearities that are also dependable enough to be useful. In this connection, consider some examples from education in the United States, where objections to experimentation are probably the most prevalent and virulent. Few educational researchers seem to object to the following substantive conclusions of the form that A dependably causes B: small schools are better than large ones; time-on-task raises achievement; summer school raises test scores; school desegregation hardly affects achievement but does increase White flight; and assigning and grading homework raises achievement. The critics also do not seem to object to other conclusions involving very simple causal contingencies: reducing class size increases achievement, but only if the amount of change is "sizable" and to a level under 20; or Catholic schools are superior to public ones, but only in the inner city and not in the suburbs and then most noticeably in graduation rates rather than in achievement test scores.

The primary justification for such oversimplifications—and for the use of the experiments that test them—is that some moderators of effects are of minor relevance to policy and theory, even if they marginally improve explanation. The most important contingencies are usually those that modify the sign of a causal relationship rather than its magnitude. Sign changes imply that a treatment is beneficial in some circumstances but might be harmful in others. This is quite different from identifying circumstances that influence just how positive an effect might be. Policy-makers are often willing to advocate an overall change, even if they suspect it has different-sized positive effects for different groups, as long as the effects are rarely negative. But if some groups will be positively affected and others negatively, political actors are loath to prescribe different treatments for different groups because rivalries and jealousies often ensue. Theoreticians also probably pay more attention to causal relationships that differ in causal sign because this result implies that one can identify the boundary conditions that impel such a disparate data pattern.

Of course, we do not advocate ignoring all causal contingencies. For example, physicians routinely prescribe one of several possible interventions for a given diagnosis. The exact choice may depend on the diagnosis, test results, patient preferences, insurance resources, and the availability of treatments in the patient's area. However, the costs of such a contingent system are high. In part to limit the number of relevant contingencies, physicians specialize, and within their own specialty they undergo extensive training to enable them to make these contingent decisions. Even then, substantial judgment is still required to cover the many situations in which causal contingencies are ambiguous or in dispute. In many other policy domains it would also be costly to implement the financial, management, and cultural changes that a truly contingent system would require even if the requisite knowledge were available. Taking such a contingent approach to its logical extremes would entail in education, for example, that individual tutoring become the order of the day. Students and instructors would have to be carefully matched for overlap in teaching and learning skills and in the curriculum supports they would need.

Within limits, some moderators can be studied experimentally, either by measuring the moderator so it can be tested during analysis or by deliberately varying it in the next study in a program of research. In conducting such experiments, one moves away from the black-box experiments of yesteryear toward taking causal contingencies more seriously and toward routinely studying them by, for example, disaggregating the treatment to examine its causally effective components, disaggregating the effect to examine its causally impacted components, conducting analyses of demographic and psychological moderator variables, and exploring the causal pathways through which (parts of) the treatment affects (parts of) the outcome. To do all of this well in a single experiment is not possible, but to do some of it well is possible and desirable.

Epistemological Criticisms of Experiments

In highlighting statistical conclusion validity and in selecting examples, we have often linked causal description to quantitative methods and hypothesis testing. Many critics will (wrongly) see this as implying a discredited theory of positivism. As a philosophy of science first outlined in the early 19th century, positivism rejected metaphysical speculations, especially about unobservables, and equated knowledge with descriptions of experienced phenomena. A narrower school of logical positivism emerged in the early 20th century that also rejected realism while also emphasizing the use of data-theory connections in predicate logic form and a preference for predicting phenomena over explaining them. Both these related epistemologies were long ago discredited, especially as explanations of how science operates. So few critics seriously criticize experiments on this basis. However, many critics use the term *positivism* with less historical fidelity to attack quantitative social science methods in general (e.g., Lincoln & Guba, 1985). Building on the rejection of logical positivism, they reject the use of quantification and formal logic in observation, measurement, and hypothesis testing. Because these last features are part of experiments, to reject this loose conception of positivism entails rejecting experiments. However, the errors in such criticisms are numerous. For example, to reject a specific feature of positivism (like the idea that quantification and predicate logic are the only permissible links between data and theory) does not necessarily imply rejecting all related and more general propositions (such as the notion that some kinds of quantification and hypothesis testing may be useful for knowledge growth). We and others have outlined more such errors elsewhere (Phillips, 1990; Shadish, 1995a).

Other epistemological criticisms of experimentation cite the work of historians of science such as Kuhn (1962), of sociologists of science such as Latour and Woolgar (1979) and of philosophers of science such as Harré (1981). These critics tend to focus on three things. One is the incommensurability of theories, the notion that theories are never perfectly specified and so can always be reinterpreted. As a result, when disconfirming data seem to imply that a theory should be rejected, its postulates can instead be reworked in order to make the theory and observations consistent with each other. This is usually done by adding new contingencies to the theory that limit the conditions under which it is thought to hold. A second critique is of the assumption that experimental observations can be used as truth tests. We would like observations to be objective assessments that can adjudicate between different theoretical explanations of a phenomenon. But in practice, observations are not theory neutral; they are open to multiple interpretations that include such irrelevancies as the researcher's hopes, dreams, and predilections. The consequence is that observations rarely result in definitive hypothesis tests. The final criticism follows from the many behavioral and cognitive inconsistencies between what scientists do in practice and what scientific norms prescribe they should do. Descriptions of scientists' behavior in laboratories reveal them as choosing to do particular experiments because they have an intuition about a relationship, or they are simply curious to see what happens, or they want to play with a new piece of equipment they happen to find lying around. Their impetus, therefore, is not a hypothesis carefully deduced from a theory that they then test by means of careful observation.

Although these critiques have some credibility, they are overgeneralized. Few experimenters believe that their work yields definitive results even after it has been subjected to professional review. Further, though these philosophical, historical, and social critiques complicate what a "fact" means for *any* scientific method, nonetheless many relationships have stubbornly recurred despite changes associated with the substantive theories, methods, and researcher biases that first generated them. Observations may never achieve the status of "facts," but many of them are so stubbornly replicable that they may be considered as though they were facts. For experimenters, the trick is to make sure that observations are not impregnated with just one theory, and this is done by building multiple theories into observations and by valuing independent replications, especially those of substantive critics—what we have elsewhere called **critical multiplism** (Cook, 1985; Shadish, 1989, 1994).

Although causal claims can never be definitively tested and proven, individual experiments still manage to probe such claims. For example, if a study produces negative results, it is often the case that program developers and other advocates then bring up methodological and substantive contingencies that might have changed the result. For instance, they might contend that a different outcome measure or population would have led to a different conclusion. Subsequent studies then probe these alternatives and, if they again prove negative, lead to yet another round of probes of whatever new explanatory possibilities have emerged. After a time, this process runs out of steam, so particularistic are the contingencies that remain to be examined. It is as though a consensus emerges: "The causal relationship was not obtained under many conditions. The conditions that remain to be examined are so circumscribed that the intervention will not be worth much even if it is effective under these conditions." We agree that this process is as much or more social than logical. But the reality of elastic theory does not mean that decisions about causal hypotheses are only social and devoid of all empirical and logical content.

The criticisms noted are especially useful in highlighting the limited value of individual studies relative to reviews of research programs. Such reviews are better because the greater diversity of study features makes it less likely that the same theoretical biases that inevitably impregnate any one study will reappear across all the studies under review. Still, a dialectic process of point, response, and counterpoint is needed even with reviews, again implying that no single review is definitive. For example, in response to Smith and Glass's (1977) meta-analytic claim that psychotherapy was effective, Eysenck (1977) and Presby (1977) pointed out methodological and substantive contingencies that challenged the original reviewers' results. They suggested that a different answer would have been achieved if Smith and Glass had not combined randomized and nonrandomized experiments or if they had used narrower categories in which to classify types of therapy. Subsequent studies probed these challenges to Smith and Glass or brought forth novel ones (e.g., Weisz et al., 1992). This process of challenging causal claims with specific alternatives has now slowed in reviews of psychotherapy as many major contingencies that might limit effectiveness have been explored. The current consensus from reviews of many experiments in many kinds of settings is that psychotherapy is effective; it is not just the product of a regression process (spontaneous remission) whereby those who are temporarily in need seek professional help and get better, as they would have even without the therapy.

Neglected Ancillary Questions

Our focus on causal questions within an experimental framework neglects many other questions that are relevant to causation. These include questions about how to decide on the importance or leverage of any single causal question. This could entail exploring whether a causal question is even warranted, as it often is not at the early stage of development of an issue. Or it could entail exploring what type of causal question is more important—one that fills an identified hole in some literature, or one that sets out to identify specific boundary conditions limiting a causal connection, or one that probes the validity of a central assumption held by all the theorists and researchers within a field, or one that reduces uncertainty about an important decision when formerly uncertainty was high. Our approach also neglects the reality that how one formulates a descriptive causal question usually entails meeting some stakeholders' interests in the social research more than those of others. Thus to ask about the effects of a national program meets the needs of Congressional staffs, the media, and policy wonks to learn about whether the program works. But it can fail to meet the needs of local practitioners who usually want to know about the effectiveness of microelements within the program so that they can use this knowledge to improve their daily practice. In more theoretical work, to ask how some intervention affects personal self-efficacy is likely to promote individuals' autonomy needs, whereas to ask about the effects of a persuasive communication designed to change attitudes could well cater to the needs of those who would limit or manipulate such autonomy. Our narrow technical approach to causation also neglected issues related to how such causal knowledge might be used and misused. It gave short shrift to a systematic analysis of the kinds of causal questions that can and cannot be answered through experiments. What about the effects of abortion, divorce, stable cohabitation, birth out of wedlock, and other possibly harmful events that we cannot ethically manipulate? What about the effects of class, race, and gender that are not amenable to experimentation? What about the effects of historical occurrences that can be studied only by using time-series methods on whatever variables might or might not be in the archives? Of what use, one might ask, is a method that cannot get at some of the most important phenomena that shape our social world, often over generations, as in the case of race, class, and gender?

Many statisticians now consider questions about things that cannot be manipulated as being beyond causal analysis, so closely do they link manipulation to causation. To them, the cause must be at least potentially manipulable, even if it is not actually manipulated in a given observational study. Thus they would not consider race a cause, though they would speak of the causal analysis of race in studies in which Black and White couples are, say, randomly assigned to visiting rental units in order to see if the refusal rates vary, or that entail chemically changing skin color to see how individuals are responded to differently as a function of pigmentation, or that systematically varied the racial mix of students in schools or classrooms in order to study teacher responses and student performance. Many critics do not like so tight a coupling of manipulation and causation. For example, those who do status attainment research consider it obvious that race causally influences how teachers treat individual minority students and thus affects how well these children do in school and therefore what jobs they get and what prospects their own children will subsequently have. So this coupling of cause to manipulation is a real limit of an experimental approach to causation. Although we like the coupling of causation and manipulation for purposes of defining experiments, we do not see it as necessary to all useful forms of cause.

VALIDITY

Objections to Internal Validity

There are several criticisms of Campbell's (1957) validity typology and its extensions (Gadenne, 1976; Kruglanski & Kroy, 1976; Hultsch & Hickey, 1978; Cronbach, 1982; Cronbach et al., 1980). We start first with two criticisms of internal validity raised by Cronbach (1982) and to a lesser extent by Kruglanski and Kroy (1976): (1) an atheoretically defined internal validity (A causes B) is trivial without reference to constructs; and (2) causation in single instances is impossible, including in single experiments.

Internal Validity Is Trivial

Cronbach (1982) writes:

> I consider it pointless to speak of causes when all that can be validly meant by reference to a cause in a particular instance is that, on one trial of a partially specified manipulation under conditions A, B, and C, along with other conditions not named, phenomenon P was observed. To introduce the word cause seems pointless. Campbell's writings make internal validity a property of trivial, past-tense, and local statements. (p. 137)

Hence, "causal language is superfluous" (p. 140). Cronbach does not retain a specific role for causal inference in his validity typology at all. Kruglanski and Kroy (1976) criticize internal validity similarly, saying:

> The concrete events which constitute the treatment within a specific research are meaningful only as members of a general conceptual category. . . . Thus, it is simply impossible to draw strictly specific conclusions from an experiment: our concepts are general and each presupposes an implicit general theory about resemblance between different concrete cases. (p. 167)

All these authors suggest collapsing internal with construct validity in different ways.

Of course, we agree that researchers conceptualize and discuss treatments and outcomes in conceptual terms. As we said in Chapter 3, constructs are so basic to language and thought that it is impossible to conceptualize scientific work without them. Indeed, in many important respects, the constructs we use constrain what we experience, a point agreed to by theorists ranging from Quine (1951, 1969) to the postmodernists (Conner, 1989; Tester, 1993). So when we say that internal validity concerns an atheoretical local molar causal inference, we do not mean that the researcher should conceptualize experiments or report a causal claim as "Something made a difference," to use Cronbach's (1982, p. 130) exaggerated characterization.

