Experimental and Quasi-Experimental Designs for Generalized Causal Inference

Contents

Experimental and Quasi-Experimental Designs for Generalized Causal Inference

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.

2

Statistical Conclusion Validity and Internal Validity

Val·id (văl′ĭd): [French *valide,* from Old French from Latin *validus,* strong, from *valre,* to be strong; see *wal-* in Indo-European Roots.] adj. 1. Well grounded; just: *a valid objection.* 2. Producing the desired results; efficacious: *valid methods.* 3. Having legal force; effective or binding: *a valid title.* 4. Logic. a. Containing premises from which the conclusion may logically be derived: *a valid argument.* b. Correctly inferred or deduced from a premise: *a valid conclusion.*

Ty·pol·o·gy (tī-pŏl′ə-jē): n., pl. ty·pol·o·gies. 1. The study or systematic classification of types that have characteristics or traits in common. 2. A theory or doctrine of types, as in scriptural studies.

Threat (thrĕt): [Middle English from Old English *thrat,* oppression; see *treud-* in Indo-European Roots.] n. 1. An expression of an intention to inflict pain, injury, evil, or punishment. 2. An indication of impending danger or harm. 3. One that is regarded as a possible danger; a menace.

A FAMOUS STUDY in early psychology concerned a horse named Clever Hans who seemed to solve mathematics problems, tapping out the answer with his hoof. A psychologist, Oskar Pfungst, critically examined the performance of Clever Hans and concluded that he was really responding to subtly conveyed researcher expectations about when to start and stop tapping (Pfungst, 1911). In short, Pfungst questioned the **validity** of the initial inference that Clever Hans solved math problems. All science and all experiments rely on making such inferences validly. This chapter presents the theory of validity that underlies the approach to generalized causal inference taken in this book. It begins by discussing the meaning ascribed to validity both in theory and in social science practice and then describes a validity typology that introduces the twin ideas of validity types and **threats to validity.** This chapter and the next provide an extended description of these types of validity and of threats that go with them.

VALIDITY

We use the term *validity* to refer to the approximate truth of an inference.[1] When we say something is valid, we make a judgment about the extent to which relevant evidence supports that inference as being true or correct. Usually, that evidence comes from both empirical findings and the consistency of these findings with other sources of knowledge, including past findings and theories. Assessing validity always entails fallible human judgments. We can never be certain that all of the many inferences drawn from a single experiment are true or even that other inferences have been conclusively falsified. That is why validity judgments are not absolute; various degrees of validity can be invoked. As a result, when we use terms such as *valid* or *invalid* or *true* or *false* in this book, they should always be understood as prefaced by "approximately" or "tentatively." For reasons of style we usually omit these modifiers.

Validity is a property of inferences. It is *not* a property of designs or methods, for the same design may contribute to more or less valid inferences under different circumstances. For example, using a randomized experiment does not guarantee that one will make a valid inference about the existence of a descriptive causal relationship. After all, differential attrition may vitiate randomization, power may be too low to detect the effect, improper statistics may be used to analyze the data, and sampling error might even lead us to misestimate the direction of the effect. So it is wrong to say that a randomized experiment is internally valid or has internal validity—although we may occasionally speak that way for convenience. The same criticism is, of course, true of *any* other method used in science, from the case study to the random sample survey. No method guarantees the validity of an inference.

As a corollary, because methods do not have a one-to-one correspondence with any one type of validity, the use of a method may affect more than one type of validity simultaneously. The best-known example is the decision to use a randomized experiment, which often helps internal validity but hampers external validity. But there are many other examples, such as the case in which diversifying participants improves external validity but decreases statistical conclusion validity or in which treatment standardization clarifies construct validity of the treatment but reduces external validity to practical settings in which such standardi-

[1] We might use the terms *knowledge claim* or *proposition* in place of *inference* here, the former being observable embodiments of inferences. There are differences implied by each of these terms, but we treat them interchangeably for present purposes.

zation is not common. This is the nature of practical action: our design choices have multiple consequences for validity, not always ones we anticipate. Put differently, every solution to a problem tends to create new problems. This is not unique to science but is true of human action generally (Sarason, 1978).

Still, in our theory, validity is intimately tied to the idea of truth. In philosophy, three theories of truth have traditionally dominated (Schmitt, 1995). **Correspondence theory** says that a knowledge claim is true if it corresponds to the world—for example, the claim that it is raining is true if we look out and see rain falling. **Coherence theory** says that a claim is true if it belongs to a coherent set of claims—for example, the claim that smoking marijuana causes cancer is true if it is consistent with what we know about the results of marijuana smoking on animal systems much like human ones, if cancer has resulted from other forms of smoking, if the causes of cancer include some elements that are known to follow from marijuana smoking, and if the physiological mechanisms that relate smoking tobacco to cancer are also activated by smoking marijuana. **Pragmatism** says that a claim is true if it is useful to believe that claim—for example, we say that "electrons exist" if inferring such entities brings meaning or predictability into a set of observations that are otherwise more difficult to understand. To play this role, electrons need not actually exist; rather, postulating them provides intellectual order, and following the practices associated with them in theory provides practical utility.[2]

Unfortunately, philosophers do not agree on which of these three theories of truth is correct and have successfully criticized aspects of all of them. Fortunately, we need not endorse any one of these as the single *correct definition* of truth in order to endorse each of them as part of a complete description of the *practical strategies* scientists actually use to construct, revise, and justify knowledge claims. Correspondence theory is apparent in the nearly universal scientific concern of gathering data to assess how well knowledge claims match the world. Scientists also judge how well a given knowledge claim coheres with other knowledge claims built into accepted current theories and past findings. Thus Eisenhart and Howe (1992) suggest that a case study's conclusions must cohere with existing theoretical, substantive, and practical knowledge in order to be valid, and scientists traditionally view with skepticism any knowledge claim that flatly contradicts what is already thought to be well established (Cook et al., 1979). On the pragmatic front, Latour (1987) claims that what comes to be accepted as true in science is what scientists can convince others to use, for it is by use that knowledge claims gain currency and that practical accomplishments accrue. This view is apparent in

[2] A fourth theory, **deflationism** (sometimes called the redundancy or minimalist theory of truth; Horowich, 1990), denies that truth involves correspondence to the world, coherence, or usefulness. Instead, it postulates that the word *truth* is a trivial linguistic device "for assenting to propositions expressed by sentences too numerous, lengthy, or cumbersome to utter" (Schmitt, 1995, p. 128). For example, the claim that "Euclidean geometry is true" is said instead of repeating one's assent to all the axioms of Euclidean geometry, and the claim means no more than that list of axioms.

Mishler's (1990) assertion that qualitative methods are validated by "a functional criterion—whether findings are relied upon for further work" (p. 419) and in a recent response to a statistical-philosophical debate that "in the interest of science, performance counts for more than rigid adherence to philosophical principles" (Casella & Schwartz, 2000, p. 427).

Our theory of validity similarly makes some use of each of these approaches to truth—as we believe all practical theories of validity must do. Our theory clearly appeals to the correspondence between empirical evidence and abstract inferences. It is sensitive to the degree to which an inference coheres with relevant theory and findings. And it has a pragmatic emphasis in emphasizing the utility of ruling out the alternative explanations that practicing scientists in a given research area believe could compromise knowledge claims, even though such threats are, in logic, just a subset of all possible alternatives to the claim. Thus a mix of strategies characterizes how we will proceed, reluctantly eschewing a single, royal road to truth, for each of these single roads is compromised. Correspondence theory is compromised because the data to which a claim is compared are themselves theory laden and so cannot provide a theory-free test of that claim (Kuhn, 1962). Coherence theory is vulnerable to the criticism that coherent stories need not bear any exact relationship to the world. After all, effective swindlers' tales are often highly coherent, even though they are, in fact, false in some crucial ways. Finally, pragmatism is vulnerable because many beliefs known to be true by other criteria have little utility—for example, knowledge of the precise temperature of small regions in the interior of some distant star. Because philosophers do not agree among themselves about which theory of truth is best, practicing scientists should not have to choose among them in justifying a viable approach to the validity of inferences about causation and its generalization.

Social and psychological forces also profoundly influence what is accepted as true in science (Bloor, 1997; Latour, 1987; Pinch, 1986; Shapin, 1994). This is illustrated by Galileo's famous tribulations with the Inquisition and by the history of the causes of ulcers that we covered in Chapter 1. But following Shapin's (1994) distinction between an evaluative and a social theory of truth, we

> want to preserve . . . the loose equation between truth, knowledge and the facts of the matter, while defending the practical interest and legitimacy of a more liberal notion of truth, a notion in which there is indeed a socio-historical story to be told about truth. (Shapin, 1994, p. 4)

As Bloor (1997) points out, science is not a zero-sum game whose social and cognitive-evaluative influences detract from each other; instead, they complement each other. Evaluative theories deal with factors influencing what we *should* accept as true and, for the limited realm of causal inferences and their generality, our theory of validity tries to be evaluative in this normative sense. The social theory tells about external factors influencing what we *do* accept as true, including how we come to believe that one thing causes another (Heider, 1944)—so a social theory of truth might be based on insight, on findings from psychology, or on features in the social, political, and economic environment (e.g., Cordray, 1986). Social theory about truth is not a central topic of this book, though we touch on it in several places. However, truth is manifestly a social construction, and it depends on more than evaluative theories of truth such as correspondence, coherence, and pragmatism. But we believe that truth *does* depend on these in part, and it is this part we develop most thoroughly.

A Validity Typology

A little history will place the current typology in context. Campbell (1957) first defined **internal validity** as the question, "did in fact the experimental stimulus make some significant difference in this specific instance?" (p. 297) and **external validity** as the question, "to what populations, settings, and variables can this effect be generalized?" (p. 297).[3] Campbell and Stanley (1963) followed this lead closely. Internal validity referred to inferences about whether "the experimental treatments make a difference in this specific experimental instance" (Campbell & Stanley, 1963, p. 5). External validity asked "to what populations, settings, treatment variables, and measurement variables can this effect be generalized" (Campbell & Stanley, 1963, p. 5).[4]

Cook and Campbell (1979) elaborated this validity typology into four related components: **statistical conclusion validity**, internal validity, construct validity, and external validity. Statistical conclusion validity referred to the appropriate use of statistics to infer whether the presumed independent and dependent variables covary. Internal validity referred to whether their covariation resulted from a causal relationship. Both construct and external validity referred to generalizations—the former from operations to constructs (with particular emphasis on cause and effect constructs) and the latter from the samples of persons,

[3] Campbell (1986) suggests that the distinction was partly motivated by the emphasis in the 1950s on Fisherian randomized experiments, leaving students with the erroneous impression that randomization took care of all threats to validity. He said that the concept of external validity was originated to call attention to those threats that randomization did not reduce and that therefore "backhandedly, threats to internal validity were, initially and implicitly, those for which random assignment did control" (p. 68). Though this cannot be literally true—attrition was among his internal validity threats, but it is not controlled by random assignment—this quote does provide useful insight into the thinking that initiated the distinction.