Still, it is both sensible and useful to differentiate internal from construct validity. The task of sorting out constructs is demanding enough to warrant separate attention from the task of sorting out causes. After all, operations are concept laden, and it is very rare for researchers to know fully what those concepts are. In fact, the researcher almost certainly cannot know them fully because paradigmatic concepts are so implicitly and universally imbued that those concepts and their assumptions are sometimes entirely unrecognized by research communities for years. Indeed, the history of science is replete with examples of famous series of experiments in which a causal relationship was demonstrated early, but it took years for the cause (or effect) to be consensually and stably named. For instance, in psychology and linguistics many causal relationships originally emanated from a behaviorist paradigm but were later relabeled in cognitive terms; in the early Hawthorne study, illumination effects were later relabeled as effects of obtrusive observers; and some cognitive dissonance effects have been reinterpreted as attribution effects. In the history of a discipline, relationships that are correctly identified as causal can be important even when the cause and effect constructs are incorrectly labeled. Such examples exist because the reasoning used to draw causal inferences (e.g., requiring evidence that treatment preceded outcome) differs from the reasoning used to generalize (e.g., matching operations to prototypical characteristics of constructs). Without understanding what is meant by descriptive causation, we have no means of telling whether a claim to have established such causation is justified.

Cronbach's (1982) prose makes clear that he understands the importance of causal logic; but in the end, his sporadically expressed craft knowledge does not add up to a coherent theory of judging the validity of descriptive causal inferences. His equation of internal validity as part of reproducibility (under replication) misses the point that one can replicate incorrect causal conclusions. His solution to such questions is simply that "the force of each question can be reduced by suitable controls" (1982, p. 233). This is inadequate, for a complete analysis of the problem of descriptive causal inference requires concepts we can use to recognize suitable controls. If a suitable control is one that reduces the plausibility of, say, history or maturation, as Cronbach (1982, p. 233) suggests, this is little more than internal validity as we have formulated it. If one needs the concepts enough to use them, then they should be part of a validity typology for cause-probing methods.

For completeness, we might add that a similar boundary question arises between construct validity and external validity and between construct validity and statistical conclusion validity. In the former case, no scientist ever frames an external validity question without couching the question in the language of constructs. In the latter case, researchers never conceptualize or discuss their results solely in terms of statistics. Constructs are ubiquitous in the process of doing research because they are essential for conceptualizing and reporting operations. But again, the answer to this objection is the same. The strategies for making inferences about a construct are not the same as strategies for making inferences about whether a causal relationship holds over variation in persons, settings, treatments, and outcomes in external validity or for drawing valid statistical conclusions in the case of statistical conclusion validity. Construct validity requires a theoretical argument and an assessment of the correspondence between samples and constructs. External validity requires analyzing whether causal relationships hold over variations in persons, settings, treatments, and outcomes. Statistical conclusion validity requires close examination of the statistical procedures and assumptions used. And again, one can be wrong about construct labels while being right about external or statistical conclusion validity.

Objections to Causation in Single Experiments

A second criticism of internal validity denies the possibility of inferring causation in a single experiment. Cronbach (1982) says that the important feature of causation is the "progressive localization of a cause" (Mackie, 1974, p. 73) over multiple experiments in a program of research in which the uncertainties about the essential features of the cause are reduced to the point at which one can characterize exactly what the cause is and is not. Indeed, much philosophy of causation asserts that we only recognize causes through observing multiple instances of a putative causal relationship, although philosophers differ as to whether the mechanism for recognition involves logical laws or empirical regularities (Beauchamp, 1974; P. White, 1990).

However, some philosophers do defend the position that causes can be inferred in single instances (e.g., Davidson, 1967; Ducasse, 1951; Madden & Humber, 1971). A good example is causation in the law (e.g., Hart & Honore, 1985), by which we judge whether or not one person, say, caused the death of another despite the fact that the defendant may never before have been on trial for a crime. The verdict requires a plausible case that (among other things) the defendant's actions preceded the death of the victim, that those actions were related to the death, that other potential causes of the death are implausible, and that the death would not have occurred had the defendant not taken those actions—the very logic of causal relationships and counterfactuals that we outlined in Chapter 1. In fact, the defendant's criminal history will often be specifically excluded from consideration in judging guilt during the trial. The lesson is clear. Although we may learn more about causation from multiple than from single experiments, we *can* infer cause in single experiments. Indeed, experimenters will do so whether we tell them to or not. Providing them with conceptual help in doing so is a virtue, not a vice; failing to do so is a major flaw in a theory of cause-probing methods.

Of course, individual experiments virtually always use prior concepts from other experiments. However, such prior conceptualizations are entirely consistent with the claim that internal validity is about causal claims in single experiments. If it were not (at least partly) about single experiments, there would be no point to doing the experiment, for the prior conceptualization would successfully predict what will be observed. The possibility that the data will not support the prior conceptualization makes internal validity essential. Further, prior conceptualizations are not logically necessary; we can experiment to discover effects that we have no prior conceptual structure to expect: "The physicist George Darwin used to say that once in a while one should do a completely crazy experiment, like blowing the trumpet to the tulips every morning for a month. Probably nothing will happen, but if something did happen, that would be a stupendous discovery" (Hacking, 1983, p. 154). But we would still need internal validity to guide us in judging if the trumpets had an effect.

Objections to Descriptive Causation

A few authors object to the very notion of descriptive causation. Typically, however, such objections are made about a caricature of descriptive causation that has not been used in philosophy or in science for many years—for example, a billiard ball model that requires a commitment to deterministic causation or that excludes reciprocal causation. In contrast, most who write about experimentation today espouse theories of probabilistic causation in which the many difficulties associated with identifying dependable causal relationships are humbly acknowledged. Even more important, these critics inevitably use causal-sounding language themselves, for example, replacing "cause" with "mutual simultaneous shaping" (Lincoln & Guba, 1985, p. 151). These replacements seem to us to avoid the word but keep the concept, and for good reason. As we said at the end of Chapter 1, if we wiped the slate clean and constructed our knowledge of the world anew, we believe we would end up reinventing the notion of descriptive causation all over again, so greatly does knowledge of causes help us to survive in the world.

Objections Concerning the Discrimination Between Construct Validity and External Validity

Although we traced the history of the present validity system briefly in Chapter 2, readers may want additional historical perspective on why we made the changes we made in the present book regarding construct and external validity. Both Campbell (1957) and Campbell and Stanley (1963) only used the phrase external validity, which they defined as inferring to what populations, settings, treatment variables, and measurement variables an effect can be generalized. They did not refer at all to construct validity. However, from his subsequent writings (Campbell, 1986), it is clear Campbell thought of construct validity as being part of external validity. In Campbell and Stanley, therefore, external validity subsumed generalizing from research operations about persons, settings, causes, and effects for the purposes of labeling these particulars in more abstract terms, and also generalizing by identifying sources of variation in causal relationships that are attributable to person, setting, cause, and effect factors. All subsequent conceptualizations also share the same generic strategy based on sampling instances of persons, settings, causes, and effects and then evaluating them for their presumed correspondence to targets of inference.

In Campbell and Stanley's formulation, person, setting, cause, and effect categories share two basic similarities despite their surface differences—to wit, all of them have both ostensive qualities and construct representations. Populations of persons or settings are composed of units that are obviously individually ostensive. This capacity to point to individual persons and settings, especially when they are known to belong in a referent category, permits them to be readily enumerated and selected for study in the formal ways that sampling statisticians prefer. By contrast, although individual measures (e.g., the Beck Depression Inventory) and treatments (e.g., a syringe full of a vaccine) are also ostensive, efforts to enumerate all existing ways of measuring or manipulating such measures and treatments are much more rare (e.g., Bloom, 1956; Ciarlo et al., 1986; Steiner & Gingrich, 2000). The reason is that researchers prefer to use substantive theory to determine which attributes a treatment or outcome measure should contain in any given study, recognizing that scholars often disagree about the relevant attributes of the higher order entity and of the supposed best operations to represent them. None of this negates the reality that populations of persons or settings are also defined in part by the theoretical constructs used to refer to them, just like treatments and outcomes; they also have multiple attributes that can be legitimately contested. What, for instance, is the American population? While a legal definition surely exists, it is not inviolate. The German conception of nationality allows that the great grandchildren of a German are Germans even if their parents and grandparents have not claimed German nationality. This is not possible for Americans. And why privilege a legal definition? A cultural conception might admit as American all those illegal immigrants who have been in the United States for decades and it might exclude those American adults with passports who have never lived in the United States. Given that persons, settings, treatments, and outcomes all have both construct and ostensive qualities, it is no surprise that Campbell and Stanley did not distinguish between construct and external validity.

Cook and Campbell, however, did distinguish between the two. Their unstated rationale for the distinction was mostly pragmatic—to facilitate memory for the very long list of threats that, with the additions they made, would have had to fit under Campbell and Stanley's umbrella conception of external validity. In their theoretical discussion, Cook and Campbell associated construct validity with generalizing to causes and effects, and external validity with generalizing to and across persons, settings, and times. Their choice of terms explicitly referenced Cronbach and Meehl (1955) who used construct and construct validity in measurement theory to justify inferences "about higher-order constructs from research operations" (Cook & Campbell, 1979, p. 38). Likewise, Cook and Campbell associated the terms *population* and *external validity* with sampling theory and the formal and purposive ways in which researchers select instances of persons and settings. But to complicate matters, Cook and Campbell also briefly acknowledged that "all aspects of the research require naming samples in generalizable terms, including samples of peoples and settings as well as samples of measures or manipulations" (p. 59). And in listing their external validity threats as statistical interactions between a treatment and population, they linked external validity more to generalizing across populations than to generalizing to them. Also, their construct validity threats were listed in ways that emphasized generalizing to cause and effect constructs. Generalizing across different causes and effects was listed as external validity because this task does not involve attributing meaning to a particular measure or manipulation. To read the threats in Cook and Campbell, external validity is about generalizing across populations of persons and settings and across different cause and effect constructs, while construct validity is about generalizing to causes and effects. Where, then, is generalizing from samples of persons or settings to their referent populations? The text discusses this as a matter of external validity, but this classification is not apparent in the list of validity threats. A system is needed that can improve on Cook and Campbell's partial confounding between objects of generalization (causes and effects versus persons and settings) and functions of generalization (generalizing to higher-order constructs from research operations versus inferring the degree of replication across different constructs and populations).

This book uses such a functional approach to differentiate construct validity from external validity. It equates construct validity with labeling research operations, and external validity with sources of variation in causal relationships. This new formulation subsumes all of the old. Thus, Cook and Campbell's understanding of construct validity as generalizing from manipulations and measures to cause and effect constructs is retained. So is external validity understood as generalizing across samples of persons, settings, and times. And generalizing across different cause or effect constructs is now even more clearly classified as part of external validity. Also highlighted is the need to label samples of persons and settings in abstract terms, just as measures and manipulations need to be labeled. Such labeling would seem to be a matter of construct validity, given that construct validity is functionally defined in terms of labeling. However, labeling human samples might have been read as being a matter of external validity in Cook and Campbell, given that their referents were human populations and their validity types were organized more around referents than functions. So, although the new formulation in this book is definitely more systematic than its predecessors, we are unsure whether that systematization will ultimately result in greater terminological clarity or confusion. To keep the latter to a minimum, the following discussion reflects issues pertinent to the demarcation of construct and external validity that have emerged either in deliberations between the first two authors or in classes that we have taught using pre-publication versions of this book.

Is Construct Validity a Prerequisite for External Validity?

In this book, we equate external validity with variation in causal relationships and construct validity with labeling research operations. Some readers might see this as suggesting that successful generalization of a causal relationship requires the accurate labeling of each population of persons and each type of setting to which generalization is sought, even though we can never be certain that anything is labeled with perfect accuracy. The relevant task is to achieve the most accurate assessment available under the circumstances. Technically, we can test generalization across entities that are already known to be confounded and thus not labeled well—e.g., when causal data are broken out by gender but the females in the sample are, on average, more intelligent than the males and therefore score higher on everything else correlated with intelligence. This example illustrates how dangerous it is to rely on measured surface similarity alone (i.e., gender differences) for determining how a sample should be labeled in population terms. We might more accurately label gender differences if we had a random sample of each gender taken from the same population. But this is not often found in experimental work, and even this is not perfect because gender is known to be confounded with other attributes (e.g., income, work status) even in the population, and those other attributes may be pertinent labels for some of the inferences being made. Hence, we usually have to rely on the assumption that, because gender samples come from the same physical setting, they are comparable on all background characteristics that might be correlated with the outcome. Because this assumption cannot be fully tested and is anyway often false—as in the hypothetical example above—this means that we could and should measure all the potential confounds within the limits of our theoretical knowledge to suggest them, and that we should also use these measures in the analysis to reduce confounding.

Even with acknowledged confounding, sample-specific differences in effect sizes may still allow us to conclude that a causal relationship varies by something associated with gender. This is a useful conclusion for preventing premature overgeneralization. With more breakdowns, confounded or not, one can even get a sense of the percentage of contrasts across which a causal relationship does and does not hold. But without further work, the populations across which the relationship varies are incompletely identified. The value of identifying them better is particularly salient when some effect sizes cannot be distinguished from zero. Although this clearly identifies a nonuniversal causal relationship, it does not advance theory or practice by specifying the labeled boundary conditions over which a causal relationship fails to hold. Knowledge gains are also modest from generalization strategies that do not explicitly contrast effect sizes. Thus, when different populations are lumped together in a single hypothesis test, researchers can learn how large a causal relationship is despite the many unexamined sources of variation built into the analysis. But they cannot accurately identify which constructs do and do not co-determine the relationship's size. Construct validity adds useful specificity to external validity concerns, but it is not a necessary condition for external validity. We can generalize across entities known to be confounded, albeit less usefully than across accurately labeled entities.

This last point is similar to the one raised earlier to counter the assertion of Gadenne (1976) and Kruglanski and Kroy (1976) that internal validity requires the high construct validity of both cause and effect. They assert that all science is about constructs, and so it has no value to conclude that "something caused something else"—the result that would follow if we did a technically exemplary randomized experiment with correspondingly high internal validity, but the cause and effect were not labeled. Nonetheless, a causal relationship is demonstrably entailed, and the finding that "something reliably caused something else" might lead to further research to refine whatever clues are available about the cause and effect constructs. A similar argument holds for the relationship of construct to external validity. Labels with high construct validity are not necessary for internal or for external validity, but they are useful for both.

Researchers necessarily use the language of constructs (including human and setting population ones) to frame their research questions and select their representations of constructs in the samples and measures chosen. If they have designed their work well and have had some luck, the constructs they begin and end with will be the same, though critics can challenge any claims they make. However, the samples and constructs might not match well, and then the task is to examine the samples and ascertain what they might alternatively stand for. As critics like Gadenne, Kruglanski, and Kroy have pointed out, such reliance on the operational level seems to legitimize operations as having a life independent of constructs. This is not the case, though, for operations are intimately dependent on interpretations at all stages of research. Still, every operation fits some interpretations, however tentative that referent may be due to poor research planning or to nature turning out to be more complex than the researcher's initial theory.

How Does Variation Across Different Operational Representations of the Same Intended Cause or Effect Relate to Construct and External Validity?

In Chapter 3 we emphasized how the valid labeling of a cause or effect benefits from multiple operational instances, and also that these various instances can be fruitfully analyzed to examine how a causal relationship varies with the definition used. If each operational instance is indeed of the same underlying construct, then the same causal relationship should result regardless of how the cause or effect is operationally defined. Yet data analysis sometimes reveals that a causal relationship varies by operational instance. This means that the operations are not in fact equivalent, so that they presumably tap both into different constructs and into different causal relationships. Either the same causal construct is differently related to what now must be seen as two distinct outcomes, or the same effect construct is differently related to two or more unique causal agents. So the intention to promote the construct validity of causes and effects by using multiple operations has now facilitated conclusions about the external validity of causes or effects; that is, when the external validity of the cause and effect are in play, the data analysis has revealed that more than one causal relationship needs to be invoked.

Fortunately, when we find that a causal relationship varies over different causes or different effects, the research and its context often provide clues as to how the causal elements in each relationship might be (re)labeled. For example, the researcher will generally examine closely how the operations differ in their particulars, and will also study which unique meanings have been attached to variants like these in the existing literature. While the meanings that are achieved might be less successful because they have been devised post hoc to fit novel findings, they may in some circumstances still attain an acceptable level of accuracy, and will certainly prompt continued discussion to account for the findings. Thus, we come full circle. We began with multiple operational representations of the same cause or effect when testing a single causal relationship; then the data forced us to invoke more than one relationship; and finally the pattern of the outcomes and their relationship to the existing literature can help improve the labeling of the new relationships achieved. A construct validity exercise begets an external validity conclusion that prompts the need for relabeling constructs. Demonstrating effect size variation across operations presumed to represent the same cause or effect can enhance external validity by showing that more constructs and causal relationships are involved than was originally envisaged; and in that case, it can eventually increase construct validity by preventing any mislabeling of the cause or effect inherent in the original choice of measures and by providing clues from details of the causal relationships about how the elements in each relationship should be labeled. We see here analytic tasks that flow smoothly between construct and external validity concerns, involving each.

Should Generalizing from a Single Sample of Persons or Settings Be Classified as External or Construct Validity?

If a study has a single sample of persons or settings, this sample must represent a population. How this sample should be labeled is an issue. Given that construct validity is about labeling, is labeling the sample an issue of construct validity? After all, external validity hardly seems relevant since with a single sample it is not immediately obvious what comparison of variation in causal relationships would be involved. So if generalizing from a sample of persons or settings is treated as a matter of construct validity analogous to generalizing from treatment and outcome operations, two problems arise. First, this highlights a potential conflict in usage in the general social science community, some parts of which say that generalizations from a sample of people to its population are a matter of external validity, even when other parts say that labeling people is a matter of construct validity. Second, this does not fit with the discussion in Cook and Campbell that treats generalizing from individual samples of persons and settings as an external validity matter, though their list of external validity threats does not explicitly deal with this and only mentions interactions between the treatment and attributes of the setting and person.

The issue is most acute when the sample was randomly selected from the population. Consider why sampling statisticians are so keen to promote random sampling for representing a well-designated universe. Such sampling ensures that the sample and population distributions are identical on all measured and unmeasured variables within the limits of sampling error. Notice that this includes the population label (whether more or less accurate), which random sampling guarantees also applies to the sample. Key to the usefulness of random sampling is having a well bounded population from which to sample, a requirement in sampling theory and something often obvious in practice. Given that many well bounded populations are also well labeled, random sampling then guarantees that a valid population label can equally validly be applied to the sample. For instance, the population of telephone prefixes used in the city of Chicago is known and is obviously correctly labeled. Hence, it would be difficult to use random digit dialing from that list of Chicago prefixes and then mislabel the resulting sample as representing telephone owners in Detroit or only in the Edgewater section of Chicago. Given a clearly bounded population and random sampling, the sample label is the population label, which is why sampling statisticians believe that no method is superior to random selection for labeling samples when the population label is known.

With purposive sample selection, this elegant rationale cannot be used, whether or not the population label is known. Thus, if respondents were selected haphazardly from shopping malls all over Chicago, many of the people studied would belong in the likely population of interest—residents of Chicago. But many would not because some Chicago residents do not go to malls at the hours interviewing takes place, and because many persons in these malls are not from Chicago. Lacking random sampling, we could not even confidently call this sample "people walking in Chicago malls," for other constructs such as volunteering to be interviewed may be systematically confounded with sample membership. So, mere membership in the sample is not sufficient for accurately representing a population, and by the rationale in the previous paragraph, it is also not sufficient for accurately labeling the sample. All this leads to two conclusions worth elaborating: (1) that random sampling can sometimes promote construct validity, and (2) that external validity is in play when inferring that a single causal relationship from a sample would hold in a population, whether from a random sample or not.

On the first point, the conditions under which random sampling can sometimes promote the construct validity of single samples are straightforward. Given a well bounded universe, sampling statisticians have justified random sampling as a way of clearly representing in the sample all population attributes. This must include the population label, and so random sampling results in labeling the sample in the same terms that apply to the population. Random sampling does not, of course, tell us whether the population label is itself reasonably accurate; random sampling will also replicate in the sample any mistakes that are made in labeling the population. However, given that many populations are already reasonably well-labeled based on past research and theory and that such situations are often intuitively obvious for researchers experienced in an area, random sampling can, under these circumstances, be counted on to promote construct validity. However, when random selection has not occurred or when the population label is itself in doubt, this book has explicated other principles and methods that can be used for labeling study operations, including labeling the samples of persons and settings in a study.

On the second point, when the question concerns the validity of generalizing from a causal relationship in a single sample to its population, the reader may also wonder how external validity can be in play at all. After all, we have framed external validity as being about whether the causal relationship holds over *variation* in persons, settings, treatment variables, and measurement variables. If there is only one random sample from a population, where is the variation over which to examine that causal relationship? The answer is simple: the variation is between sampled and unsampled persons in that population. As we said in Chapter 2 (and as was true in our predecessor books), external validity questions can be about whether a causal relationship holds (a) over variations in persons, settings, treatments, and outcomes that *were* in the experiment, and (b) for persons, settings, treatments, and outcomes that *were not* in the experiment. Those persons in a population who were not randomly sampled fall into the latter category. Nothing about external validity, either in the present book or in its predecessors, requires that all possible variations of external validity interest actually be observed in the study—indeed, it would be impossible to do so, and we provided several arguments in Chapter 2 about why it would not be wise to limit external validity questions only to variations actually observed in a study. Of course, in most cases external validity generalizations to things that were not studied are difficult, having to rely on the concepts and methods we outlined in our grounded theory of generalized causal inference in Chapters 11 through 13. But it is the great beauty of random sampling that it guarantees that this generalization will hold over both sampled and unsampled persons. So it is indeed an external validity question whether a causal relationship that has been observed in a single random sample would hold for those units that were in the population but not in the random sample.

In the end, this book treats the labeling of a single sample of persons or settings as a matter of construct validity, whether or not random sampling is used. It also treats the generalization of causal relationships from a single sample to unobserved instances as a matter of external validity—again, whether or not random sampling was used. The fact that random sampling (which is associated with external validity in this book) sometimes happens to facilitate the construct labeling of a sample is incidental to the fact that the population label is already known. Though many population labels are indeed well-known, many more are still matters of debate, as reflected in the examples we gave in Chapter 3 of whether persons should be labeled schizophrenic or settings labeled as hostile work environments. In these latter cases, random sampling makes no contribution to resolving debates about the applicability of those labels. Instead, the principles and methods we outlined in Chapters 11 through 13 will have to be brought to bear. And when random sampling has not been used, those principles and methods will also have to be brought to bear on the external validity problem of generalizing causal relationships from single samples to unobserved instances.

Objections About the Completeness of the Typology

The first objection of this kind is that our lists of particular threats to validity are incomplete. Bracht and Glass (1968), for example, added new external validity threats that they thought were overlooked by Campbell and Stanley (1963); and more recently Aiken and West (1991) pointed to new reactivity threats. These challenges are important because the key to the most confident causal conclusions in our theory of validity is the ability to construct a persuasive argument that every plausible and identified threat to validity has been identified and ruled out. However, there is no guarantee that all relevant threats to validity have been identified. Our lists are not divinely ordained, as can be observed from the changes in the threats from Campbell (1957) to Campbell and Stanley (1963) to Cook and

Campbell (1979) to this book. Threats are better identified from insider knowledge than from abstract and nonlocal lists of threats.

A second objection is that we may have left out particular validity types or organized them suboptimally. Perhaps the best illustration that this is true is Sackett's (1979) treatment of bias in case-control studies. Case-control studies do not commonly fall under the rubric of experimental or quasi-experimental designs; but they are cause-probing designs, and in that sense a general interest in generalized causal inference is at least partly shared. Yet Sackett created a different typology. He organized his list around seven stages of research at which bias can occur: (1) in reading about the field, (2) in sample specification and selection, (3) in defining the experimental exposure, (4) in measuring exposure and outcome, (5) in data analysis, (6) in interpretation of analyses, and (7) in publishing results. Each of these could generate a validity type, some of which would overlap considerably with our validity types. For example, his concept of biases "in executing the experimental manoeuvre" (p. 62) is quite similar to our internal validity, whereas his withdrawal bias mirrors our attrition. However, his list also suggests new validity types, such as biases in reading the literature, and biases he lists at each stage are partly orthogonal to our lists. For example, biases in reading include biases of rhetoric in which "any of several techniques are used to convince the reader without appealing to reason" (p. 60).

In the end, then, our claim is only that the present typology is reasonably well informed by knowledge of the nature of generalized causal inference and of some of the problems that are frequently salient about those inferences in field experimentation. It can and hopefully will continue to be improved both by addition of threats to existing validity types and by thoughtful exploration of new validity types that might pertain to the problem of generalized causal inference that is our main concern.[1]

[1] We are acutely aware of, and modestly dismayed at, the many different usages of these validity labels that have developed over the years and of the risk that poses for terminological confusion—even though we are responsible for many of these variations ourselves. After all, the understandings of validity in this book differ from those in Campbell and Stanley (1963), whose only distinction was between internal and external validity. They also differ from Cook and Campbell (1979), in which external validity was concerned with generalizing to and across populations of persons and settings, whereas all issues of generalizing from the cause and effect operations constituted the domain of construct validity. Further, Campbell (1986) himself relabeled internal validity and external validity as local molar causal validity and the principle of proximal similarity, respectively. Stepping outside Campbell's tradition, Cronbach (1982) used these labels with yet other meanings. He said internal validity is the problem of generalizing from samples to the domain about which the question is asked, which sounds much like our construct validity except that he specifically denied any distinction between construct validity and external validity, using the latter term to refer to generalizing results to unstudied populations, an issue of extrapolation beyond the data at hand. Our understanding of external validity includes such extrapolations as one case, but it is not limited to that because it also has to do with empirically identifying sources of variation in an effect size when existing data allow doing so. Finally, many other authors have casually used all these labels in completely different ways (Goetz & LeCompte, 1984; Kleinbaum, Kupper, & Morgenstern, 1982; Menard, 1991). So in view of all these variations, we urge that these labels be used only with descriptions that make their intended understandings clear.

Objections Concerning the Nature of Validity

We defined validity as the approximate truth of an inference. Others define it differently. Here are some alternatives and our reasons for not using them.

Validity in the New Test Theory Tradition

Test theorists discussed validity (e.g., Cronbach, 1946; Guilford, 1946) well before Campbell (1957) invented his typology. We can only begin to touch on the many issues pertinent to validity that abound in that tradition. Here we outline a few key points that help differentiate our approach from that of test theory. The early emphasis in test theory was mostly on inferences about what a test measured, with a pinnacle being reached in the notion of construct validity. Cronbach (1989) credits Cook and Campbell for giving "proper breadth to the notion of constructs" (p. 152) in construct validity through their claim that construct validity is not just limited to inferences about outcomes but also about causes and about other features of experiments. In addition, early test theory tied validity to the truth of such inferences: "The literature on validation has concentrated on the truthfulness of test interpretation" (Cronbach, 1988, p. 5).