[4] External validity is sometimes confused with ecological validity. The latter is used in many different ways (e.g., Bronfenbrenner, 1979; Brunswick, 1943, 1956). However, in its original meaning it is not a validity type but a method that calls for research with samples of settings and participants that reflect the ecology of application (although Bronfenbrenner understood it slightly differently; 1979, p. 29). The internal-external validity distinction is also sometimes confused with the laboratory-field distinction. Although the latter distinction did help motivate Campbell's (1957) thinking, the two are logically orthogonal. In principle, the causal inference from a field experiment can have high internal validity, and one can ask whether a finding first identified in the field would generalize to the laboratory setting.

TABLE 2.1 Four Types of Validity

Statistical Conclusion Validity: The validity of inferences about the correlation (covariation) between treatment and outcome.

Internal Validity: The validity of inferences about whether observed covariation between A (the presumed treatment) and B (the presumed outcome) reflects a causal relationship from A to B as those variables were manipulated or measured.

Construct Validity: The validity of inferences about the higher order constructs that represent sampling particulars.

External Validity: The validity of inferences about whether the cause-effect relationship holds over variation in persons, settings, treatment variables, and measurement variables.

settings, and times achieved in a study to and across populations about which questions of generalization might be raised.

In this book, the definitions of statistical conclusion and internal validity remain essentially unchanged from Cook and Campbell (1979), extending the former only to consider the role of effect sizes in experiments. However, we modify construct and external validity to accommodate Cronbach's (1982) points that both kinds of causal generalizations (representations and extrapolations) apply to all elements of a study (units, treatments, observations and settings; see Table 2.1). Hence construct validity is now defined as the degree to which inferences are warranted from the observed persons, settings, and cause and effect operations included in a study to the constructs that these instances might represent. External validity is now defined as the validity of inferences about whether the causal relationship holds over variation in persons, settings, treatment variables, and measurement variables.

In Cook and Campbell (1979), construct validity was mostly limited to inferences about higher order constructs that represent the treatments and observations actually studied;[5] in our current usage, we extend this definition of construct validity to cover persons and settings, as well. In Cook and Campbell (1979), external validity referred only to inferences about how a causal relationship would generalize to and across populations of persons and settings; here we extend their definition of external validity to include treatments and observations, as well. Creating a separate construct validity label only for cause and effect issues was justi-

[5] However, Cook and Campbell (1979) explicitly recognized the possibility of inferences about constructs regarding other study features such as persons and settings: "In the discussion that follows we shall restrict ourselves to the construct validity of presumed causes and effects, since these play an especially crucial role in experiments whose raison d'etre is to test causal propositions. But it should be clearly noted that construct validity concerns are not limited to cause and effect constructs. 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).

fied pragmatically in Cook and Campbell because of the attention it focused on a central issue in causation: how the cause and effect should be characterized theoretically. But this salience was sometimes interpreted to imply that characterizing populations of units and settings is trivial. Because it is not, construct validity should refer to them also. Similarly, we should not limit external generalizations to persons and settings, for it is worth assessing whether a particular cause-and-effect relationship would hold if different variants of the causes or effects were used—those differences are often small variations but can sometimes be substantial. We will provide examples of these inferences in Chapter 3.

Our justification for discussing these four slightly reformulated validity types remains pragmatic, however, based on their correspondence to four major questions that practicing researchers face when interpreting causal studies: (1) How large and reliable is the covariation between the presumed cause and effect? (2) Is the covariation causal, or would the same covariation have been obtained without the treatment? (3) Which general constructs are involved in the persons, settings, treatments, and observations used in the experiment? and (4) How generalizable is the locally embedded causal relationship over varied persons, treatments, observations, and settings? Although these questions are often highly interrelated, it is worth treating them separately because the inferences drawn about them often occur independently and because the reasoning we use to construct each type of inference differs in important ways. In the end, however, readers should always remember that "A validity typology can greatly aid . . . design, but it does not substitute for critical analysis of the particular case or for logic" (Mark, 1986 p. 63).

Threats to Validity

Threats to validity are specific reasons why we can be partly or completely wrong when we make an inference about covariance, about causation, about constructs, or about whether the causal relationship holds over variations in persons, settings, treatments, and outcomes. In this chapter we describe threats to statistical conclusion validity and internal validity; in the following chapter we do the same for construct and external validity. The threats we present to each of the four validity types have been identified through a process that is partly conceptual and partly empirical. In the former case, for example, many of the threats to internal validity are tied to the nature of reasoning about descriptive causal inferences outlined in Chapter 1. In the latter case, Campbell (1957) identified many threats from critical commentary on past experiments, most of those threats being theoretically mundane. The empirically based threats can, should, and do change over time as experience indicates both the need for new threats and the obsolescence of former ones. Thus we add a new threat to the traditional statistical conclusion validity threats. We call it "Inaccurate Effect Size Estimation" in order to reflect the reality that social scientists now emphasize estimating the size of causal effects, in addition to running the usual statistical significance tests. Conversely, although each of the threats we describe do indeed occur in experiments, the likelihood that they will occur varies across contexts. Lists of validity threats are heuristic aids; they are not etched in stone, and they are not universally relevant across all research areas in the social sciences.

These threats serve a valuable function: they help experimenters to anticipate the likely criticisms of inferences from experiments that experience has shown occur frequently, so that the experimenter can try to rule them out.[6] The primary method we advocate for ruling them out is to use design controls that minimize the number and plausibility of those threats that remain by the end of a study. This book is primarily about how to conduct such studies, particularly with the help of design rather than statistical adjustment controls. The latter are highlighted in presentations of causal inference in much of economics, say, but less so in statistics itself, in which the design controls we prefer also tend to be preferred. Random assignment is a salient example of good design control. This book describes the experimental design elements that generally increase the quality of causal inferences by ruling out more alternative interpretations to a causal claim. Chapter 8 shows how and when random assignment to treatment and comparison conditions can enhance causal inference, whereas Chapters 4 through 7 show what design controls can be used when random assignment is not possible or has broken down.

However, many threats to validity cannot be ruled out by design controls, either because the logic of design control does not apply (e.g., with some threats to construct validity such as inadequate construct explication) or because practical constraints prevent available controls from being used. In these cases, the appropriate method is to identify and explore the role and influence of the threat in the study. In doing this, three questions are critical: (1) How would the threat apply in this case? (2) Is there evidence that the threat is plausible rather than just possible? (3) Does the threat operate in the same direction as the observed effect, so that it could partially or totally explain the observed findings? For example, suppose a critic claims that history (other events occurring at the same time as treatment that could have caused the same outcome) is a threat to the internal validity of a quasi-experiment you have conducted on the effects of the federal Women, Infants, and Children (WIC) Program to improve pregnancy outcome among eligible low-income women compared with a control group of ineligible women. First, we need to know how "history" applies in this case, for example, whether other social programs are available and whether women who are eligible for WIC are also eligible for these other programs. A little thought shows that the food stamps program might be such a threat. Second, we need to know if there is evi-

[6] We agree with Reichardt (2000) that it would be better to speak of "taking account of threats to validity" than to say "ruling out threats to validity," for the latter implies a finality that can rarely be achieved in either theory or practice. Talking about "ruling out" threats implies an all-or-none quality in which threats either do or do not apply; but in many cases threats are a matter of degree rather than being absolute. However, we also agree with Reichardt that the term "ruling out" has such a strong foothold in this literature that we can continue to use the term for stylistic reasons.

dence—or at least a reasonable expectation given past findings or background knowledge—that more women who are eligible for WIC are getting food stamps than women who are ineligible for WIC. If not, then although this particular history threat is possible, it may not be plausible. In this case, background knowledge suggests that the threat is plausible because both the WIC Program and the food stamps program use similar eligibility criteria. Third, if the threat is plausible, we need to know if the effects of food stamps on pregnancy outcome would be similar to the effects of the WIC Program. If not, then this history threat could not explain the observed effect, and so it does not threaten it. In this case, the threat would be real, for food stamps could lead to better nutrition, which could also improve pregnancy outcome. Throughout this book, we will emphasize these three crucial questions about threats in the examples we use.

The previous example concerns a threat identified by a critic after a study was done. Given the difficulties all researchers have in criticizing their own work, such post hoc criticisms are probably the most common source of identified threats to studies. However, it is better if the experimenter can anticipate such a threat before the study has begun. If he or she can anticipate it but cannot institute design controls to prevent the threat, the best alternative is to measure the threat directly to see if it actually operated in a given study and, if so, to conduct statistical analyses to examine whether it can plausibly account for the obtained cause-effect relationship. We heartily endorse the direct assessment of possible threats, whether done using quantitative or qualitative observations. It will sometimes reveal that a specific threat that might have operated did not in fact do so or that the threat operated in a way opposite to the observed effect and so could not account for the effect (e.g., Gastwirth, Krieger, & Rosenbaum, 1994). However, we are cautious about using such direct measures of threats in statistical analyses that claim to rule out the threat. The technical reasons for this caution are explained in subsequent chapters, but they have to do with the need for full knowledge of how a threat operates and for perfect measurement of the threat. The frequent absence of such knowledge is why we usually prefer design over statistical control, though in practice most studies will achieve a mix of both. We want to tilt the mix more in the design direction, and to this end this book features a large variety of practical design elements that, in different real-world circumstances, can aid in causal inference while limiting the need for statistical adjustment.

In doing all this, the experimenter must remember that ruling out threats to validity is a falsificationist enterprise, subject to all the criticisms of falsificationism that we outlined in Chapter 1. For example, ruling out plausible threats to validity in experiments depends on knowing the relevant threats. However, this knowledge depends on the quality of the relevant methodological and substantive theories available and on the extent of background information available from experience with the topic on hand. It also depends on the existence of a widely accepted theory of "plausibility," so that we know which of the many possible threats are plausible in this particular context. Without such a theory, most researchers rely on their own all-too-fallible judgment (Mark, 1986; Rindskopf, 2000). And it depends on measuring the threats in unbiased ways that do not include the theories, wishes, expectations, hopes, or category systems of the observers. So the process of ruling out threats to validity exemplifies the fallible falsificationism that we described in Chapter 1.

STATISTICAL CONCLUSION VALIDITY

Statistical conclusion validity concerns two related statistical inferences that affect the covariation component of causal inferences:[7] (1) whether the presumed cause and effect covary and (2) how strongly they covary. For the first of these inferences, we can incorrectly conclude that cause and effect covary when they do not (a **Type I error**) or incorrectly conclude that they do not covary when they do (a **Type II error**). For the second inference, we can overestimate or underestimate the magnitude of covariation, as well as the degree of confidence that magnitude estimate warrants. In this chapter, we restrict ourselves to classical statistical conceptions of covariation and its magnitude, even though qualitative analyses of covariation are both plausible and important.[8] We begin with a brief description of the nature of covariation statistics and then discuss the specific threats to those inferences.