However, the years have brought change to this early understanding. In one particularly influential definition of validity in test theory, Messick (1989) said, "Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p. 13); and later he says that "Validity is broadly defined as nothing less than an evaluative summary of both the evidence for and the actual—as well as potential—consequences of score interpretation and use" (1995, p. 742). Whereas our understanding of validity is that *inferences* are the subject of validation, this definition suggests that *actions* are also subject to validation and that validation is actually evaluation. These extentions are far from our view.

A little history will help here. Tests are designed for practical use. Commercial test developers hope to profit from sales to those who use tests; employers hope to use tests to select better personnel; and test takers hope that tests will tell them something useful about themselves. These practical applications generated concern in the American Psychological Association (APA) to identify the characteristics of better and worse tests. APA appointed a committee chaired by Cronbach to address the problem. The committee produced the first in a continuing series of test standards (APA, 1954); and this work also led to Cronbach and Meehl's (1955) classic article on construct validity. The test standards have been frequently revised, most recently cosponsored by other professional associations (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1985, 1999). Requirements to adhere to the standards became part of professional ethical codes. The standards were also influential in legal and regulatory proceedings and have been cited, for example, in U.S. Supreme Court cases about alleged misuses of testing practices (e.g., Albermarle Paper Co. v. Moody, 1975; Washington v. Davis, 1976) and have influenced the "Uniform Guidelines" for personnel selection by the Equal Employment Opportunity Commission (EEOC) et al. (1978). Various validity standards were particularly salient in these uses.

Because of this legal, professional, and regulatory concern with the use of testing, the research community concerned with measurement validity began to use the word *validity* more expansively, for example, "as one way to justify the use of a test" (Cronbach, 1989, p. 149). It is only a short distance from validating use to validating action, because most of the relevant uses were actions such as hiring or firing someone or labeling someone retarded. Actions, in turn, have consequences—some positive, such as efficiency in hiring and accurate diagnosis that allows better tailoring of treatment, and some negative, such as loss of income and stigmatization. So Messick (1989, 1995) proposed that validation also evaluate those consequences, especially the social justice of consequences. Thus evaluating the consequences of test use became a key feature of validity in test theory. The net result was a blurring of the line between validity-as-truth and validity-as-evaluation, to the point where Cronbach (1988) said "Validation of a test or test use is evaluation" (p. 4).

We strongly endorse the legitimacy of questions about the use of both tests and experiments. Although scientists have frequently avoided value questions in the mistaken belief that they cannot be studied scientifically or that science is value free, we cannot avoid values even if we try. The conduct of experiments involves values at every step, from question selection through the interpretation and reporting of results. Concerns about the uses to which experiments and their results are put and the value of the consequences of those uses are all important (e.g., Shadish et al., 1991), as we illustrated in Chapter 9 in discussing ethical concerns with experiments.

However, if validity is to retain its primary association with the truth of knowledge claims, then it is fundamentally impossible to validate an action because actions are not knowledge claims. Actions are more properly evaluated, not validated. Suppose an employer administers a test, intending to use it in hiring decisions. Suppose the action is that a person is hired. The action is not itself a knowledge claim and therefore cannot be either true or false. Suppose that person then physically assaults a subordinate. That consequence is also not a knowledge claim and so also cannot be true or false. The action and the consequences merely exist; they are ontological entities, not epistemological ones. Perhaps Messick (1989) really meant to ask whether *inferences* about actions and consequences are true or false. If so, the inclusion of action in his (1989) definition of validity is entirely superfluous, for validity-as-truth is already about evidence in support of inferences, including those about action or consequences.[2]

[2] Perhaps partly in recognition of this, the most recent version of the test standards (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999) helps resolve some of the problems outlined herein by removing reference to validating action from the definition of validity: "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9).

Alternatively, perhaps Messick (1989, 1995) meant his definition to instruct test validators to *evaluate* the action or its consequences, as intimated in: "Validity is broadly defined as nothing less than an evaluative summary of both the evidence for and the actual—as well as potential—consequences of score interpretation and use" (1995, p. 742). Validity-as-truth certainly plays a role in evaluating tests and experiments. But we must be clear about what that role is and is not. Philosophers (e.g., Scriven, 1980; Rescher, 1969) tell us that a judgment about the value of something requires that we (1) select criteria of merit on which the thing being evaluated would have to perform well, (2) set standards of performance for how well the thing must do on each criterion to be judged positively, (3) gather pertinent data about the thing's performance on the criteria, and then (4) integrate the results into one or more evaluative conclusions. Validity-as-truth is one (but only one) criterion of merit in evaluation; that is, it is good if inferences about a test are true, just as it is good for the causal inference made from an experiment to be true. However, validation is not isomorphic with evaluation. First, criteria of merit for tests (or experiments) are not limited to validity-as-truth. For example, a good test meets other criteria, such as having a test manual that reports norms, being affordable for the contexts of application, and protecting confidentiality as appropriate. Second, the theory of validity Messick proposed gives no help in accomplishing some of the other steps in the four-step evaluation process outlined previously. To evaluate a test, we need to know something about how much validity the inference should have to be judged good; and we need to know how to integrate results from all the other criteria of merit along with validity into an overall evaluation. It is not a flaw in validity theory that these other steps are not addressed, for they are the domain of evaluation theory. The latter tells us something about how to execute these steps (e.g., Scriven, 1980, 1991) and also about other matters to be taken into account in the evaluation. Validation is not evaluation; truth is not value.

Of course, the definition of terms is partly arbitrary. So one might respond that one should be able to conflate validity-as-truth and validity-as-evaluation if one so chooses. However:

> The very fact that terms must be supplied with arbitrary meanings requires that words be used with a great sense of responsibility. This responsibility is twofold: first, to established usage; second, to the limitations that the definitions selected impose on the user. (Goldschmidt, 1982, p. 642)

We need the distinction between truth and value because true inferences can be about bad things (the fact that smoking causes cancer does not make smoking or cancer good); and false inferences can lead to good things (the astrologer's advice to Pisces to "avoid alienating your coworkers today" may have nothing to do with heavenly bodies, but may still be good advice). Conflating truth and value can be actively harmful. Messick (1995) makes clear that the social consequences of testing are to be judged in terms of "bias, fairness, and distributive justice" (p. 745). We agree with this statement, but this is test evaluation, not test validity. Messick notes that his intention is not to open the door to the social policing of truth (i.e., a test is valid if its social consequences are good), but ambiguity on this issue has nonetheless opened this very door. For example, Kirkhart (1995) cites Messick as justification for judging the validity of evaluations by their social consequences: "Consequential validity refers here to the soundness of change exerted on systems by evaluation and the extent to which those changes are just" (p. 4). This notion is risky because the most powerful arbiter of the soundness and justice of social consequences is the sociopolitical system in which we live. Depending on the forces in power in that system at any given time, we may find that what counts as valid is effectively determined by the political preferences of those with power.

Validity in the Qualitative Traditions

One of the most important developments in recent social research is the expanded use of qualitative methods such as ethnography, ethnology, participant observation, unstructured interviewing, and case study methodology (e.g., Denzin & Lincoln, 2000). These methods have unrivaled strengths for the elucidation of meanings, the in-depth description of cases, the discovery of new hypotheses, and the description of how treatment interventions are implemented or of possible causal explanations. Even for those purposes for which other methods are usually preferable, such as for making the kinds of descriptive causal inferences that are the topic of this book, qualitative methods can often contribute helpful knowledge and on rare occasions can be sufficient (Campbell, 1975; Scriven, 1976). Whenever resources allow, field experiments will benefit from including qualitative methods both for the primary benefits they are capable of generating and also for the assistance they provide to the descriptive causal task itself. For example, they can uncover important site-specific threats to validity and also contribute to explaining experimental results in general and perplexing outcome patterns in particular.

However, the flowering of qualitative methods has often been accompanied by theoretical and philosophical controversy, often referred to as the qualitative-quantitative debates. These debates concern not just methods but roles and rewards within science, ethics and morality, and epistemologies and ontologies. As part of the latter, the concept of validity has received considerable attention (e.g., Eisenhart & Howe, 1992; Goetz & LeCompte, 1984; Kirk & Miller, 1986; Kvale, 1989; J. Maxwell, 1992; J. Maxwell & Lincoln, 1990; Mishler, 1990; Phillips, 1987; Wolcott, 1990). Notions of validity that are different from ours have occasionally resulted from qualitative work, and sometimes validity is rejected entirely. However, before we review those differences we prefer to emphasize the commonalities that we think dominate on all sides of the debates.

Commonalities. As we read it, the predominant view among qualitative theorists is that validity is a concept that is and should be applicable to their work. We start with examples of discussions of validity by qualitative theorists that illustrate these similarities because they are surprisingly more common than some portrayals in the qualitative-quantitative debates suggest and because they demonstrate an underlying unity of interest in producing valid knowledge that we believe is widely shared by most social scientists. For example, Maxwell (1990) says, "qualitative researchers are just as concerned as quantitative ones about 'getting it wrong,' and validity broadly defined simply refers to the possible ways one's account might be wrong, and how these 'validity threats' can be addressed" (p. 505). Even those qualitative theorists who say they reject the word *validity* will admit that they "go to considerable pains not to get it all wrong" (Wolcott, 1990, p. 127). Kvale (1989) ties validity directly to truth, saying "concepts of validity are rooted in more comprehensive epistemological assumptions of the nature of true knowledge" (p. 11); and later that validity "refers to the truth and correctness of a statement" (p. 73). Kirk and Miller (1986) say "the technical use of the term 'valid' is as a properly hedged weak synonym for 'true' " (p. 19). Maxwell (1992) says "Validity, in a broad sense, pertains to this relationship between an account and something outside that account" (p. 283). All these seem quite compatible with our understanding of validity.

Maxwell's (1992) account points to other similarities. He claims that validity is always relative to "the kinds of understandings that accounts can embody" (p. 284) and that different communities of inquirers are interested in different kinds of understandings. He notes that qualitative researchers are interested in five kinds of understandings about: (1) the descriptions of what was seen and heard, (2) the meaning of what was seen and heard, (3) theoretical constructions that characterize what was seen and heard at higher levels of abstraction, (4) generalization of accounts to other persons, times, or settings than originally studied, and (5) evaluations of the objects of study (Maxwell, 1992; he says that the last two understandings are of interest relatively rarely in qualitative work). He then proposes a five-part validity typology for qualitative researchers, one for each of the five understandings. We agree that validity is relative to understanding, though we usually refer to inference rather than understanding. And we agree that different communities of inquirers tend to be interested in different kinds of understandings, though common interests are illustrated by the apparently shared concerns that both experimenters and qualitative researchers have in how best to characterize what was seen and heard in a study (Maxwell's theoretical validity and our construct validity). Our extended discussion of internal validity reflects the interest of the community of experimenters in understanding descriptive causes, proportionately more so than is relevant to qualitative researchers, even when their reports are necessarily replete with the language of causation. This observation is not a criticism of qualitative researchers, nor is it a criticism of experimenters as being less interested than qualitative researchers in thick description of an individual case.

On the other hand, we should not let differences in prototypical tendencies across research communities blind us to the fact that when a particular understanding *is* of interest, the pertinent validity concerns are the same no matter what the methodology used to develop the knowledge claim. It would be wrong for a qualitative researcher to claim that internal validity is irrelevant to qualitative methods. Validity is not a property of methods but of inferences and knowledge claims. On those infrequent occasions in which a qualitative researcher has a strong interest in a local molar causal inference, the concerns we have outlined under internal validity pertain. This argument cuts both ways, of course. An experimenter who wonders what the experiment means to participants could learn a lot from the concerns that Maxwell outlines under interpretive validity.

Maxwell (1992) also points out that his validity typology suggests threats to validity about which qualitative researchers seek "evidence that would allow them to be ruled-out . . . using a logic similar to that of quasi-experimental researchers such as Cook and Campbell" (p. 296). He does not outline such threats himself, but his description allows one to guess what some might look like. To judge from Maxwell's prose, threats to descriptive validity include errors of commission (describing something that did not occur), errors of omission (failing to describe something that did occur), errors of frequency (misstating how often something occurred), and interrater disagreement about description. Threats to the validity of knowledge claims have also been invoked by qualitative theorists other than Maxwell—for example, by Becker (1979), Denzin (1989), and Goetz and LeCompte (1984). Our only significant disagreement with Maxwell's discussion of threats is his claim that qualitative researchers are less able to use "design features" (p. 296) to deal with threats to validity. For instance, his preferred use of multiple observers *is* a qualitative design feature that helps to reduce errors of omission, commission, and frequency. The repertoire of design features that qualitative researchers use will usually be quite different from those used by researchers in other traditions, but they are design features (methods) all the same.

Differences. These agreements notwithstanding, many qualitative theorists approach validity in ways that differ from our treatment. A few of these differences are based on arguments that are simply erroneous (Heap, 1995; Shadish, 1995a). But many are thoughtful and deserve more attention than our space constraints allow. Following is a sample.

Some qualitative theorists either mix together evaluative and social theories of truth (Eisner, 1979, 1983) or propose to substitute the social for the evaluative. So Jensen (1989) says that validity refers to whether a knowledge claim is "meaningful and relevant" (p. 107) to a particular language community; and Guba and Lincoln (1982) say that truth can be reduced to whether an account is credible to those who read it. Although we agree that social and evaluative theories complement each other and are both helpful, replacing the evaluative with the social is misguided. These social alternatives allow for devastating counterexamples (Phillips, 1987): the swindler's story is coherent but fraudulent; cults convince members of beliefs that have little or no apparent basis otherwise; and an account of an interaction between teacher and student might be true even if neither found it to be credible. Bunge (1992) shows how one cannot define the basic idea of erqualitative researcher to claim that internal validity is irrelevant to qualitative methods. Validity is not a property of methods but of inferences and knowledge claims. On those infrequent occasions in which a qualitative researcher has a strong interest in a local molar causal inference, the concerns we have outlined under internal validity pertain. This argument cuts both ways, of course. An experimenter who wonders what the experiment means to participants could learn a lot from the concerns that Maxwell outlines under interpretive validity.