Reporting Results of Statistical Tests of Covariation

The most widely used way of addressing whether cause and effect covary is **null hypothesis significance testing** (NHST). An example is that of an experimenter who computes a t-test on treatment and comparison group means at posttest, with the usual **null hypothesis** being that the difference between the population means from which these samples were drawn is zero.[9] A test of this hypothesis is typically accompanied by a statement of the probability that a difference of the size obtained (or larger) would have occurred by chance (e.g., $p = .036$) in a popula-

[7] We use covariation and correlation interchangeably, the latter being a standardized version of the former. The distinction can be important for other purposes, however, such as when we model explanatory processes in Chapter 12.

[8] Qualitative researchers often make inferences about covariation based on their observations, as when they talk about how one thing seems related to another. We can think about threats to the validity of those inferences, too. Psychological theory about biases in covariation judgments might have much to offer to this program (e.g., Crocker, 1981; Faust, 1984), as with the "illusory correlation" bias in clinical psychology (Chapman & Chapman, 1969). But we do not know all or most of these threats to qualitative inferences about covariation; and some we know have been seriously criticized (e.g., Gigerenzer, 1996) because they seem to operate mostly with individuals' first reactions. Outlining threats to qualitative covariation inferences is a task best left to qualitative researchers whose contextual familiarity with such work makes them better suited to the task than we are.

[9] Cohen (1994) suggests calling this zero-difference hypothesis the "nil" hypothesis to emphasize that the hypothesis of zero difference is not the only possible hypothesis to be nullified. We discuss other possible null hypotheses shortly. Traditionally, the opposite of the null hypothesis has been called the **alternative hypothesis**, for example, that the difference between group means is not zero.

tion in which no between-group difference exists. Following a tradition first suggested by Fisher (1926, p. 504), it has unfortunately become customary to describe this result dichotomously—as statistically significant if $p < .05$ or as nonsignificant otherwise. Because the implication of nonsignificance is that a cause and effect do not covary—a conclusion that can be wrong and have serious consequences—threats to statistical conclusion validity are partly about why a researcher might be wrong in claiming not to find a significant effect using NHST.

However, problems with this kind of NHST have been known for decades (Meehl, 1967, 1978; Rozeboom, 1960), and the debate has intensified recently (Abelson, 1997; Cohen, 1994; Estes, 1997; Frick, 1996; Harlow, Mulaik, & Steiger, 1997; Harris, 1997; Hunter, 1977; Nickerson, 2000; Scarr, 1997; Schmidt, 1996; Shrout, 1997; Thompson, 1993). Some critics even want to replace NHST totally with other options (Hunter, 1997; Schmidt, 1996). The arguments are beyond the scope of this text, but primarily they reduce to two: (1) scientists routinely misunderstand NHST, believing that *p* describes the chances that the null hypothesis is true or that the experiment would replicate (Greenwald, Gonzalez, Harris, & Guthrie, 1996); and (2) NHST tells us little about the size of an effect. Indeed, some scientists wrongly think that nonsignificance implies a zero effect when it is more often true that such **effect sizes** are different from zero (e.g., Lipsey & Wilson, 1993).

This is why most parties to the debate about statistical significance tests prefer reporting results as effect sizes bounded by confidence intervals, and even the advocates of NHST believe it should play a less prominent role in describing experimental results. But few parties to the debate believe that NHST should be banned outright (e.g., Howard, Maxwell, & Fleming, 2000; Kirk, 1996). It can still be useful for understanding the role that chance may play in our findings (Krantz, 1999; Nickerson, 2000). So we prefer to see results reported first as effect size estimates accompanied by 95% confidence intervals, followed by the exact probability level of a Type I error from a NHST.[10] This is feasible for any focused comparison between two conditions (e.g., treatment versus control); Rosenthal and Rubin (1994) suggest methods for contrasts involving more than two conditions.

The effect size and 95% confidence interval contain all the information provided by traditional NHST but focus attention on the magnitude of covariation and the precision of the effect size estimate; for example, "the 95% confidence interval of 6 ± 2 shows more precision than the 95% confidence interval of 6 ± 5"

[10] The American Psychological Association's Task Force on Statistical Inference concluded, *"It is hard to imagine a situation in which a dichotomous accept-reject decision is better than reporting an actual* p *value or, better still, a confidence interval. . . . Always provide some effect-size estimate when reporting a* p *value. . . . Interval estimates should be given for any effect sizes involving principal outcomes"* (Wilkinson and the Task Force on Statistical Inference, 1999, p. 599). Cohen (1994) suggests reporting "confidence curves" (Birnbaum, 1961) from which can be read all confidence intervals from 50% to 100% so that just one confidence interval need not be chosen; a computer program for generating these curves is available (Borenstein, Cohen, & Rothstein, in press).

(Frick, 1996, p. 383). Confidence intervals also help to distinguish between situations of low statistical power, and hence wide confidence intervals, and situations with precise but small effect sizes—situations that have quite different implications. Reporting the preceding statistics would also decrease current dependence on speciously precise point estimates, replacing them with more realistic ranges that better reflect uncertainty even though they may complicate public communication. Thus the statement "the average increase in income was $1,000 per year" would be complemented by "the likely outcome is an average increase ranging between $400 and $1600 per year."

In the classic interpretation, exact Type I probability levels tell us the probability that the results that were observed in the experiment could have been obtained by chance from a population in which the null hypothesis is true (Cohen, 1994). In this sense, NHST provides some information that the results could have arisen due to chance—perhaps not the most interesting hypothesis but one about which it has become customary to provide the reader with information. A more interesting interpretation (Frick, 1996; Harris, 1997; Tukey, 1991) is that the probability level tells us about the confidence we can have in deciding among three claims: (1) the sign of the effect in the population is positive (Treatment A did better than Treatment B); (2) the sign is negative (Treatment B did better than Treatment A); or (3) the sign is uncertain. The smaller the *p* value, the less likely it is that our conclusion about the sign of the population effect is wrong; and if $p >$ .05 (or, equivalently, if the confidence interval contains zero), then our conclusion about the sign of the effect is too close to call.

In any case, whatever interpretation of the *p* value from NHST one prefers, all this discourages the overly simplistic conclusion that either "there is an effect" or "there is no effect." We believe that traditional NHST will play an increasingly small role in social science, though no new approach will be perfect.[11] As Abelson recently said:

> Whatever else is done about null-hypothesis tests, let us stop viewing statistical analysis as a sanctification process. We are awash in a sea of uncertainty, caused by a flood tide of sampling and measurement errors, and there are no objective procedures that avoid human judgment and guarantee correct interpretations of results. (1997, p. 13)

[11] An alternative (more accurately, a complement) to both NHST and reporting effect sizes with confidence intervals is the use of Bayesian statistics (Etzioni & Kadane, 1995; Howard et al., 2000). Rather than simply accept or reject the null hypothesis, Bayesian approaches use the results from a study to update existing knowledge on an ongoing basis, either prospectively by specifying expectations about study outcomes before the study begins (called prior probabilities) or retrospectively by adding results from an experiment to an existing corpus of experiments that has already been analyzed with Bayesian methods to update results. The latter is very close to random effects meta-analytic procedures (Hedges, 1998) that we cover in Chapter 13. Until recently, Bayesian statistics have been used sparingly, partly because of ambiguity about how prior probabilities should be obtained and partly because Bayesian methods were computationally intensive with few computer programs to implement them. The latter objection is rapidly dissipating as more powerful computers and acceptable programs are developed (Thomas, Spiegelhalter, & Gilks, 1992), and the former is beginning to be addressed in useful ways (Howard et al., 2000). We expect to see increasing use of Bayesian statistics in the next few decades, and as their use becomes more frequent, we will undoubtedly find threats to the validity of them that we do not yet include here.

TABLE 2.2 Threats to Statistical Conclusion Validity: Reasons Why Inferences About Covariation Between Two Variables May Be Incorrect

1. *Low Statistical Power:* An insufficiently powered experiment may incorrectly conclude that the relationship between treatment and outcome is not significant.
2. *Violated Assumptions of Statistical Tests:* Violations of statistical test assumptions can lead to either overestimating or underestimating the size and significance of an effect.
3. *Fishing and the Error Rate Problem:* Repeated tests for significant relationships, if uncorrected for the number of tests, can artifactually inflate statistical significance.
4. *Unreliability of Measures:* Measurement error weakens the relationship between two variables and strengthens or weakens the relationships among three or more variables.
5. *Restriction of Range:* Reduced range on a variable usually weakens the relationship between it and another variable.
6. *Unreliability of Treatment Implementation:* If a treatment that is intended to be implemented in a standardized manner is implemented only partially for some respondents, effects may be underestimated compared with full implementation.
7. *Extraneous Variance in the Experimental Setting:* Some features of an experimental setting may inflate error, making detection of an effect more difficult.
8. *Heterogeneity of Units:* Increased variability on the outcome variable within conditions increases error variance, making detection of a relationship more difficult.
9. *Inaccurate Effect Size Estimation:* Some statistics systematically overestimate or underestimate the size of an effect.

Threats to Statistical Conclusion Validity

Table 2.2 presents a list of threats to statistical conclusion validity, that is, reasons why researchers may be wrong in drawing valid inferences about the existence and size of covariation between two variables.

Low Statistical Power

Power refers to the ability of a test to detect relationships that exist in the population, and it is conventionally defined as the probability that a statistical test will reject the null hypothesis when it is false (Cohen, 1988; Lipsey, 1990; Maxwell & Delaney, 1990). When a study has low power, effect size estimates will be less precise (have wider confidence intervals), and traditional NHST may incorrectly conclude that cause and effect do not covary. Simple computer programs can calculate power if we know or can estimate the sample size, the Type I and Type II error rates, and the effect sizes (Borenstein & Cohen, 1988; Dennis, Lennox, & Foss, 1997; Hintze, 1996; Thomas & Krebs, 1997). In social science practice, Type I error rates are usually set at $\alpha = .05$, although good reasons often exist to deviate from this

(Makuch & Simon, 1978)—for example, when testing a new drug for harmful side effects, a higher Type I error rate might be fitting (e.g., $\alpha = .20$). It is also common to set the Type II error rate (β) at .20, and power is then $1 - \beta = .80$. The target effect size is often inferred from what is judged to be a practically important or theoretically meaningful effect (Cohen, 1996; Lipsey, 1990), and the standard deviation needed to compute effect sizes is usually taken from past research or pilot work. If the power is too low for detecting an effect of the specified size, steps can be taken to increase power. Given the central importance of power in practical experimental design, Table 2.3 summarizes the many factors that affect power that will be discussed in this book and provides comments about such matters as their feasibility, application, exceptions to their use, and disadvantages.

TABLE 2.3 Methods to Increase Power

Method	Comments
Use matching, stratifying, blocking	1. Be sure the variable used for matching, stratifying, or blocking is correlated with outcome (Maxwell, 1993), or use a variable on which subanalyses are planned. 2. If the number of units is small, power can decrease when matching is used (Gail et al., 1996).
Measure and correct for covariates	1. Measure covariates correlated with outcome and adjust for them in statistical analysis (Maxwell, 1993). 2. Consider cost and power tradeoffs between adding covariates and increasing sample size (Allison, 1995; Allison et al., 1997). 3. Choose covariates that are nonredundant with other covariates (McClelland, 2000). 4. Use covariance to analyze variables used for blocking, matching, or stratifying.
Use larger sample sizes	1. If the number of treatment participants is fixed, increase the number of control participants. 2. If the budget is fixed and treatment is more expensive than control, compute optimal distribution of resources for power (Orr, 1999). 3. With a fixed total sample size in which aggregates are assigned to conditions, increase the number of aggregates and decrease the number of units within aggregates.
Use equal cell sample sizes	1. Unequal cell splits do not affect power greatly until they exceed 2:1 splits (Pocock, 1983). 2. For some effects, unequal sample size splits can be more powerful (McClelland, 1997).

TABLE 2.3 Continued

Method	Comments
Improve measurement	1. Increase measurement reliability or use latent variable modeling. 2. Eliminate unnecessary restriction of range (e.g., rarely dichotomize continuous variables). 3. Allocate more resources to posttest than to pretest measurement (Maxwell, 1994). 4. Add additional waves of measurement (Maxwell, 1998). 5. Avoid floor or ceiling effects.
Increase the strength of treatment	1. Increase dose differential between conditions. 2. Reduce diffusion over conditions. 3. Ensure reliable treatment delivery, receipt, and adherence.
Increase the variability of treatment	1. Extend the range of levels of treatment that are tested (McClelland, 2000). 2. In some cases, oversample from extreme levels of treatment (McClelland, 1997).
Use a within-participants design	1. Less feasible outside laboratory settings. 2. Subject to fatigue, practice, contamination effects.
Use homogenous participants selected to be responsive to treatment	1. Can compromise generalizability.
Reduce random setting irrelevancies	1. Can compromise some kinds of generalizability.
Ensure that powerful statistical tests are used and their assumptions are met	1. Failure to meet test assumptions sometimes increases power (e.g., treating dependent units as independent), so you must know the relationship between assumption and power. 2. Transforming data to meet normality assumptions can improve power even though it may not affect Type I error rates much (McClelland, 2000). 3. Consider alternative statistical methods (e.g., Wilcox, 1996).

To judge from reviews, low power occurs frequently in experiments. For instance, Kazdin and Bass (1989) found that most psychotherapy outcome studies comparing two treatments had very low power (see also Freiman, Chalmers, Smith, & Kuebler, 1978; Lipsey, 1990; Sedlmeier & Gigerenzer, 1989). So low power is a major cause of false null conclusions in individual studies. But when effects are small, it is frequently impossible to increase power sufficiently using the methods in Table 2.3. This is one reason why the synthesis of many studies (see Chapter 13) is now so routinely advocated as a path to more powerful tests of small effects.

Violated Assumptions of the Test Statistics

Inferences about covariation may be inaccurate if the assumptions of a statistical test are violated. Some assumptions can be violated with relative impunity. For instance, a two-tailed *t*-test is reasonably robust to violations of normality if group sample sizes are large and about equal and only Type I error is at issue (Judd, McClelland, & Culhane, 1995; but for Type II error, see Wilcox, 1995). However, violations of other assumptions are more serious. For instance, inferences about covariation may be inaccurate if observations are not independent—for example, children in the same classroom may be more related to each other than randomly selected children are; patients in the same physician's practice or workers in the same workplace may be more similar to each other than randomly selected individuals are.[12] This threat occurs often and violates the assumption of independently distributed errors. It can introduce severe bias to the estimation of standard errors, the exact effects of which depend on the design and the kind of dependence (Judd et al., 1995). In the most common case of units nested within aggregates (e.g., children in some schools get one treatment and children in other schools get the comparison condition), the bias is to increase the Type I error rate dramatically so that researchers will conclude that there is a "significant" treatment difference far more often than they should. Fortunately, recent years have seen the development of relevant statistical remedies and accompanying computer programs (Bryk & Raudenbush, 1992; Bryk, Raudenbush, & Congdon, 1996; DeLeeuw & Kreft, 1986; Goldstein, 1987).

Fishing and the Error Rate Problem

An inference about covariation may be inaccurate if it results from fishing through the data set to find a "significant" effect under NHST or to pursue leads suggested by the data themselves, and this inaccuracy can also occur when multiple investigators reanalyze the same data set (Denton, 1985). When the Type I error rate for a single test is $\alpha = .05$, the error rate for a set of tests is quite different and increases with more tests. If three tests are done with a nominal $\alpha = .05$, then the actual alpha (or the probability of making a Type I error over all three tests) is .143; with twenty tests it is .642; and with fifty tests it is .923 (Maxwell & Delaney, 1990). Especially if only a subset of results are reported (e.g., only the significant ones), the research conclusions can be misleading.

[12] Violations of this assumption used to be called the "unit of analysis" problem; we discuss this problem in far more detail in Chapter 8.

The simplest corrective procedure is the very conservative Bonferroni correction, which divides the overall target Type I error rate for a set (e.g., $\alpha = .05$) by the number of tests in the set and then uses the resulting Bonferroni-corrected α in all individual tests. This ensures that the error rate over all tests will not exceed the nominal $\alpha = .05$. Other corrections include the use of conservative multiple comparison follow-up tests in analysis of variance (ANOVA) or the use of a multivariate ANOVA if multiple dependent variables are tested (Maxwell & Delaney, 1990). Some critics of NHST discourage such corrections, arguing that we already tend to overlook small effects and that conservative corrections make this even more likely. They argue that reporting effect sizes, confidence intervals, and exact *p* values shifts the emphasis from "significant-nonsignificant" decisions toward confidence about the likely sign and size of the effect. Other critics argue that if results are reported for all statistical tests, then readers can assess for themselves the chances of spuriously "significant" results by inspection (Greenwald et al., 1996). However, it is unlikely that complete reporting will occur because of limited publication space and the tendency of authors to limit reports to the subset of results that tell an interesting story. So in most applications, fishing will still lead researchers to have more confidence in associations between variables than they should.

Unreliability of Measures

A conclusion about covariation may be inaccurate if either variable is measured unreliably (Nunnally & Bernstein, 1994). **Unreliability** always attenuates bivariate relationships. When relationships involve three or more variables, the effects of unreliability are less predictable. Maxwell and Delaney (1990) showed that unreliability of a covariate in an analysis of covariance can produce significant treatment effects when the true effect is zero or produce zero effects in the presence of true effects. Similarly, Rogosa (1980) showed that the effects of unreliability in certain correlational designs depended on the pattern of relationships among variables and the differential **reliability** of the variables, so that nearly any effect or null effect could be found no matter what the true effect might be. Special reliability issues arise in longitudinal studies that assess rates of change, acceleration, or other features of development (Willett, 1988). So reliability should be assessed and reported for each measure. Remedies for unreliability include increasing the number of measurements (e.g., using more items or more raters), improving the quality of measures (e.g., better items, better training of raters), using special kinds of growth curve analyses (Willett, 1988), and using techniques like **latent variable** modeling of several observed measures to parcel out true score from error variance (Bentler, 1995).

Restriction of Range

Sometimes variables are restricted to a narrow range; for instance, in experiments two highly similar treatments might be compared or the outcome may have only two values or be subject to floor or ceiling effects. This restriction also lowers power and attenuates bivariate relations. Restriction on the independent variable can be decreased by, for example, studying distinctly different treatment doses or even full-dose treatment versus no treatment. This is especially valuable early in a research program when it is important to test whether large effects can be found under circumstances most favorable to its emergence. Dependent variables are restricted by **floor effects** when all respondents cluster near the lowest possible score, as when most respondents score normally on a scale measuring pathological levels of depression, and by **ceiling effects** when all respondents cluster near the highest score, as when a study is limited to the most talented students. When continuous measures are dichotomized (or trichotomized, etc.), range is again restricted, as when a researcher uses the median weight of a sample to create high- and low-weight groups. In general, such splits should be avoided.[13] Pilot testing measures and selection procedures help detect range restriction, and item response theory analyses can help to correct the problem if a suitable calibration sample is available (Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980).

Unreliability of Treatment Implementation

Conclusions about covariation will be affected if treatment is implemented inconsistently from site to site or from person to person within sites (Boruch & Gomez, 1977; Cook, Habib, Philips, Settersten, Shagle, & Degirmencioglu, 1999; Lipsey, 1990). This threat is pervasive in field experiments, in which controlling the treatment is less feasible than in the laboratory. Lack of standardized implementation is commonly thought to decrease an effect size, requiring more attention to other design features that increase power, such as sample size. However, some authors note that variable implementation may reflect a tailoring of the intervention to the recipient in order to increase its effects (Scott & Sechrest, 1989; Sechrest, West, Phillips, Redner, & Yeaton, 1979; Yeaton & Sechrest, 1981). Further, lack of standardization is also not a problem if the desired inference is to a treatment that is supposed to differ widely across units. Indeed, a lack of standardization is intrinsic to some real-world interventions. Thus, in studies of the Comprehensive Child Development Program (Goodson, Layzer, St. Pierre, Bernstein & Lopez, 2000) and Early Head Start (Kisker & Love, 1999), poor parents of young children were provided with different packages of services depending on the varying nature of their needs. Thus some combinations of job training, formal education, parent training, counseling, or emergency housing might be needed, creating a very heterogeneous treatment across the families studied. In all these cases, however, efforts should be made to measure the components of the treatment package and to explore how the various components are related to changes

[13] Counterintuitively, Maxwell and Delaney (1990) showed that dichotomizing two continuous independent variables to create a factorial ANOVA design can sometimes increase power (by increasing Type I error rate).

in outcomes. Because this issue is so important, in Chapters 10 and 12 we discuss methods for improving, measuring, and analyzing treatment implementation that help reduce this threat.

Extraneous Variance in the Experimental Setting

Conclusions about covariation can be inaccurate if features of an experimental setting artifactually inflate error. Examples include distracting noises, fluctuations in temperature due to faulty heating/cooling systems, or frequent fiscal or administrative changes that distract practitioners. A solution is to control these factors or to choose experimental procedures that force respondents' attention on the treatment or that lower environmental salience. But in many field settings, these suggestions are impossible to implement fully. This situation entails the need to measure those sources of extraneous variance that cannot otherwise be reduced, using them later in the statistical analysis. Early qualitative monitoring of the experiment will help suggest what these variables might be.