Maxwell (1992) also points out that his validity typology suggests threats to validity about which qualitative researchers seek "evidence that would allow them to be ruled-out . . . using a logic similar to that of quasi-experimental researchers such as Cook and Campbell" (p. 296). He does not outline such threats himself, but his description allows one to guess what some might look like. To judge from Maxwell's prose, threats to descriptive validity include errors of commission (describing something that did not occur), errors of omission (failing to describe something that did occur), errors of frequency (misstating how often something occurred), and interrater disagreement about description. Threats to the validity of knowledge claims have also been invoked by qualitative theorists other than Maxwell—for example, by Becker (1979), Denzin (1989), and Goetz and LeCompte (1984). Our only significant disagreement with Maxwell's discussion of threats is his claim that qualitative researchers are less able to use "design features" (p. 296) to deal with threats to validity. For instance, his preferred use of multiple observers *is* a qualitative design feature that helps to reduce errors of omission, commission, and frequency. The repertoire of design features that qualitative researchers use will usually be quite different from those used by researchers in other traditions, but they are design features (methods) all the same.

Differences. These agreements notwithstanding, many qualitative theorists approach validity in ways that differ from our treatment. A few of these differences are based on arguments that are simply erroneous (Heap, 1995; Shadish, 1995a). But many are thoughtful and deserve more attention than our space constraints allow. Following is a sample.

Some qualitative theorists either mix together evaluative and social theories of truth (Eisner, 1979, 1983) or propose to substitute the social for the evaluative. So Jensen (1989) says that validity refers to whether a knowledge claim is "meaningful and relevant" (p. 107) to a particular language community; and Guba and Lincoln (1982) say that truth can be reduced to whether an account is credible to those who read it. Although we agree that social and evaluative theories complement each other and are both helpful, replacing the evaluative with the social is misguided. These social alternatives allow for devastating counterexamples (Phillips, 1987): the swindler's story is coherent but fraudulent; cults convince members of beliefs that have little or no apparent basis otherwise; and an account of an interaction between teacher and student might be true even if neither found it to be credible. Bunge (1992) shows how one cannot define the basic idea of error using social theories of truth. Kirk and Miller (1986) capture the need for an evaluative theory of truth in qualitative methods:

> In response to the propensity of so many nonqualitative research traditions to use such hidden positivist assumptions, some social scientists have tended to overreact by stressing the possibility of alternative interpretations of everything to the exclusion of any effort to choose among them. This extreme relativism ignores the other side of objectivity—that there is an external world at all. It ignores the distinction between knowledge and opinion, and results in everyone having a separate insight that cannot be reconciled with anyone else's. (p. 15)

A second difference refers to equating the validity of knowledge claims with their evaluation, as we discussed earlier with test theory (e.g., Eisenhart & Howe, 1992). This is most explicit in Salner (1989), who suggested that much of validity in qualitative methodology concerns the criteria "that are useful for evaluating competing claims" (p. 51); and she urges researchers to expose the moral and value implications of research, much as Messick (1989) said in reference to test theory. Our response is the same as for test theory. We endorse the need to evaluate knowledge claims broadly, including their moral implications; but this is not the same as saying that the claim is true. Truth is just one criterion of merit for a good knowledge claim.

A third difference makes validity a result of the process by which truth emerges. For instance, emphasizing the dialectic process that gives rise to truth, Salner (1989) says: "Valid knowledge claims emerge . . . from the conflict and differences between the contexts themselves as these differences are communicated and negotiated among people who share decisions and actions" (p. 61). Miles and Huberman (1984) speak of the problem of validity in qualitative methods being an insufficiency of "analysis procedures for qualitative data" (p. 230). Guba and Lincoln (1989) argue that trustworthiness emerges from communication with other colleagues and stakeholders. The problem with all these positions is the error of thinking that validity is a property of methods. Any procedure for generating knowledge can generate invalid knowledge, so in the end it is the knowledge claim itself that must be judged. As Maxwell (1992) says, "The validity of an account is inherent, not in the procedures used to produce and validate it, but in its relationship to those things it is intended to be an account of" (p. 281).

A fourth difference suggests that traditional approaches to validity must be reformulated for qualitative methods because validity "historically arose in the context of experimental research" (Eisenhart & Howe, 1992, p. 644). Others reject validity for similar reasons except that they say that validity arose in test theory (e.g., Wolcott, 1990). Both are incorrect, for validity concerns probably first arose systematically in philosophy, preceding test theory and experimental science by hundreds or thousands of years. Validity is pertinent to any discussion of the warrant for believing knowledge and is not specific to particular methods.

A fifth difference concerns the claim that there is no ontological reality at all, so there is no truth to correspond to it. The problems with this perspective are enormous (Schmitt, 1995). First, even if it were true, it would apply only to correspondence theories of truth; coherence and pragmatist theories would be unaffected. Second, the claim contradicts our experience. As Kirk and Miller (1986) put it:

> There is a world of empirical reality out there. The way we perceive and understand that world is largely up to us, but the world does not tolerate all understandings of it equally (so that the individual who believes he or she can halt a speeding train with his or her bare hands may be punished by the world for acting on that understanding). (p. 11)

Third, the claim ignores evidence about the problems with people's constructions. Maxwell notes that "one of the fundamental insights of the social sciences is that people's constructions are often systematic distortions of their actual situation" (p. 506). Finally, the claim is self-contradictory because it implies that the claim itself cannot be true.

A sixth difference is the claim that it makes no sense to speak of truth because there are many different realities, with multiple truths to match each (Filstead, 1979; Guba & Lincoln, 1982; Lincoln & Guba, 1985). Lincoln (1990), for example, says that "a realist philosophical stance requires, indeed demands, a singular reality, and therefore a singular truth" (p. 502), which she juxtaposes against her own assumption of multiple realities with multiple truths. Whatever the merits of the underlying ontological arguments, this is not an argument against validity. Ontological realism (a commitment that "something" does exist) does not require a singular reality, but merely a commitment that there be at least one reality. To take just one example, physicists have speculated that there may be circumstances under which multiple physical realities could exist in parallel, as in the case of Schrodinger's cat (Davies, 1984; Davies & Brown, 1986). Such circumstances would in no way constitute an objection to pursuing valid characterizations of those multiple realities. Nor for that matter would the existence of multiple realities require multiple truths; physicists use the same principles to account for the multiple realities that might be experienced by Schrodinger's cat. Epistemological realism (a commitment that our knowledge reflects ontological reality) does not require only one true account of that world(s), but only that there not be two contradictory accounts that are both true of the same ontological referent.[3] How many realities there might be, and how many truths it takes to account for them, should not be decided by fiat.

A seventh difference objects to the belief in a monolithic or absolute Truth (with capital T). Wolcott (1990) says, "What I seek is something else, a quality that points more to identifying critical elements and wringing plausible interpretations from them, something one can pursue without becoming obsessed with

[3] The fact that different people might have different beliefs about the same referent is sometimes cited as violating this maxim, but it need not do so. For example, if the knowledge claim being validated is "John views the program as effective but Mary views it as ineffective," the claim can be true even though the views of John and Mary are contradictory.

finding the right or ultimate answer, the correct version, the Truth" (p. 146). He describes "the critical point of departure between quantities-oriented and qualities-oriented research [as being that] we cannot 'know' with the former's satisfying levels of certainty" (p. 147). Mishler (1990) objects that traditional approaches to validation are portrayed "as universal, abstract guarantors of truth" (p. 420). Lincoln (1990) thinks that "the realist position demands absolute truth" (p. 502). However, it is misguided to attribute beliefs in certainty or absolute truth to approaches to validity such as that in this book. We hope we have made clear by now that there are no guarantors of valid inferences. Indeed, the more experience that most experimenters gain, the more they appreciate the ambiguity of their results. Albert Einstein once said, "An experiment is something everybody believes except the person who made it" (Holton, 1986, p. 13). Like Wolcott, most experimenters seek only to wring plausible interpretations from their work, believing that "prudence sat poised between skepticism and credulity" (Shapin, 1994, p. xxix). We need not, should not, and frequently cannot decide that one account is absolutely true and the other completely false. To the contrary, tolerance for multiple knowledge constructions is a virtual necessity (Lakatos, 1978) because evidence is frequently inadequate to distinguish between two well-supported accounts (is light a particle or wave?), and sometimes accounts that appear to be unsupported by evidence for many years turn out to be true (do germs cause ulcers?).

An eighth difference claims that traditional understandings of validity have moral shortcomings. The arguments here are many, for example, that it "forces issues of politics, values (social and scientific), and ethics to be submerged" (Lincoln, 1990, p. 503) and implicitly empowers "social science 'experts' . . . whose class preoccupations (primarily White, male, and middle-class) ensure status for some voices while marginalizing . . . those of women, persons of color, or minority group members" (Lincoln, 1990, p. 502). Although these arguments may be overstated, they contain important cautions. Recall the example in Chapter 3 that "Even the rats were white males" in health research. No doubt this bias was partly due to the dominance of White males in the design and execution of health research. None of the methods discussed in this book are intended to redress this problem or are capable of it. The purpose of experimental design is to elucidate causal inferences more than moral inferences. What is less clear is that this problem requires abandoning notions of validity or truth. The claim that traditional approaches to truth forcibly submerge political and ethical issues is simply wrong. To the extent that morality is reflected in the questions asked, the assumptions made, and the outcomes examined, experimenters can go a long way by ensuring a broad representation of stakeholder voices in study design. Further, moral social science requires commitment to truth. Moral righteousness without truthful analysis is the stuff of totalitarianism. Moral diversity helps prevent totalitarianism, but without the discipline provided by truth-seeking, diversity offers no means to identify those options that are good for the human condition, which is, after all, the essence of morality. In order to have a moral social science, we must have both the capacity to elucidate personal constructions and the capacity to see how those constructions reflect and distort reality (Maxwell, 1992). We embrace the moral aspirations of scholars such as Lincoln, but giving voice to those aspirations simply does not require us to abandon such notions as validity and truth.

QUASI-EXPERIMENTATION

Criteria for Ruling Out Threats: The Centrality of Fuzzy Plausibility

In a randomized experiment in which all groups are treated in the same way except for treatment assignment, very few assumptions need to be made about sources of bias. And those that are made are clear and can be easily tested, particularly as concerns the fidelity of the original assignment process and its subsequent maintenance. Not surprisingly, statisticians prefer methods in which the assumptions are few, transparent, and testable. Quasi-experiments, however, rely heavily on researcher judgments about assumptions, especially on the fuzzy but indispensable concept of plausibility. Judgments about plausibility are needed for deciding which of the many threats to validity are relevant in a given study, for deciding whether a particular design element is capable of ruling out a given threat, for estimating by how much the bias might have been reduced, and for assessing whether multiple threats that might have been only partially adjusted for might add up to a total bias greater than the effect size the researcher is inclined to claim. With quasi-experiments, the relevant assumptions are numerous, their plausibility is less evident, and their single and joint effects are less easily modeled. We acknowledge the fuzzy way in which particular internal validity threats are often ruled out, and it is because of this that we too prefer randomized experiments (and regression discontinuity designs) over most of their quasi-experimental alternatives.

But quasi-experiments vary among themselves with respect to the number, transparency, and testability of assumptions. Indeed, we deliberately ordered the chapters on quasi-experiments to reflect the increase in inferential power that comes from moving from designs without a pretest or without a comparison group to those with both, to those based on an interrupted time series, and from there to regression discontinuity and random assignment. Within most of these chapters we also illustrated how inferences can be improved by adding design elements—more pretest observation points, better stable matching, replication and systematic removal of the treatment, multiple control groups, and nonequivalent dependent variables. In a sense, the plan of the four chapters on quasi-experiments reflects two purposes. One is to show how the number, transparency, and testability of assumptions varies by type of quasi-experimental design so that, in the best of quasi-experiments, internal validity is not much worse than with the randomized experiment. The other is to get students of quasi-experiments to be more sparing with the use of this overly general label, for it threatens to tar all quasi-experiments with the same negative brush. As scholars who have contributed to the institutionalization of the term *quasi-experiment,* we feel a lot of ambivalence about our role. Scholars need to think critically about alternatives to the randomized experiment, and from this need arises the need for the quasi-experimental label. But all instances of quasi-experimental design should not be brought under the same unduly broad quasi-experimental umbrella if attributes of the best studies do not closely match the weaker attributes of the field writ large.