Heterogeneity of Units (Respondents)

The more the units in a study are heterogeneous within conditions on an outcome variable, the greater will be the standard deviations on that variable (and on any others correlated with it). Other things being equal, this heterogeneity will obscure systematic covariation between treatment and outcome. Error also increases when researchers fail to specify respondent characteristics that interact with a cause-and-effect relationship, as in the case of some forms of depression that respond better to a psychotherapeutic treatment than others. Unless they are specifically measured and modeled, these **interactions** will be part of error, obscuring systematic covariation. A solution is to sample respondents who are homogenous on characteristics correlated with major outcomes. However, such selection may reduce external validity and can cause restriction of range if it is not carefully monitored. Sometimes a better solution is to measure relevant respondent characteristics and use them for **blocking** or as covariates. Also, within-participant designs can be used in which the extent of the advantage depends on the size of the correlation between pre- and posttest scores.

Inaccurate Effect Size Estimation

Covariance estimates can be inaccurate when the size of the effect is measured poorly. For example, when outliers cause a distribution to depart even a little from normality, this can dramatically decrease effect sizes (Wilcox, 1995). Wilcox (in press) suggests alternative effect size estimation methods for such data (along with Minitab computer programs), though they may not fit well with standard statistical techniques. Also, analyzing dichotomous outcomes with effect size measures designed for continuous variables (i.e., the correlation coefficient or standardized mean difference statistic) will usually underestimate effect size; **odds ratios** are usually a better choice (Fleiss, 1981, p. 60). Effect size estimates are also implicit in common statistical tests. For example, if an ordinary *t*-test is computed on a dichotomous outcome, it implicitly uses the standardized mean difference statistic and will have lower power. As researchers increasingly report effect size and confidence intervals, more causes of inaccurate effect size estimation will undoubtedly be found.

The Problem of Accepting the Null Hypothesis

Although we hope to discourage researchers from describing a failure to reject the null hypothesis as "no effect," there are circumstances in which they must consider such a conclusion. One circumstance is that in which the true hypothesis of interest is a no-effect one, for example, that a new treatment does as well as the accepted standard, that a feared side effect does not occur (Makuch & Simon, 1978), that extrasensory perception experiments have no effect (Rosenthal, 1986), or that the result of a first coin toss has no relationship to the result of a second if the coin is fair (Frick, 1996). Another is that in which a series of experiments yields results that are all "too close to call," leading the experimenter to wonder whether to continue to investigate the treatment. A third is the case in which the analyst wants to show that groups do not differ on various threats to validity, as when group equivalence on pretests is examined for **selection bias** (Yeaton & Sechrest, 1986). Each of these situations requires testing whether the obtained covariation can be reliably distinguished from zero. However, it is very hard to prove that covariation is exactly zero because power theory suggests that, even when an effect is very small, larger sample sizes, more reliable measures, better treatment implementation, or more accurate statistics might distinguish it from zero. From this emerges the maxim that we cannot prove the null hypothesis (Frick, 1995).

To cope with situations such as these, the first thing to do is to maximize power so as to avoid "too close to call" conclusions. Table 2.3 listed many ways in which this can be done, though each differs in its feasibility for any given study and some may not be desirable if they conflict with other goals of the experiment. Nonetheless, examining studies against these power criteria will often reveal whether it is desirable and practical to conduct new experiments with more powerful designs.

A second thing to do is to pay particular attention to identifying the size of an effect worth pursuing, for example, the maximum acceptable harm or the smallest effect that makes a practical difference (Fowler, 1985; Prentice & Miller, 1992; Rouanet, 1996; Serlin & Lapsley, 1993). Aschenfelter's (1978) study of the effects of manpower training programs on subsequent earnings estimated that an increase in earnings of $200 would be adequate for declaring the program a success. He could then use power analysis to ensure a sufficient sample to detect this effect. However, specifying such an effect size is a political act, because a reference point is then created against which an innovation can be evaluated. Thus, even if an innovation has a partial effect, it may not be given credit for this if the promised effect size has not been achieved. Hence managers of educational programs learn to assert, "We want to increase achievement" rather than stating, "We want to increase achievement by two years for every year of teaching." However, even when such factors mitigate against specifying a minimally acceptable effect size, presenting the absolute magnitude of an obtained treatment effect allows readers to infer for themselves whether an effect is so small as to be practically unimportant or whether a nonsignificant effect is so large as to merit further research with more powerful analyses.

Third, if the hypothesis concerns the equivalency of two treatments, biostatisticians have developed equivalency testing techniques that could be used in place of traditional NHST. These methods test whether an observed effect falls into a range that the researcher judges to be equivalent for practical purposes, even if the difference between treatments is not zero (Erbland, Deupree, & Niewoehner, 1999; Rogers, Howard, & Vessey, 1993; Westlake, 1988).

A fourth option is to use quasi-experimental analyses to see if larger effects can be located under some important conditions—for example, subtypes of participants who respond to treatment more strongly or naturally occurring dosage variations that are larger than average in an experiment. Caution is required in interpreting such results because of the risk of capitalizing on chance and because individuals will often have self-selected themselves into treatments differentially. Nonetheless, if sophisticated quasi-experimental analyses fail to show minimally interesting covariation between treatment and outcome measures, then the analyst's confidence that the effect is too small to pursue increases.

INTERNAL VALIDITY

We use the term *internal validity* to refer to inferences about whether observed covariation between A and B reflects a causal relationship from A to B in the form in which the variables were manipulated or measured. To support such an inference, the researcher must show that A preceded B in time, that A covaries with B (already covered under statistical conclusion validity) and that no other explanations for the relationship are plausible. The first problem is easily solved in experiments because they force the manipulation of A to come before the measurement of B. However, causal order is a real problem in nonexperimental research, especially in cross-sectional work.

Although the term *internal validity* has been widely adopted in the social sciences, some of its uses are not faithful to the concept as first described by Campbell (1957). Internal validity was not about reproducibility (Cronbach, 1982), nor inferences to the target population (Kleinbaum, Kupper, & Morgenstern, 1982), nor measurement validity (Menard, 1991), nor whether researchers measure what they think they measure (Goetz & LeCompte, 1984). To reduce such misunderstandings, Campbell (1986) proposed relabeling internal validity as **local molar causal validity**, a relabeling that is instructive to explicate even though it is so cumbersome that we will not use it, sticking with the older but more memorable and widely accepted term (internal validity).

The word *causal* in *local molar causal validity* emphasizes that internal validity is about causal inferences, not about other types of inference that social scientists make. The word *local* emphasizes that causal conclusions are limited to the context of the particular treatments, outcomes, times, settings, and persons studied. The word *molar* recognizes that experiments test treatments that are a complex package consisting of many components, all of which are tested as a whole within the treatment condition. Psychotherapy, for example, consists of different verbal interventions used at different times for different purposes. There are also nonverbal cues both common to human interactions and specific to provider-client relationships. Then there is the professional placebo provided by prominently displayed graduate degrees and office suites modeled on medical precedents, financial arrangements for reimbursing therapists privately or through insurance, and the physical condition of the psychotherapy room (to name just some parts of the package). A client assigned to psychotherapy is assigned to all parts of this molar package and others, not just to the part that the researcher may intend to test. Thus the causal inference from an experiment is about the effects of being assigned to the whole molar package. Of course, experiments can and should break down such molar packages into molecular parts that can be tested individually or against each other. But even those molecular parts are packages consisting of many components. Understood as local molar causal validity, internal validity is about whether a complex and inevitably multivariate treatment package caused a difference in some variable-as-it-was-measured within the particular setting, time frames, and kinds of units that were sampled in a study.

Threats to Internal Validity

In what may be the most widely accepted analysis of causation in philosophy, Mackie (1974) stated: "Typically, we infer from an effect to a cause (inus condition) by eliminating other possible causes" (p. 67). Threats to internal validity are those other possible causes—reasons to think that the relationship between A and B is not causal, that it could have occurred even in the absence of the treatment, and that it could have led to the same outcomes that were observed for the treatment. We present these threats (Table 2.4) separately even though they are not totally independent. Enough experience with this list has accumulated to suggest that it applies to any descriptive molar causal inference, whether generated from experiments, correlational studies, observational studies, or case studies. After all, validity is not the property of a method; it is a characteristic of knowledge claims (Shadish, 1995b)—in this case, claims about causal knowledge.

TABLE 2.4 Threats to Internal Validity: Reasons Why Inferences That the Relationship Between Two Variables Is Causal May Be Incorrect

1. *Ambiguous Temporal Precedence:* Lack of clarity about which variable occurred first may yield confusion about which variable is the cause and which is the effect.
2. *Selection:* Systematic differences over conditions in respondent characteristics that could also cause the observed effect.
3. *History:* Events occurring concurrently with treatment could cause the observed effect.
4. *Maturation:* Naturally occurring changes over time could be confused with a treatment effect.
5. *Regression:* When units are selected for their extreme scores, they will often have less extreme scores on other variables, an occurrence that can be confused with a treatment effect.
6. *Attrition:* Loss of respondents to treatment or to measurement can produce artifactual effects if that loss is systematically correlated with conditions.
7. *Testing:* Exposure to a test can affect scores on subsequent exposures to that test, an occurrence that can be confused with a treatment effect.
8. *Instrumentation:* The nature of a measure may change over time or conditions in a way that could be confused with a treatment effect.
9. *Additive and Interactive Effects of Threats to Internal Validity:* The impact of a threat can be added to that of another threat or may depend on the level of another threat.

Ambiguous Temporal Precedence

Cause must precede effect, but sometimes it is unclear whether A precedes B or vice versa, especially in correlational studies. But even in correlational studies, one direction of causal influence is sometimes implausible (e.g., an increase in heating fuel consumption does not cause a decrease in outside temperature). Also, some correlational studies are longitudinal and involve data collection at more than one time. This permits analyzing as potential causes only those variables that occurred before their possible effects. However, the fact that A occurs before B does not justify claiming that A causes B; other conditions of causation must also be met.

Some causation is bidirectional (reciprocal), as with the criminal behavior that causes incarceration that causes criminal behavior that causes incarceration, or with high levels of school performance that generate self-efficacy in a student that generates even higher school performance. Most of this book is about testing unidirectional causation in experiments. Experiments were created for this purpose precisely because it is known which factor was deliberately manipulated before another was measured. However, separate experiments can test first whether A causes B and second whether B causes A. So experiments are not irrelevant to causal reciprocation, though simple experiments are. Other methods for testing reciprocal causation are discussed briefly in Chapter 12.

Selection

Sometimes, at the start of an experiment, the average person receiving one experimental condition already differs from the average person receiving another condition. This difference might account for any result after the experiment ends that the analysts might want to attribute to treatment. Suppose that a compensatory education program is given to children whose parents volunteer them and that the comparison condition includes only children who were not so volunteered. The volunteering parents might also read to their children more, have more books at home, or otherwise differ from nonvolunteers in ways that might affect their child's achievement. So children in the compensatory education program might do better even without the program.[14] When properly implemented, random assignment definitionally eliminates such selection bias because randomly formed groups differ only by chance. Of course, faulty randomization can introduce selection bias, as can a successfully implemented randomized experiment in which subsequent attrition differs by treatment group. Selection is presumed to be pervasive in quasi-experiments, given that they are defined as using the structural attributes of experiments but without random assignment. The key feature of selection bias is a confounding of treatment effects with population differences. Much of this book will be concerned with selection, both when individuals select themselves into treatments and when administrators place them in different treatments.