Statisticians seek to make their assumptions transparent through the use of formal models laid out as formulae. For the most part, we have resisted this strategy because it backfires with so many readers, alienating them from the very conceptual issues the formulae are designed to make evident. We have used words instead. There is a cost to this, and not juşt in the distaste of statistical cognoscenti, particularly those whose own research has emphasized statistical models. The main cost is that our narrative approach makes it more difficult to formally demonstrate how much fewer and more evident and more testable the alternative interpretations became as we moved from the weaker to the stronger quasi-experiments, both within the relevant quasi-experimental chapters and across the set of them. We regret this, but do not apologize for the accessibility we tried to create by minimizing the use of Greek symbols and Roman subscripts. Fortunately, this deficit is not absolute, as both we and others have worked to develop methods that can be used to measure the size of particular threats, both in particular studies (e.g., Gastwirth et al., 1994; Shadish et al., 1998; Shadish, 2000) and in sets of studies (e.g., Kazdin & Bass, 1989; Miller, Turner, Tindale, Posavac, & Dugoni, 1991; Rosenthal & Rubin, 1978; Willson & Putnam, 1982). Further, our narrative approach has a significant advantage over a more narrowly statistical emphasis—it allows us to address a broader array of qualitatively different threats to validity, threats for which no statistical measure is yet available and that therefore might otherwise be overlooked with too strict an emphasis on quantification. Better to have imprecise attention to plausibility than to have no attention at all paid to many important threats just because they cannot be well measured.

Pattern Matching as a Problematic Criterion

This book is more explicit than its predecessors about the desirability of imbuing a causal hypothesis with multiple testable implications in the data, provided that they serve to reduce the viability of alternative causal explanations. In a sense, we have sought to substitute a pattern-matching methodology for the usual assessment of whether a few means, often only two, reliably differ. We do this not because complexity itself is a desideratum in science. To the contrary, simplicity in the number of questions asked and methods used is highly prized in science. The simplicity of randomized experiments for descriptive causal inference illustrates this well. However, the same simple circumstance does not hold with quasi-experiments. With them, we have asserted that causal inference is improved the more specific, generating these lists. The main concern was to have a consensus of education researchers endorsing each practice; and he guessed that the number of these best practices that depended on randomized experiments would be zero. Several nationally known educational researchers were present, agreed that such assignment probably played no role in generating the list, and felt no distress at this. So long as the belief is widespread that quasi-experiments constitute the summit of what is needed to support causal conclusions, the support for experimentation that is currently found in health, agriculture, or health in schools is unlikely to occur. Yet randomization is possible in many educational contexts within schools if the will exists to carry it out (Cook et al., 1999; Cook et al., in press). An unfortunate and inadvertent side effect of serious discussion of quasi-experiments may sometimes be the practical neglect of randomized experiments. That is a pity.

RANDOMIZED EXPERIMENTS

This section lists objections that have been raised to doing randomized experiments, and our analysis of the more and less legitimate issues that these objections raise.

Experiments Cannot Be Successfully Implemented

Even a little exposure to large-scale social experimentation shows that treatments are often improperly or incompletely implemented and that differential attrition often occurs. Organizational obstacles to experiments are many. They include the reality that different actors vary in the priority they attribute to random assignment, that some interventions seem disruptive at all levels of the organization, and that those at the point of service delivery often find the treatment requirements a nuisance addition to their already overburdened daily routine. Then there are sometimes treatment crossovers, as units in the control condition adopt or adapt components from the treatment or as those in a treatment group are exposed to some but not all of these same components. These criticisms suggest that the correct comparison is not between the randomized experiment and better quasi-experiments when each is implemented perfectly but rather between the randomized experiment as it is often imperfectly implemented and better quasi-experiments. Indeed, implementation can sometimes be better in the quasi-experiment if the decision not to randomize is based on fears of treatment degradation. This argument cannot be addressed well because it depends on specifying the nature and degree of degradation and the kind of quasi-experimental alternative. But taken to its extreme it suggests that randomized experiments have no special warrant in field settings because there is no evidence that they are stronger than other designs *in practice* (only in theory).

But the situation is probably not so bleak. Methods for preventing and coping with treatment degradation are improving rapidly (see Chapter 10, this volume; Boruch, 1997; Gueron, 1999; Orr, 1999). More important, random assignment may still create a superior counterfactual to its alternatives even with the flaws mentioned herein. For example, Shadish and Ragsdale (1996) found that, compared with randomized experiments without attrition, randomized experiments with attrition still yielded better effect size estimates than did nonrandomized experiments. Sometimes, of course, an alternative to severely degraded randomization will be best, such as a strong interrupted time series with a control. But routine rejection of degraded randomized experiments is a poor rule to follow; it takes careful study and judgment to decide. Further, many alternatives to experimentation are themselves subject to treatment implementation flaws that threaten the validity of inferences from them. Attrition and treatment crossovers also occur in them. We also suspect that implementation flaws are salient in experimentation because experiments have been around so long and experimenters are so critical of each other's work. By contrast, criteria for assessing the quality of implementation and results from other methods are far more recent (e.g., Datta, 1997), and they may therefore be less well developed conceptually, less subjected to peer criticism, and less improved by the lessons of experience.

Experimentation Needs Strong Theory and Standardized Treatment Implementation

Many critics claim that experimentation is more fruitful when an intervention is based on strong substantive theory, when implementation of treatment details is faithful to that theory, when the research setting is well managed, and when implementation does not vary much between units. In many field experiments, these conditions are not met. For example, schools are large, complex, social organizations with multiple programs, disputatious politics, and conflicting stakeholder goals. Many programs are implemented variably across school districts, as well as across schools, classrooms, and students. There can be no presumption of standard implementation or fidelity to program theory (Berman & McLaughlin, 1977).

But these criticisms are, in fact, misplaced. Experiments do not require well-specified program theories, good program management, standard implementation, or treatments that are totally faithful to theory. Experiments make a contribution when they simply probe whether an intervention-as-implemented makes a marginal improvement beyond other background variability. Still, the preceding factors can reduce statistical power and so cloud causal inference. This suggests that in settings in which more of these conditions hold, experiments should: (1) use large samples to detect effects; (2) take pains to reduce the influence of extraneous variation either by design or through measurement and statistical manipulation; and (3) study implementation quality both as a variable worth studying in its own right in order to ascertain which settings and providers implement the intervention better and as a mediator to see how implementation carries treatment effects to outcome.

Indeed, for many purposes the lack of standardization may aid in understanding how effective an intervention will be under normal conditions of implementation. In the social world, few treatments are introduced in a standard and theory-faithful way. Local adaptations and partial implementation are the norm. If this is the case, then some experiments should reflect this variation and ask whether the treatment can continue to be effective despite all the variation within groups that we would expect to find if the treatment were policy. Program developers and social theorists may want standardization at high levels of implementation, but policy analysts should not welcome this if it makes the research conditions different from the practice conditions to which they would like to generalize. Of course, it is most desirable to be able to answer both sets of questions—about policy-relevant effects of treatments that are variably implemented and also about the more theory-relevant effects of optimal exposure to the intervention. In this regard, one might recall recent efforts to analyze the effects of the original intent to treat through traditional means but also of the effects of the actual treatment through using random assignment as an instrumental variable (Angrist et al., 1996a).

Experiments Entail Tradeoffs Not Worth Making

The choice to experiment involves a number of tradeoffs that some researchers believe are not worth making (Cronbach, 1982). Experimentation prioritizes on unbiased answers to descriptive causal questions. But, given finite resources, some researchers prefer to invest what they have not into marginal improvements in internal validity but into promoting higher construct and external validity. They might be content with a greater degree of uncertainty about the quality of a causal connection in order to purposively sample a greater range of populations of people or settings or, when a particular population is central to the research, in order to generate a formally representative sample. They might even use the resources to improve treatment fidelity or to include multiple measures of a very important outcome construct. If a consequence of this preference for construct and external validity is to conduct a quasi-experiment or even a nonexperiment rather than a randomized experiment, then so be it. Similar preferences make other critics look askance when advocates of experimentation counsel restricting a study to volunteers in order to increase the chances of being able to implement and maintain random assignment or when these same advocates advise close monitoring of the treatment to ensure its fidelity, thereby creating a situation of greater obtrusiveness than would pertain if the same treatment were part of some ongoing social policy (e.g., Heckman, 1992). In the language of Campbell and Stanley (1963), the claim was that experimentation traded off external validity in favor of internal validity. In the parlance of this book and of Cook and Campbell (1979), it is that experimentation trades off both external and construct validity for internal validity, to its detriment.

Critics also claim that experiments overemphasize conservative standards of scientific rigor. These include (1) using a conservative criterion to protect against wrongly concluding a treatment is effective ($p < .05$) at the risk of failing to detect true treatment effects; (2) recommending intent-to-treat analyses that include as part of the treatment those units that have never received treatment; (3) denigrating inferences that result from exploring unplanned treatment interactions with characteristics of units, observations, settings, or times; and (4) rigidly pursuing a priori experimental questions when other interesting questions emerge during a study. Most laypersons use a more liberal risk calculus to decide about causal inferences in their own lives, as when they consider taking up some potentially lifesaving therapy. Should not science do the same, be less conservative? Should it not at least sometimes make different tradeoffs between protection against incorrect inferences and the failure to detect true effects?

Critics further object that experiments prioritize descriptive over explanatory causation. The critics in question would tolerate more uncertainty about whether the intervention works in order to learn more about any explanatory processes that have the potential to generalize across units, settings, observations, and times. Further, some critics prefer to pursue this explanatory knowledge using qualitative methods similar to those of the historian, journalist, and ethnographer than by means of, say, structural equation modeling that seems much more opaque than the narrative reports of these other fields.

Critics also dislike the priority that experiments give to providing policymakers with often belated answers about what works instead of providing real-time help to service providers in local settings. These providers are rarely interested in a long-delayed summary of what a program has achieved. They often prefer receiving continuous feedback about their work and especially about those elements of practice that they can change without undue complication. A recent letter to the *New York Times* captured this preference:

> Alan Krueger . . . claims to eschew value judgments and wants to approach issues (about educational reform) empirically. Yet his insistence on postponing changes in education policy until studies by researchers approach certainty is itself a value judgment in favor of the status quo. In view of the tragic state of affairs in parts of public education, his judgment is a most questionable one. (Petersen, 1999)

We agree with many of these criticisms. Among all possible research questions, causal questions constitute only a subset. And of all possible causal methods, experimentation is not relevant to all types of questions and all types of circumstance. One need only read the list of options and contingencies outlined in Chapters 9 and 10 to appreciate how foolhardy it is to advocate experimentation on a routine basis as a causal "gold standard" that will invariably result in clearly interpretable effect sizes. However, many of the criticisms about tradeoffs are based on artificial dichotomies, correctable problems, and even oversimplifications. Experiments can and should examine reasons for variable implementation, and they should search to uncover mediating processes. They need not use stringent alpha rates; only statistical tradition argues for the .05 level. Nor need one restrict data analyses only to the intent-to-treat, though that should definitely be one analysis. Experimenters can also explore statistical interactions to the extent that substantive theory and statistical power allow, guarding against profligate error rates and couching their conclusions cautiously. Interim results from experiments can be published. There can and should also be nonexperimental analyses of the representativeness of samples and the construct validity of assessments of persons, settings, treatments, and outcomes. There can and should be qualitative data collection aimed at discovering unintended outcomes and mediating processes. And as much information as possible about causal generalization should be generated using the methods outlined in this book. All these procedures require resources, but sometimes few of them (e.g., adding measures of mediating variables). Experiments need not be as rigid as some texts suggest, and the goal of ever-finer marginal improvements in internal validity is often a poor one.

In the latter regard, some critics have claimed that more useful information will be learned from programs of research that consist mostly or even entirely of quasi-experimental and nonexperimental studies than from programs emphasizing the stronger experimental methods (e.g., Cronbach et al., 1980; Cronbach, 1982). Although we are generally sympathetic to this point, some bounds cannot be crossed without compromising the integrity of key inferences. Unless threats to internal validity are clearly implausible on logical or evidential grounds, to have *no* strong experimental studies on the effects of an intervention is to risk drawing broad general conclusions about a causal connection that is undependable. This happens all too often, alas. It is now 30 years since school vouchers were proposed, and we still have no clear answers about their effects. It is 15 years since Henry Levin began accelerated schools, and we have no experiments and no answers. It is 30 years since James Comer began the School Development Program, and almost the same situation holds. Although premature experimentation is a real danger, such decade-long time lines without clear answers are probably even more problematic, particularly for those legislators and their staffs who want to promote effectiveness-based social policies. Finding out what works is too important to suggest that experiments require tradeoffs that are *never* worth making.

By contrast, we are impressed with the capacity of programs of experimental research to address both construct and external validity issues modestly well. Granted, individual experiments have limited reach in addressing both these issues, but as we see most clearly in meta-analysis, the capacity to address both construct and external validity issues over multiple experiments greatly exceeds what past critics have suggested. Of course, as we made clear in Chapter 2, we are not calling for any routine primacy of internal validity over construct or external validity (every validity type must have its time in the spotlight). Rather, we are calling attention to the inferential weaknesses that history suggests have emerged in programs of research that deemphasize internal validity too much and to the surprisingly broad inferential reach of programs in which internal validity plays a much more prominent role.

Experiments Assume an Invalid Model of Research Utilization

To some critics, experiments recreate a naïve rational choice model of decision making. That is, one first lays out the alternatives to choose among (the treatments); then one decides on criteria of merit (the outcomes); then one collects information on each criterion for each treatment (the data collection), and finally one makes a decision about the superior alternative. Unfortunately, empirical work on the use of social science data shows that use is not so simple as the rational choice model suggests (C. Weiss & Bucuvalas, 1980; C. Weiss, 1988).

First, even when cause and effect questions are asked in decision contexts, experimental results are still used along with other forms of information—from existing theories, personal testimony, extrapolations from surveys, consensus of a field, claims from experts with interests to defend, and ideas that have recently become trendy. Decisions are shaped partly by ideology, interests, politics, personality, windows of opportunity, and values; and they are as much made by a policy-shaping community (Cronbach et al., 1980) as by an individual or committee. Further, many decisions are not so much made as accreted over time as earlier decisions constrain later ones, leaving the final decision maker with few options (Weiss, 1980). Indeed, by the time experimental results are available, new decision makers and issues may have replaced old ones.

Second, experiments often yield contested rather than unanimous verdicts that therefore have uncertain implications for decisions. Disputes arise about whether the causal questions were correctly framed, whether results are valid, whether relevant outcomes were assessed, and whether the results entail a specific decision. For example, reexaminations of the Milwaukee educational voucher study offered different conclusions about whether and where effects occurred (H. Fuller, 2000; Greene, Peterson, & Du, 1999; Witte, 1998, 1999, 2000). Similarly, different effect sizes were generated from the Tennessee class size experiment (Finn & Achilles, 1990; Hanushek, 1999; Mosteller, Light, & Sachs, 1996). Sometimes, scholarly disagreements are at issue, but at other times the disputes reflect deeply conflicted stakeholder interests.

Third, short-term instrumental use of experimental data is more likely when the intervention is a minor variant on existing practice. For example, it is easier to change textbooks in a classroom or pills given to patients or eligibility criteria for program entry than it is to relocate hospitals to underserved locations or to open day-care centers for welfare recipients throughout an entire state. Because the more feasible changes are so modest in scope, they are less likely to dramatically affect the problem they address. So critics note that prioritizing on short-term instrumental change tends to preserve most of the status quo and is unlikely to solve trenchant social problems. Of course, there are some experiments that truly twist the lion's tail and involve bold initiatives. Thus moving families from densely poor inner-city locations to the suburbs involved a change of three standard deviations in the poverty level of the sending and receiving communities, much greater than what happens when poor families spontaneously move. Whether such a dramatic change could ever be used as a model for cleaning out the inner cities of those who want to move is a moot issue. Many would judge such a policy to be unlikely. Truly bold experiments have many important rationales; but creating new policies that look like the treatment soon after the experiment is not one of them.

Fourth, the most frequent use of research may be conceptual rather than instrumental, changing how users think about basic assumptions, how they understand contexts, and how they organize or label ideas. Some conceptual uses are intentional, as when a person deliberately reads a book on a current problem; for example, Murray's (1984) book on social policy had such a conceptual impact in the 1980s, creating a new social policy agenda. But other conceptual uses occur in passing, as when a person reads a newspaper story referring to social research. Such uses can have great long-run impact as new ways of thinking move through the system, but they rarely change particular short-term decisions.

These arguments against a naïve rational decision-making model of experimental usefulness are compelling. That model is rightly rejected. However, most of the objections are true not just of experiments but of all social science methods. Consider controversies over the accuracy of the U.S. Census, the entirely descriptive results of which enter into a decision-making process about the apportionment of resources that is complex and highly politically charged. No method offers a direct road to short-term instrumental use. Moreover, the objections are exaggerated. In settings such as the U.S. Congress, decision making is *sometimes* influenced instrumentally by social science information (Chelimsky, 1998), and experiments frequently contribute to that use as part of a research review on effectiveness questions. Similarly, policy initiatives get recycled, as happened with school vouchers, so that social science data that were not used in past years are used later when they become instrumentally relevant to a current issue (Polsby, 1984; Quirk, 1986). In addition, data about effectiveness influence many stakeholders' thinking even when they do not use the information quickly or instrumentally. Indeed, research suggests that high-quality experiments can confer extra credibility among policymakers and decision makers (C. Weiss & Bucuvalas, 1980), as happened with the Tennessee class size study. We should also not forget that the conceptual use of experiments occurs when the texts used to train professionals in a given field contain results of past studies about successful practice (Leviton & Cook, 1983). And using social science data to produce incremental change is not always trivial. Small changes can yield benefits of hundreds of millions of dollars (Fienberg, Singer, & Tanur, 1985). Sociologist Carol Weiss, an advocate of doing research for enlightenment's sake, says that 3 decades of experience and her studies of the use of social science data leave her "impressed with the utility of evaluation findings in stimulating incremental increases in knowledge and in program effectiveness. Over time, cumulative increments are not such small potatoes after all" (Weiss, 1998, p. 319). Finally, the usefulness of experiments can be increased by the actions outlined earlier in this chapter that involve complementing basic experimental design with adjuncts such as measures of implementation and mediation or qualitative methods—anything that will help clarify program process and implementation problems. In summary, invalid models of the usefulness of experimental results seem to us to be no more nor less common than invalid models of the use of any other social science methods. We have learned much in the last several decades about use, and experimenters who want their work to be useful can take advantages of those lessons (Shadish et al., 1991).

The Conditions of Experimentation Differ from the Conditions of Policy Implementation

Experiments are often done on a smaller scale than would pertain if services were implemented state- or nationwide, and so they cannot mimic all the details relevant to full policy implementation. Hence policy implementation of an intervention may yield different outcomes than the experiment (Elmore, 1996). For example, based partly on research about the benefits of reducing class size, Tennessee and California implemented statewide policies to have more classes with fewer students in each. This required many new teachers and new classrooms. However, because of a national teacher shortage, some of those new teachers may have been less qualified than those in the experiment; and a shortage of classrooms led to more use of trailers and dilapidated buildings that may have harmed effectiveness further.

Sometimes an experimental treatment is an innovation that generates enthusiastic efforts to implement it well. This is particularly frequent when the experiment is done by a charismatic innovator whose tacit knowledge may exceed that of those who would be expected to implement the program in ordinary practice and whose charisma may induce high-quality implementation. These factors may generate more successful outcomes than will be seen when the intervention is implemented as routine policy.

Policy implementation may also yield different results when experimental treatments are implemented in a fashion that differs from or conflicts with practices in real-world application. For example, experiments studying psychotherapy outcome often standardize treatment with a manual and sometimes observe and correct the therapist for deviating from the manual (Shadish et al., 2000); but these practices are rare in clinical practice. If manualized treatment is more effective (Chambless & Hollon, 1998; Kendall, 1998), experimental results might transfer poorly to practice settings.

Random assignment may also change the program from the intended policy implementation (Heckman, 1992). For example, those willing to be randomized may differ from those for whom the treatment is intended; randomization may change people's psychological or social response to treatment compared with those who self-select treatment; and randomization may disrupt administration and implementation by forcing the program to cope with a different mix of clients.

Heckman claims this kind of problem with the Job Training Partnership Act (JTPA) evaluation "calls into question the validity of the experimental estimates as a statement about the JTPA system as a whole" (Heckman, 1992, p. 221).

In many respects, we agree with these criticisms, though it is worth noting several responses to them. First, they *assume* a lack of generalizability from experiment to policy, but that is an *empirical* question. Some data suggest that generalization may be high despite differences between lab and field (C. Anderson, Lindsay, & Bushman, 1999) or between research and practice (Shadish et al., 2000). Second, it can help to implement treatment under conditions that are more characteristic of practice if it does not unduly compromise other research priorities. A little forethought can improve the surface similarity of units, treatments, observations, settings, or times to their intended targets. Third, some of these criticisms are true of *any* research methodology conducted in a limited context, such as locally conducted case studies or quasi-experiments, because local implementation issues always differ from large-scale issues. Fourth, the potentially disruptive nature of experimentally manipulated interventions is shared by many locally invented novel programs, *even when they are not studied by any research methodology at all.* Innovation inherently disrupts, and substantive literatures are rife with examples of innovations that encountered policy implementation impediments (Shadish, 1984).

However, the essential problem remains that large-scale policy implementation is a singular event, the effects of which cannot be fully known except by doing the full implementation. A single experiment, or even a small series of similar ones, cannot provide complete answers about what will happen if the intervention is adopted as policy. However, Heckman's criticism needs reframing. He fails to distinguish among validity types (statistical conclusion, internal, construct, external). Doing so makes it clear that his claim that such criticism "calls into question the validity of the experimental estimates as a statement about the JTPA system as a whole" (Heckman, 1992, p. 221) is really about external validity and construct validity, not statistical conclusion or internal validity. Except in the narrow econometrics tradition that he understandably cites (Haavelmo, 1944; Marschak, 1953; Tinbergen, 1956), few social experimenters ever claimed that experiments could describe the "system as a whole"—even Fisher (1935) acknowledged this tradeoff. Further, the econometric solutions that Heckman suggests cannot avoid the same tradeoffs between internal and external validity. For example, surveys and certain quasi-experiments can avoid some problems by observing existing interventions that have already been widely implemented, but the validity of their estimates of program effects are suspect and may themselves change if the program were imposed even more widely as policy.

Addressing these criticisms requires multiple lines of evidence—randomized experiments of efficacy and effectiveness, nonrandomized experiments that observe existing interventions, nonexperimental surveys to yield estimates of representativeness, statistical analyses that bracket effects under diverse assumptions, qualitative observation to discover potential incompatibilities between the intervention and its context of likely implementation, historical study of the fates of similar interventions when they were implemented as policy, policy analyses by those with expertise in the type of intervention at issue, and the methods for causal generalization in this book. The conditions of policy implementation will be different from the conditions characteristic of *any* research study of it, so predicting generalization to policy will always be one of the toughest problems.

Imposing Treatments Is Fundamentally Flawed Compared with Encouraging the Growth of Local Solutions to Problems

Experiments impose treatments on recipients. Yet some late 20th-century thought suggests that imposed solutions may be inferior to solutions that are locally generated by those who have the problem. Partly, this view is premised on research findings of few effects for the Great Society social programs of the 1960s in the United States (Murray, 1984; Rossi, 1987), with the presumption that a portion of the failure was due to the federally imposed nature of the programs. Partly, the view reflects the success of late 20th-century free market economics and conservative political ideologies compared with centrally controlled economies and more liberal political beliefs. Experimentally imposed treatments are seen in some quarters as being inconsistent with such thinking.

Ironically, the first objection is based on results of experiments—if it is true that imposed programs do not work, experiments provided the evidence. Moreover, these no-effect findings may have been partly due to methodological failures of experiments as they were implemented at that time. Much progress in solving practical experimental problems occurred after, and partly in response to, those experiments. If so, it is premature to assume these experiments definitively demonstrated no effect, especially given our increased ability to detect small effects today (D. Greenberg & Shroder, 1997; Lipsey, 1992; Lipsey & Wilson, 1993).

We must also distinguish between political-economic currency and the effects of interventions. We know of no comparisons of, say, the effects of locally generated versus imposed solutions. Indeed, the methodological problems in doing such comparisons are daunting, especially accurately categorizing interventions into the two categories and unconfounding the categories with correlated method differences. Barring an unexpected solution to the seemingly intractable problems of causal inference in nonrandomized designs, answering questions about the effects of locally generated solutions may require exactly the kind of high-quality experimentation being criticized. Though it is likely that locally generated solutions may indeed have significant advantages, it also is likely that some of those solutions will have to be experimentally evaluated.

CAUSAL GENERALIZATION: AN OVERLY COMPLICATED THEORY?

Internal validity is best promoted via random assignment, an omnibus mechanism that ensures that we do not have many assumptions to worry about when causal inference is our goal. By contrast, quasi-experiments require us to make explicit many assumptions—the threats to internal validity—that we then have to rule out by fiat, by design, or by measurement. The latter is a more complex and assumption-riddled process that is clearly inferior to random assignment. Something similar holds for causal generalization, in which random selection is the most parsimonious and theoretically justified method, requiring the fewest assumptions when causal generalization is our goal. But because random selection is so rarely feasible, one instead has to construct an acceptable theory of generalization out of purposive sampling, a much more difficult process. We have tried to do this with our five principles of generalized causal inference. These, we contend, are the keys to generalized inference that lie behind random sampling and that have to be identified, explicated, and assessed if we are to make better general inferences, even if they are not perfect ones. But these principles are much more complex to implement than is random sampling.

Let us briefly illustrate this with the category called American adult women. We could represent this category by random selection from a critically appraised register of all women who live in the United States and who are at least 21 years of age. Within the limits of sampling error, we could formally generalize any characteristics we measured on this sample to the population on that register. Of course, we cannot select this way because no such register exists. Instead, one does one's experiment with an opportunistic sample of women. On inspection they all turn out to be between 19 and 30 years of age, to be higher than average in achievement and ability, and to be attending school—that is, we have used a group of college women. Surface similarity suggests that each is an instance of the category *woman*. But it is obvious that the modal American woman is clearly not a college student. Such students constitute an overly homogeneous sample with respect to educational abilities and achievement, socioeconomic status, occupation, and all observable and unobservable correlates thereof, including health status, current employment, and educational and occupational aspirations and expectations. To remedy this bias, we could use a more complex purposive sampling design that selects women heterogeneously on all these characteristics. But purposive sampling for heterogeneous instances can never do this as well as random selection can, and it is certainly more complex to conceive and execute. We could go on and illustrate how the other principles facilitate generalization. The point is that any theory of generalization from purposive samples is bound to be more complicated than the simplicity of random selection.

But because random selection is rarely possible when testing causal relationships within an experimental framework, we need these purposive alternatives.

Yet most experimental work probably still relies on the weakest of these alternatives, surface similarity. We seek to improve on such uncritical practice. Unfortunately, though, there is often restricted freedom for the more careful selection of instances of units, treatments, outcomes, and settings, even when the selection is done purposively. It requires resources to sample irrelevancies so that they are heterogeneous on many attributes, to measure several related constructs that can be discriminated from each other conceptually, and to measure a variety of possible explanatory processes. This is partly why we expect more progress on causal generalization from a review context rather than from single studies. Thus, if one researcher can work with college women, another can work with female schoolteachers, and another with female retirees, this creates an opportunity to see if these sources of irrelevant homogeneity make a difference to a causal relationship or whether it holds over all these different types of women.