History

History refers to all events that occur between the beginning of the treatment and the posttest that could have produced the observed outcome in the absence of that treatment. We discussed an example of a history threat earlier in this chapter regarding the evaluation of programs to improve pregnancy outcome in which receipt of food stamps was that threat (Shadish & Reis, 1984). In laboratory research, history is controlled by isolating respondents from outside events (e.g., in a quiet laboratory) or by choosing dependent variables that could rarely be affected by the world outside (e.g., learning nonsense syllables). However, experimental isolation is rarely available in field research—we cannot and would not stop pregnant mothers from receiving food stamps and other external events that might improve pregnancy outcomes. Even in field research, though, the plausibility of history can be reduced; for example, by selecting groups from the same general location and by ensuring that the schedule for testing is the same in both groups (i.e., that one group is not being tested at a very different time than another, such as testing all control participants prior to testing treatment participants; Murray, 1998).

[14] Though it is common to discuss selection in two-group designs, such selection biases can also occur in single-group designs when the composition of the group changes over time.

Maturation

Participants in research projects experience many natural changes that would occur even in the absence of treatment, such as growing older, hungrier, wiser, stronger, or more experienced. Those changes threaten internal validity if they could have produced the outcome attributed to the treatment. For example, one problem in studying the effects of compensatory education programs such as Head Start is that normal cognitive development ensures that children improve their cognitive performance over time, a major goal of Head Start. Even in short studies such processes are a problem; for example, fatigue can occur quickly in a verbal learning experiment and cause a performance decrement. At the community level or higher, maturation includes secular trends (Rossi & Freeman, 1989), changes that are occurring over time in a community that may affect the outcome. For example, if the local economy is growing, employment levels may rise even if a program to increase employment has no specific effect. Maturation threats can often be reduced by ensuring that all groups are roughly of the same age so that their individual maturational status is about the same and by ensuring that they are from the same location so that local secular trends are not differentially affecting them (Murray, 1998).

Regression Artifacts

Sometimes respondents are selected to receive a treatment because their scores were high (or low) on some measure. This often happens in quasi-experiments in which treatments are made available either to those with special merits (who are often then compared with people with lesser merits) or to those with special needs (who are then compared with those with lesser needs). When such extreme scorers are selected, there will be a tendency for them to score less extremely on other measures, including a retest on the original measure (Campbell & Kenny, 1999). For example, the person who scores highest on the first test in a class is not likely to score highest on the second test; and people who come to psychotherapy when they are extremely distressed are likely to be less distressed on subsequent occasions, even if psychotherapy had no effect. This phenomenon is often called regression to the mean (Campbell & Stanley, 1963; Furby, 1973; Lord, 1963; Galton, 1886, called it regression toward mediocrity) and is easily mistaken for a treatment effect. The prototypical case is selection of people to receive a treatment because they have extreme pretest scores, in which case those scores will tend to be less extreme at posttest. However, regression also occurs "backward" in time. That is, when units are selected because of extreme posttest scores, their pretest scores will tend to be less extreme; and it occurs on simultaneous measures, as when extreme observations on one posttest entail less extreme observations on a correlated posttest. As a general rule, readers should explore the plausibility of this threat in detail *whenever respondents are selected (or select themselves) because they had scores that were higher or lower than average.*

Regression to the mean occurs because measures are not perfectly correlated with each other (Campbell & Kenny, 1999; Nesselroade, Stigler, & Baltes, 1980; Rogosa, 1988). **Random measurement** error is part of the explanation for this imperfect correlation. Test theory assumes that every measure has a true score component reflecting a true ability, such as depression or capacity to work, *plus* a random error component that is normally and randomly distributed around the mean of the measure. On any given occasion, high scores will tend to have more positive random error pushing them up, whereas low scores will tend to have more negative random error pulling them down. On the same measure at a later time, or on other measures at the same time, the random error is less likely to be so extreme, so the observed score (the same true score plus less extreme random error) will be less extreme. So using more reliable measures can help reduce regression.

However, it will not prevent it, because most variables are imperfectly correlated with each other by their very nature and would be imperfectly correlated even if they were perfectly measured (Campbell & Kenny, 1999). For instance, both height and weight are nearly perfectly measured; yet in any given sample, the tallest person is not always the heaviest, nor is the lightest person always the shortest. This, too, is regression to the mean. Even when the same variable is measured perfectly at two different times, a real set of forces can cause an extreme score at one of those times; but these forces are unlikely to be maintained over time. For example, an adult's weight is usually measured with very little error. However, adults who first attend a weight-control clinic are likely to have done so because their weight surged after an eating binge on a long business trip exacerbated by marital stress; their weight will regress to a lower level as those causal factors dissipate even if the weight-control treatment has no effect. But notice that in all these cases, the key clue to the possibility of regression artifacts is always present—selection based on an extreme score, whether it be the person who scored highest on the first test, the person who comes to psychotherapy when most distressed, the tallest person, or the person whose weight just reached a new high.

What should researchers do to detect or reduce statistical regression? If selection of extreme scorers is a necessary part of the question, the best solution is to create a large group of extreme scorers from within which random assignment to different treatments then occurs. This unconfounds regression and receipt of treatment so that regression occurs equally for each group. By contrast, the worst situation occurs when participants are selected into a group based on extreme scores on some unreliable variable and that group is then compared with a group selected differently. This builds in the very strong likelihood of group differences in regression that can masquerade as a treatment effect (Campbell & Erlebacher, 1970). In such cases, because regression is most apparent when inspecting standardized rather than raw scores, diagnostic tests for regression (e.g., Galton squeeze diagrams; Campbell & Kenny, 1999) should be done on standardized scores. Researchers should also increase the reliability of any selection measure by increasing the number of items on it, by averaging it over several time points, or by using a multivariate function of several variables instead of a single variable for selection. Another procedure is working with three or more time points; for example, making the selection into groups based on the Time 1 measure, implementing the treatment after the Time 2 measure, and then examining change between Time 2 and Time 3 rather than between Time 1 and Time 3 (Nesselroade et al., 1980).

Regression does not require quantitative analysis to occur. Psychologists have identified it as an illusion that occurs in ordinary cognition (Fischhoff, 1975; Gilovich, 1991; G. Smith, 1997; Tversky & Kahneman, 1974). Psychotherapists have long noted that clients come to therapy when they are more distressed than usual and tend to improve over time even without therapy. They call this spontaneous remission rather than statistical regression, but it is the same phenomenon. The clients' measured progress is partly a movement back toward their stable individual mean as the temporary shock that led them to therapy (a death, a job loss, a shift in the marriage) grows less acute. Similar examples are those alcoholics who appear for treatment when they have "hit bottom" or those schools and businesses that call for outside professional help when things are suddenly worse. Many business consultants earn their living by capitalizing on regression, avoiding institutions that are stably bad but manage to stay in business and concentrating instead on those that have recently had a downturn for reasons that are unclear.

Attrition

Attrition (sometimes called experimental mortality) refers to the fact that participants in an experiment sometimes fail to complete the outcome measures. If different kinds of people remain to be measured in one condition versus another, then such differences could produce posttest outcome differences even in the absence of treatment. Thus, in a randomized experiment comparing family therapy with discussion groups for treatment of drug addicts, addicts with the worst prognosis tend to drop out of the discussion group more often than out of family therapy. If the results of the experiment suggest that family therapy does less well than discussion groups, this might just reflect differential attrition, by which the worst addicts stayed in family therapy (Stanton & Shadish, 1997). Similarly, in a longitudinal study of a study-skills treatment, the group of college seniors that eventually graduates is only a subset of the incoming freshmen and might be systematically different from the initial population, perhaps because they are more persistent or more affluent or higher achieving. This then raises the question: Was the final grade point average of the senior class higher than that of the freshman class because of the effects of a treatment or because those who dropped out had lower scores initially? Attrition is therefore a special subset of selection bias occurring after the treatment is in place. But unlike selection, differential attrition is not controlled by random assignment to conditions.

Testing

Sometimes taking a test once will influence scores when the test is taken again. Practice, familiarity, or other forms of reactivity are the relevant mechanisms and could be mistaken for treatment effects. For example, weighing someone may cause the person to try to lose weight when they otherwise might not have done so, or taking a vocabulary pretest may cause someone to look up a novel word and so perform better at posttest. On the other hand, many measures are not reactive in this way. For example, a person could not change his or her height (see Webb, Campbell, Schwartz, & Sechrest, 1966, and Webb, Campbell, Schwartz, Sechrest, & Grove, 1981, for other examples). Techniques such as item response theory sometimes help reduce testing effects by allowing use of different tests that are calibrated to yield equivalent ability estimates (Lord, 1980). Sometimes testing effects can be assessed using a Solomon Four Group Design (Braver & Braver, 1988; Dukes, Ullman, & Stein, 1995; Solomon, 1949), in which some units receive a pretest and others do not, to see if the pretest causes different treatment effects. Empirical research suggests that testing effects are sufficiently prevalent to be of concern (Willson & Putnam, 1982), although less so in designs in which the interval between tests is quite large (Menard, 1991).

Instrumentation

A change in a measuring instrument can occur over time even in the absence of treatment, mimicking a treatment effect. For example, the spring on a bar press might become weaker and easier to push over time, artifactually increasing reaction times; the component stocks of the Dow Jones Industrial Average might have changed so that the new index reflects technology more than the old one; and human observers may become more experienced between pretest and posttest and so report more accurate scores at later time points. Instrumentation problems are especially prevalent in studies of child development, in which the measurement unit or scale may not have constant meaning over the age range of interest (Shonkoff & Phillips, 2000). Instrumentation differs from testing because the former involves a change in the instrument, the latter a change in the participant. Instrumentation changes are particularly important in longitudinal designs, in which the way measures are taken may change over time (see Figure 6.7 in Chapter 6) or in which the meaning of a variable may change over life stages (Menard, 1991).[15] Methods for investigating these changes are discussed by Cunningham (1991) and Horn (1991). Researchers should avoid switching instruments during a study; but

[15] Epidemiologists sometimes call instrumentation changes surveillance bias.

if switches are required, the researcher should retain both the old and new items (if feasible) to calibrate one against the other (Murray, 1998).