Ultimately, causal generalization will always be more complicated than assessing the likelihood that a relationship is causal. The theory is more diffuse, more recent, and less well tested in the crucible of research experience. And in some quarters there is disdain for the issue, given the belief and practice that relationships that replicate once should be considered as general until proven otherwise, not to speak of the belief that little progress and prestige can be achieved by designing the next experiment to be some minor variant on past studies. There is no point in pretending that causal generalization is as institutionalized procedurally as other methods in the social sciences. We have tried to set the theoretical agenda in a systematic way. But we do not expect to have the last word. There is still no explication of causal generalization equivalent to the empirically produced list of threats to internal validity and the quasi-experimental designs that have evolved over 40 years to rule out these threats. The agenda is set but not complete.

NONEXPERIMENTAL ALTERNATIVES

Though this book is about experimental methods for answering questions about causal hypotheses, it is a mistake to believe that only experimental approaches are used for this purpose. In the following, we briefly consider several other approaches, indicating the major reasons why we have not dwelt on them in detail. Basically, the reason is that we believe that, whatever their merits for some research purposes, they generate less clear causal conclusions than randomized experiments or even the best quasi-experiments such as regression-discontinuity or interrupted time series.

The nonexperimental alternatives we examine are the major ones to emerge in various academic disciplines. In education and parts of anthropology and sociology, one alternative is intensive qualitative case studies. In these same fields, and also in developmental psychology, there is an emerging interest in theory-based causal studies based on causal modeling practices. Across the social sciences other than economics and statistics, the word *quasi-experiment* is routinely used to justify causal inferences, even though designs so referred to are so primitive in structure that causal conclusions are often problematic. We have to challenge such advocacy of low-grade quasi-experiments as a valid alternative to the quality of studies we have been calling for in this book. And finally, in parts of statistics and epidemiology, and overwhelmingly in econometrics and those parts of sociology and political science that draw from econometrics, the emphasis is more on control through statistical manipulation than on experimental design. When descriptive causal inferences are the primary concern, all of these alternatives will usually be inferior to experiments.

Intensive Qualitative Case Studies

The call to generate causal conclusions from intensive case studies comes from several sources. One is from quantitative researchers in education who became disenchanted with the tools of their trade and subsequently came to prefer the qualitative methods of the historian and journalist and especially of the ethnographer (e.g., Guba, 1981, 1990; and more tentatively, Cronbach, 1986). Another is from those researchers originally trained in primary disciplines such as qualitative anthropology (e.g., Fetterman, 1984) or sociology (Patton, 1980).

The enthusiasm for case study methods arises for several different reasons. One is that qualitative methods often reduce enough uncertainty about causation to meet stakeholder needs. Most advocates point out that journalists, historians, ethnographers, and lay persons regularly make valid causal inferences using a qualitative process that combines reasoning, observation, and falsificationist procedures in order to rule out threats to internal validity—even if that kind of language is not explicitly used (e.g., Becker, 1958; Cronbach, 1982). A small minority of qualitative theorists go even further to claim that case studies can routinely replace experiments for nearly any causal-sounding question they can conceive (e.g., Lincoln & Guba, 1985). A second reason is the belief that such methods can also engage a broad view of causation that permits getting at the many forces in the world and human minds that together influence behavior in much more complex ways than any experiment will uncover. And the third reason is the belief that case studies are broader than experiments in the types of information they yield. For example, they can inform readers about such useful and diverse matters as how pertinent problems were formulated by stakeholders, what the substantive theories of the intervention are, how well implemented the intervention components were, what distal, as well as proximal, effects have come about in respondents' lives, what unanticipated side effects there have been, and what processes explain the pattern of obtained results. The claim is that intensive case study methods allow probes of an A to B connection, of a broad range of factors conditioning this relationship, and of a range of intervention-relevant questions that is broader than the experiment allows.

Although we agree that qualitative evidence can reduce some uncertainty about cause—sometimes substantially—the conditions under which this occurs are usually rare (Campbell, 1975). In particular, qualitative methods usually produce unclear knowledge about the counterfactual of greatest importance, how those who received treatment would have changed without treatment. Adding design features to case studies, such as comparison groups and pretreatment observations, clearly improves causal inference. But it does so by melding case-study data collection methods with experimental design. Although we consider this as a valuable addition to ways of thinking about case studies, many advocates of the method would no longer recognize it as still being a case study. To our way of thinking, case studies are very relevant when causation is at most a minor issue; but in most other cases when substantial uncertainty reduction about causation is required, we value qualitative methods within experiments rather than as alternatives to them, in ways similar to those we outlined in Chapter 12.

Theory-Based Evaluations

This approach has been formulated relatively recently and is described in various books or special journal issues (Chen & Rossi, 1992; Connell, Kubisch, Schorr, & Weiss, 1995; Rogers, Hacsi, Petrosino, & Huebner, 2000). Its origins are in path analysis and causal modeling traditions that are much older. Although advocates have some differences with each other, basically they all contend that it is useful: (1) to explicate the theory of a treatment by detailing the expected relationships among inputs, mediating processes, and short- and long-term outcomes; (2) to measure all the constructs specified in the theory; and (3) to analyze the data to assess the extent to which the postulated relationships actually occurred. For shorter time periods, the available data may address only the first part of a postulated causal chain; but over longer periods the complete model could be involved. Thus, the priority is on highly specific substantive theory, high-quality measurement, and valid analysis of multivariate explanatory processes as they unfold in time (Chen & Rossi, 1987, 1992).

Such theoretical exploration is important. It can clarify general issues with treatments of a particular type, suggest specific research questions, describe how the intervention functions, spell out mediating processes, locate opportunities to remedy implementation failures, and provide lively anecdotes for reporting results (Weiss, 1998). All these serve to increase the knowledge yield, even when such theoretical analysis is done within an experimental framework. There is nothing about the approach that makes it an alternative to experiments. It can clearly be a very important adjunct to such studies, and in this role we heartily endorse the approach (Cook, 2000).

However, some authors (e.g., Chen & Rossi, 1987, 1992; Connell et al., 1995) have advocated theory-based evaluation as an attractive alternative to experiments when it comes to testing causal hypotheses. It is attractive for several reasons. First, it requires only a treatment group, not a comparison group whose agreement to be in the study might be problematic and whose participation increases research costs. Second, demonstrating a match between theory and data suggests the validity of the causal theory without having to go through a laborious process of explicitly considering alternative explanations. Third, it is often impractical to measure distant end points in a presumed causal chain. So confirmation of attaining proximal end points through theory-specified processes can be used in the interim to inform program staff about effectiveness to date, to argue for more program resources if the program seems to be on theoretical track, to justify claims that the program might be effective in the future on the as-yet-not-assessed distant criteria, and to defend against premature summative evaluations that claim that an intervention is ineffective before it has been demonstrated that the processes necessary for the effect have actually occurred.

However, major problems exist with this approach for high-quality descriptive causal inference (Cook, 2000). First, our experience in writing about the theory of a program with its developer (Anson et al., 1991) has shown that the theory is not always clear and could be clarified in diverse ways. Second, many theories are linear in their flow, omitting reciprocal feedback or external contingencies that might moderate the entire flow. Third, few theories specify how long it takes for a given process to affect an indicator, making it unclear if null results disconfirm a link or suggest that the next step did not yet occur. Fourth, failure to corroborate a model could stem from partially invalid measures as opposed to invalidity of the theory. Fifth, many different models can fit a data set (Glymour et al., 1987; Stelzl, 1986), so our confidence in any given model may be small. Such problems are often fatal to an approach that relies on theory to make strong causal claims. Though some of these problems are present in experiments (e.g., failure to incorporate reciprocal causation, poor measures), they are of far less import because experiments do not require a well-specified theory in constructing causal knowledge. Experimental causal knowledge is less ambitious than theory-based knowledge, but the more limited ambition is attainable.

Weaker Quasi-Experiments

For some researchers, random assignment is undesirable for practical or ethical reasons, so they prefer quasi-experiments. Clearly, we support thoughtful use of quasi-experimentation to study descriptive causal questions. Both interrupted time series and regression discontinuity often yield excellent effect estimates. Slightly weaker quasi-experiments can also yield defensible estimates, especially when they involve control groups with careful matching on stable pretest attributes combined with other design features that have been thoughtfully chosen to address contextually plausible threats to validity. However, when a researcher can choose, randomized designs are usually superior to nonrandomized designs.

This is especially true of nonrandomized designs in which little thought is given to such matters as the quality of the match when creating control groups, including multiple hypothesis tests rather than a single one, generating data from several pretreatment time points rather than one, or having several comparison groups to create controls that bracket performance in the treatment groups. Indeed, when results from typical quasi-experiments are compared with those from randomized experiments on the same topic, several findings emerge. Quasi-experiments frequently misestimate effects (Heinsman & Shadish, 1996; Shadish & Ragsdale, 1996). These biases are often large and plausibly due to selection biases such as the self-selection of more distressed clients into psychotherapy treatment conditions (Shadish et al., 2000) or of patients with a poorer prognosis into controls in medical experiments (Kunz & Oxman, 1998). These biases are especially prevalent in quasi-experiments that use poor quality control groups and have higher attrition (Heinsman & Shadish, 1996; Shadish & Ragsdale, 1996). So, if the answers obtained from randomized experiments are more credible than those from quasi-experiments on theoretical grounds and are more accurate empirically, then the arguments for randomized experiments are even stronger whenever a high degree of uncertainty reduction is required about a descriptive causal claim.

Because all quasi-experiments are not equal in their ability to reduce uncertainty about cause, we want to draw attention again to a common but unfortunate practice in many social sciences—to say that a quasi-experiment is being done in order to provide justification that the resulting inference will be valid. Then a quasi-experimental design is described that is so deficient in the desirable structural features noted previously, which promote better inference, that it is probably not worth doing. Indeed, over the years we have repeatedly noted the term *quasi-experiment* being used to justify designs that fell into the class that Campbell and Stanley (1963) labeled as uninterpretable and that Cook and Campbell (1979) labeled as generally uninterpretable. These are the simplest forms of the designs discussed in Chapters 4 and 5. Quasi-experiments cannot be an alternative to randomized experiments when the latter are feasible, and poor quasi-experiments can never be a substitute for stronger quasi-experiments when the latter are also feasible. Just as Gueron (1999) has reminded us about randomized experiments, good quasi-experiments have to be fought for, too. They are rarely handed out as though on a silver plate.

Statistical Controls

In this book, we have advocated that statistical adjustments for group nonequivalence are best used *after* design controls have already been used to the maximum in order to reduce nonequivalence to a minimum. So we are not opponents of statistical adjustment techniques such as those advocated by the statisticians and econometricians described in the appendix to Chapter 5. Rather, we want to use them as the last resort. The position we do not like is the assumption that statistical controls are so well developed that they can be used to obtain confident results in nonexperimental and weak quasi-experimental contexts. As we saw in Chapter 5, research in the past 2 decades has not much supported the notion that a control group can be constructed through matching from some national or state registry when the treatment group comes from a more circumscribed and local setting. Nor has research much supported the use of statistical adjustments in longitudinal national surveys in which individuals with different experiences are explicitly contrasted in order to estimate the effects of this experience difference. Undermatching is a chronic problem here, as are consequences of unreliability in the selection variables, not to speak of specification errors due to incomplete knowledge of the selection process. In particular, endogeneity problems are a real concern. We are heartened that more recent work on statistical adjustments seems to be moving toward the position we represent, with greater emphasis being placed on internal controls, on stable matching within such internal controls, on the desirability of seeking cohort controls through the use of siblings, on the use of pretests collected on the same measures as the posttest, on the utility of such pretest measures collected at several different times, and on the desirability of studying interventions that are clearly exogenous shocks to some ongoing system. We are also heartened by the progress being made in the statistical domain because it includes progress on design considerations, as well as on analysis per se (e.g., Rosenbaum, 1999a). We are agnostic at this time as to the virtues of the propensity score and instrumental variable approaches that predominate in discussions of statistical adjustment. Time will tell how well they pan out relative to the results from randomized experiments. We have surely not heard the last word on this topic.

CONCLUSION

We cannot point to one new development that has revolutionized field experimentation in the past few decades, yet we have seen a very large number of incremental improvements. As a whole, these improvements allow us to create far better field experiments than we could do 40 years ago when Campbell and Stanley (1963) first wrote. In this sense, we are very optimistic about the future. We believe that we will continue to see steady, incremental growth in our knowledge about how to do better field experiments. The cost of this growth, however, is that field experimentation has become a more specialized topic, both in terms of knowledge development and of the opportunity to put that knowledge into practice in the conduct of field experiments. As a result, nonspecialists who wish to do a field experiment may greatly benefit by consulting with those with the expertise, especially for large experiments, for experiments in which implementation problems may be high, or for cases in which methodological vulnerabilities will greatly reduce credibility. The same is true, of course, for many other methods. Case-study methods, for example, have become highly enough developed that most researchers would do an amateurish job of using them without specialized training or supervised practice. Such Balkanization of methodology is, perhaps, inevitable, though none the less regrettable. We can ease the regret somewhat by recognizing that with specialization may come faster progress in solving the problems of field experimentation.