Additive and Interactive Effects of Threats to Internal Validity

Validity threats need not operate singly. Several can operate simultaneously. If they do, the net bias depends on the direction and magnitude of each individual bias plus whether they combine additively or multiplicatively (interactively). In the real world of social science practice, it is difficult to estimate the size of such net bias. We presume that inaccurate causal inferences are more likely the more numerous and powerful are the simultaneously operating validity threats and the more homogeneous their direction. For example, a **selection-maturation** additive effect may result when nonequivalent experimental groups formed at the start of treatment are also maturing at different rates over time. An illustration might be that higher achieving students are more likely to be given National Merit Scholarships and also likely to be improving their academic skills at a more rapid rate. Both initial high achievement and more rapid achievement growth serve to doubly inflate the perceived effects of National Merit Scholarships. Similarly, a **selection-history** additive effect may result if nonequivalent groups also come from different settings and each group experiences a unique local history. A **selection-instrumentation** additive effect might occur if nonequivalent groups have different means on a test with unequal intervals along its distribution, as would occur if there is a ceiling or floor effect for one group but not for another.[16]

Estimating Internal Validity in Randomized Experiments and Quasi-Experiments

Random assignment eliminates selection bias definitionally, leaving a role only to chance differences. It also reduces the plausibility of other threats to internal validity. Because groups are randomly formed, any initial group differences in maturational rates, in the experience of simultaneous historical events, and in regression artifacts ought to be due to chance. And so long as the researcher administers the same tests in each condition, pretesting effects and instrumentation changes should be experienced equally over conditions within the limits of chance. So random assignment and treating groups equivalently in such matters as pretesting and instrumentation improve internal validity.

[16] Cook and Campbell (1979) previously called these interactive effects; but they are more accurately described as additive. Interactions among threats are also possible, including higher order interactions, but describing examples of these accurately can be more complex than needed here.

Given random assignment, inferential problems about causation arise in only two situations. In the first, attrition from the experiment is differential by treatment group, in which case the outcome differences between groups might be due to differential attrition rather than to treatment. Techniques have recently been advanced for dealing with this problem (e.g., Angrist et al., 1996a), and we review them in Chapter 10. In the second circumstance, testing must be different in each group, as when the expense or response burden of testing on participants is so high that the experimenter decides to administer pretests only to a treatment group that is more likely to be cooperative if they are getting, say, a desirable treatment. Experimenters should monitor a study to detect any differential attrition early and to try to correct it before it goes too far, and they should strive to make testing procedures as similar as possible across various groups.

With quasi-experiments, the causal situation is more murky, because differences between groups will be more systematic than random. So the investigator must rely on other options to reduce internal validity threats. The main option is to modify a study's design features. For example, regression artifacts can be reduced by not selecting treatment units on the basis of extreme and partially unreliable scores, provided that this restriction does not trivialize the research question. History can be made less plausible to the extent that experimental isolation is feasible. Attrition can be reduced using many methods to be detailed in Chapter 10. But it is not always feasible to implement these design features, and doing so sometimes subtly changes the nature of the research question. This is why the omnibus character of random assignment is so desirable.

Another option is to make all the threats explicit and then try to rule them out one by one. Identifying each threat is always context specific; for example, what may count as history in one context (e.g., the introduction of *Sesame Street* during an experiment on compensatory education in the 1970s) may not count as a threat at all in another context (e.g., watching *Sesame Street* is an implausible means of reducing unwanted pregnancies). Once identified, the presence of a threat can be assessed either quantitatively by measurement or qualitatively by observation or interview. In both cases, the presumed effect of the threat can then be compared with the outcome to see if the *direction* of the threat's bias is the same as that of the observed outcome. If so, the threat may be plausible, as with the example of the introduction of *Sesame Street* helping to improve reading rather than a contemporary education program helping to improve it. If not, the threat may still be implausible, as in the discovery that the healthiest mothers are more likely to drop out of a treatment but that the treatment group still performs better than the controls. When the threat is measured quantitatively, it might be addressed by state-of-the-art statistical adjustments, though this is problematic because those adjustments have not always proven very accurate and because it is not easy to be confident that all the context-specific threats to internal validity have been identified. Thus the task of individually assessing the plausibility of internal validity threats is definitely more laborious and less certain than relying on experimental design, randomization in particular but also the many design elements we introduce throughout this book.

THE RELATIONSHIP BETWEEN INTERNAL VALIDITY AND STATISTICAL CONCLUSION VALIDITY

These two validity types are closely related. Both are primarily concerned with study operations (rather than with the constructs those operations reflect) and with the relationship between treatment and outcome. Statistical conclusion validity is concerned with errors in assessing statistical covariation, whereas internal validity is concerned with causal-reasoning errors. Even when all the statistical analyses in a study are impeccable, errors of causal reasoning may still lead to the wrong causal conclusion. So statistical covariation does not prove causation. Conversely, when a study is properly implemented as a randomized experiment, statistical errors can still occur and lead to incorrect judgments about statistical significance and misestimated effect sizes. Thus, in quantitative experiments, internal validity depends substantially on statistical conclusion validity.

However, experiments need not be quantitative in how either the intervention or the outcome are conceived and measured (Lewin, 1935; Lieberson, 1985; Mishler, 1990), and some scholars have even argued that the statistical analysis of quantitative data is detrimental (e.g., Skinner, 1961). Moreover, examples of qualitative experiments abound in the physical sciences (e.g., Drake, 1981; Hacking, 1983; Naylor, 1989; Schaffer, 1989), and there are even some in the social sciences. For instance, Sherif's famous Robber's Cave Experiment (Sherif, Harvey, White, Hood, & Sherif, 1961) was mostly qualitative. In that study, boys at a summer camp were divided into two groups of eleven each. Within-group cohesion was fostered for each group separately, and then intergroup conflict was introduced. Finally, conflict was reduced using an intervention to facilitate equal status cooperation and contact while working on common goals. Much of the data in this experiment was qualitative, including the highly cited effects on the reduction of intergroup conflict. In such cases, internal validity no longer depends directly on statistical conclusion validity, though clearly an assessment that treatment covaried with the effect is still necessary, albeit a qualitative assessment.

Indeed, given such logic, Campbell (1975) recanted his previous rejection (Campbell & Stanley, 1963) of using case studies to investigate causal inferences because the reasoning of causal inference *is* qualitative and because all the logical requirements for inferring cause apply as much to qualitative as to quantitative work. Scriven (1976) has made a similar argument. Although each makes clear that causal inferences from case studies are likely to be valid only under limited circumstances (e.g., when isolation of the cause from other confounds is feasible), neither believes that causation requires quantitatively scaled treatments or outcomes. We agree.

3

Construct Validity and External Validity

Re·la·tion·ship (rĭ-lā´shən-shĭp´): n. 1. The condition or fact of being related; connection or association. 2. Connection by blood or marriage; kinship.

Trade·off or **Trade-off** (trād´ôf´, -ŏf´): n. An exchange of one thing in return for another, especially relinquishment of one benefit or advantage for another regarded as more desirable: *"a fundamental trade-off between capitalist prosperity and economic security" (David A. Stockman).*

Pri·or·i·ty (prī-ôr´ĭ-tē, -ŏr´-): [Middle English *priorite,* from Old French from Medieval Latin *pririts,* from Latin *prior,* first; see prior.] n., pl. pri·or·i·ties. 1. Precedence, especially established by order of importance or urgency. 2. a. An established right to precedence. b. An authoritative rating that establishes such precedence. 3. A preceding or coming earlier in time. 4. Something afforded or deserving prior attention.

IN THIS chapter, we continue the consideration of validity by discussing both construct and external validity, including threats to each of them. We then end with a more general discussion of relationships, tradeoffs, and priorities among validity types.

CONSTRUCT VALIDITY

A recent report by the National Academy of Sciences on research in early childhood development succinctly captured the problems of construct validity:

> In measuring human height (or weight or lung capacity, for example), there is little disagreement about the meaning of the construct being measured, or about the units of measurement (e.g., centimeters, grams, cubic centimeters). . . . Measuring growth in psychological domains (e.g., vocabulary, quantitative reasoning, verbal memory, hand–eye coordination, self-regulation) is more problematic. Disagreement is more

> likely to arise about the definition of the constructs to be assessed. This occurs, in part, because there are often no natural units of measurement (i.e., nothing comparable to the use of inches when measuring height). (Shonkoff & Phillips, 2000, pp. 82–83)

Here we see the twin problems of construct validity—understanding constructs and assessing them. In this chapter, we elaborate on how these problems occur in characterizing and measuring the persons, settings, treatments, and outcomes used in an experiment.

Scientists do empirical studies with specific instances of units, treatments, observations, and settings; but these instances are often of interest only because they can be defended as measures of general constructs. Construct validity involves making inferences from the sampling particulars of a study to the higher-order constructs they represent. Regarding the persons studied, for example, an economist may be interested in the construct of unemployed, disadvantaged workers; but the sample of persons actually studied may be those who have had family income below the poverty level for 6 months before the experiment begins or who participate in government welfare or food stamp programs. The economist intends the match between construct and operations to be close, but sometimes discrepancies occur—in one study, some highly skilled workers who only recently lost their jobs met the preceding criteria and so were included in the study, despite not really being disadvantaged in the intended sense (Heckman, Ichimura, & Todd, 1997). Similar examples apply to the treatments, outcomes, and settings studied. Psychotherapists are rarely concerned only with answers to the 21 items on the Beck Depression Inventory; rather, they want to know if their clients are depressed. When agricultural economists study farming methods in the foothills of the Atlas Mountains in Morocco, they are frequently interested in arid agriculture in poor countries. When physicians study 5-year mortality rates among cancer patients, they are interested in the more general concept of survival.

As these examples show, research cannot be done without using constructs. As Albert Einstein once said, "Thinking without the positing of categories and concepts in general would be as impossible as breathing in a vacuum" (Einstein, 1949, pp. 673–674). Construct validity is important for three other reasons, as well. First, constructs are the central means we have for connecting the operations used in an experiment to pertinent theory and to the language communities that will use the results to inform practical action. To the extent that experiments contain construct errors, they risk misleading both theory and practice. Second, construct labels often carry social, political, and economic implications (Hopson, 2000). They shape perceptions, frame debates, and elicit support and criticism. Consider, for example, the radical disagreements that stakeholders have about the label of a "hostile work environment" in sexual or racial harassment litigation, disagreements about what that construct means, how it should be measured, and whether it applies in any given setting. Third, the creation and defense of basic constructs is a fundamental task of all science. Examples from the physical sciences include "the development of the periodic table of elements, the identification of the composition of water, the laying out of different genera and species of plants and animals, and the discovery of the structure of genes" (Mark, 2000, p. 150)—though such taxonomic work is considerably more difficult in the social sciences, for reasons which we now discuss.

Why Construct Inferences Are a Problem

The naming of things is a key problem in all science, for names reflect category memberships that themselves have implications about relationships to other concepts, theories, and uses. This is true even for seemingly simple labeling problems. For example, a recent newspaper article reported a debate among astronomers over what to call 18 newly discovered celestial objects ("Scientists Quibble," 2000). The Spanish astronomers who discovered the bodies called them planets, a choice immediately criticized by some other astronomers: "I think this is probably an inappropriate use of the 'p' word," said one of them. At issue was the lack of a match between some characteristics of the 18 objects (they are drifting freely through space and are only about 5 million years old) and some characteristics that are prototypical of planets (they orbit a star and require tens of millions of years to form). Critics said these objects were more reasonably called brown dwarfs, objects that are too massive to be planets but not massive enough to sustain the thermonuclear processes in a star. Brown dwarfs would drift freely and be young, like these 18 objects. The Spanish astronomers responded that these objects are too small to be brown dwarfs and are so cool that they could not be that young. All this is more than just a quibble: If these objects really are planets, then current theories of how planets form by condensing around a star are wrong! And this is a simple case, for the category of planets is so broadly defined that, as the article pointed out, "Gassy monsters like Jupiter are in, and so are icy little spitwads like Pluto." Construct validity is a much more difficult problem in the field experiments that are the topic of this book.

Construct validity is fostered by (1) starting with a clear explication of the person, setting, treatment, and outcome constructs of interest; (2) carefully selecting instances that match those constructs; (3) assessing the match between instances and constructs to see if any slippage between the two occurred; and (4) revising construct descriptions accordingly. In this chapter, we primarily deal with construct explication and some prototypical ways in which researchers tend to pick instances that fail to represent those constructs well. However, throughout this book, we discuss methods that bear on construct validity. Chapter 9, for example, devotes a section to ensuring that enough of the intended participants exist to be recruited into an experiment and randomized to conditions; and Chapter 10 devotes a section to ensuring that the intended treatment is well conceptualized, induced, and assessed.

There is a considerable literature in philosophy and the social sciences about the problems of construct explication (Lakoff, 1985; Medin, 1989; Rosch, 1978; Smith & Medin, 1981; Zadeh, 1987). In what is probably the most common theory, each construct has multiple features, some of which are more central than others and so are called prototypical. To take a simple example, the prototypical features of a tree are that it is a tall, woody plant with a distinct main stem or trunk that lives for at least 3 years (a perennial). However, each of these attributes is associated with some degree of fuzziness in application. For example, their height and distinct trunk distinguish trees from shrubs, which tend to be shorter and have multiple stems. But some trees have more than one main trunk, and others are shorter than some tall shrubs, such as rhododendrons. No attributes are foundational. Rather, we use a **pattern-matching** logic to decide whether a given instance sufficiently matches the prototypical features to warrant using the category label, especially given alternative category labels that could be used.

But these are only surface similarities. Scientists are often more concerned with deep similarities, prototypical features of particular scientific importance that may be visually peripheral to the layperson. To the layperson, for example, the difference between deciduous (leaf-shedding) and coniferous (evergreen) trees is visually salient; but scientists prefer to classify trees as angiosperms (flowering trees in which the seed is encased in a protective ovary) or gymnosperms (trees that do not bear flowers and whose seeds lie exposed in structures such as cones). Scientists value this discrimination because it clarifies the processes by which trees reproduce, more crucial to understanding forestation and survival than is the lay distinction between deciduous and coniferous. It is thus difficult to decide which features of a thing are more peripheral or more prototypical, but practicing researchers always make this decision, either explicitly or implicitly, when selecting participants, settings, measures, and treatment manipulations.

This difficulty arises in part because deciding which features are prototypical depends on the context in which the construct is to be used. For example, it is not that scientists are right and laypersons wrong about how they classify trees. To a layperson who is considering buying a house on a large lot with many trees, the fact that the trees are deciduous means that substantial annual fall leaf cleanup expenses will be incurred. Medin (1989) gives a similar example, asking what label should be applied to the category that comprises children, money, photo albums, and pets. These are not items we normally see as sharing prototypical construct features, but in one context they do—when deciding what to rescue from a fire.

Deciding which features are prototypical also depends on the particular language community doing the choosing. Consider the provocative title of Lakoff's (1985) book *Women, Fire, and Dangerous Things*. Most of us would rarely think of women, fire, and dangerous things as belonging to the same category. The title provokes us to think of what these things have in common: Are women fiery and dangerous? Are both women and fires dangerous? It provokes us partly because we do not have a natural category that would incorporate all these elements. In the language community of natural scientists, fire might belong to a category having to do with oxidation processes, but women are not in that category. In the language community of ancient philosophy, fire might belong to a category of basic elements along with air, water, and earth, but dangerous things are not among those elements. But in the Australian aboriginal language called Dyirbal, women, fire, and dangerous things are all part of one category.[1]

All these difficulties in deciding which features are prototypical are exacerbated in the social sciences. In part, this is because so many important constructs are still being discovered and developed, so that strong consensus about prototypical construct features is as much the exception as the rule. In the face of only weak consensus, slippage between instance and construct is even greater than otherwise. And in part, it is because of the abstract nature of the entities with which social scientists typically work, such as violence, incentive, decision, plan, and intention. This renders largely irrelevant a theory of categorization that is widely used in some areas—the theory of natural kinds. This theory postulates that nature cuts things at the joint, and so we evolve names and shared understandings for the entities separated by joints. Thus we have separate words for a tree's trunk and its branches, but no word for the bottom left section of a tree. Likewise, we have words for a twig and leaf, but no word for the entity formed by the bottom half of a twig and the attached top third of a leaf. There are many fewer "joints" (or equivalents thereof) in the social sciences—what would they be for intentions or aggression, for instance?

By virtue of all these difficulties, it is never possible to establish a one-to-one relationship between the operations of a study and corresponding constructs. Logical positivists mistakenly assumed that it would be possible to do this, creating a subtheory around the notion of definitional operationalism—that a thing is only what its measure captures, so that each measure is a perfect representation of its own construct. Definitional operationalism failed for many reasons (Bechtel, 1988; H. I. Brown, 1977). Indeed, various kinds of definitional operationalism are threats to construct validity in our list below. Therefore, a theory of constructs must emphasize (1) operationalizing each construct several ways within and across studies; (2) probing the pattern match between the multivariate characteristics of instances and the characteristics of the target construct, and (3) acknowledging legitimate debate about the quality of that match given the socially constructed nature of both operations and constructs. Doing all this is facilitated by detailed description of the studied instances, clear explication of the prototypical elements of the target construct, and valid observation of relationships among the instances, the target construct, and any other pertinent constructs.[2]

[1] The explanation is complex, occupying a score of pages in Lakoff (1985), but a brief summary follows. The Dyirbal language classifies words into four categories (much as the French language classifies nouns as masculine or feminine): (1) Bayi: (human) males; animals; (2) Balan: (human) females; water; fire; fighting; (3) Balam: nonflesh food; (4) Bala: everything not in the other three classes. The moon is thought to be husband to the sun, and so is included in the first category as male; hence the sun is female and in the second category. Fire reflects the same domain of experience as the sun, and so is also in the second category. Because fire is associated with danger, dangerousness in general is also part of the second category.

[2] Cronbach and Meehl (1955) called this set of theoretical relationships a nomological net. We avoid this phrase because its dictionary definition (nomological: the science of physical and logical laws) fosters an image of lawful relationships that is incompatible with field experimentation as we understand it.

Assessment of Sampling Particulars

Good construct explication is essential to construct validity, but it is only half the job. The other half is good assessment of the sampling particulars in a study, so that the researcher can assess the match between those assessments and the constructs. For example, the quibble among astronomers about whether to call 18 newly discovered celestial objects "planets" required *both* a set of prototypical characteristics of planets versus brown dwarfs *and* measurements of the 18 objects on these characteristics—their mass, position, trajectory, radiated heat, and likely age. Because the prototypical characteristics of planets are well-established and accepted among astronomers, critics tend first to target the accuracy of the measurements in such debates, for example, speculating that the Spanish astronomers measured the mass or radiated heat of these objects incorrectly. Consequently, other astronomers try to replicate these measurements, some using the same methods and others using different ones. If the measurements prove correct, then the prototypical characteristics of the construct called planets will have to be changed, or perhaps a new category of celestial object will be invented to account for the anomalous measurements.

Not surprisingly, this attention to measurement was fundamental to the origins of construct validity (Cronbach & Meehl, 1955), which grew out of concern with the quality of psychological tests. The American Psychological Association's (1954) Committee on Psychological Tests had as its job to specify the qualities that should be investigated before a test is published. They concluded that one of those qualities was construct validity. For example, Cronbach and Meehl (1955) said that the question addressed by construct validity is, "What constructs account for variance in test performance?" (p. 282) and also that construct validity involved "how one is to defend a proposed interpretation of a test" (p. 284). The measurement and the construct are two sides of the same construct validity coin.

Of course, Cronbach and Meehl (1955) were not writing about experiments. Rather, their concern was with the practice of psychological testing of such matters as intelligence, personality, educational achievement, or psychological pathology, a practice that blossomed in the aftermath of World War II with the establishment of the profession of clinical psychology. However, those psychological tests were used frequently in experiments, especially as outcome measures in, say, experiments on the effects of educational interventions. So it was only natural that critics of particular experimental findings might question the construct validity of inferences about what is being measured by those outcome measurements. In adding construct validity to the D. T. Campbell and Stanley (1963) validity typology, Cook and Campbell (1979) recognized this usage; and they extended this usage from outcomes to treatments, recognizing that it is just as important to characterize accurately the nature of the treatments that are applied in an experiment. In this book, we extend this usage two steps further to cover persons and settings, as well. Of course, our categorization of experiments as consisting of units (persons), settings, treatments, and outcomes is partly arbitrary, and we could have chosen to treat, say, time as a separate feature of each experiment, as we occasionally have in some of our past work. Such additions would not change the key point. Construct validity involves making inferences from assessments of *any* of the sampling particulars in a study to the higher-order constructs they represent.

Most researchers probably understand and accept the rationale for construct validity of outcome measures. It may help, however, to give examples of construct validity of persons, settings, and treatments. A few of the simplest person constructs that we use require no sophisticated measurement procedures, as when we classify persons as males or females, usually done with no controversy on the basis of either self-report or direct observation. But many other constructs that we use to characterize people are less consensually agreed upon or more controversial. For example, consider the superficially simple problem of racial and ethnic identity for descendants of the indigenous peoples of North America. The labels have changed over the years (Indians, Native Americans, First Peoples), and the ways researchers have measured whether someone merits any one of these labels have varied from self-report (e.g., on basic U.S. Census forms) to formal assessments of the percentage of appropriate ancestry (e.g., by various tribal registries). Similarly, persons labeled schizophrenic will differ considerably depending on whether their diagnosis was measured by criteria of the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders* (1994), by one of the earlier editions of that manual, by the recorded diagnosis in a nursing home chart, or by the Schizophrenia subscale of the Minnesota Multiphasic Personality Inventory-2 (Hathaway & McKinley, 1989). When one then turns to common but very loosely applied terms such as *the disadvantaged* (as with the Heckman et al., 1996, example earlier in this chapter), it is not surprising to find dramatically different kinds of persons represented under the same label, especially across studies, but often within studies, too.

Regarding settings, the constructs we use again range from simple to complex and controversial. Frequently the settings investigated in a study are a sample of convenience, described as, say, "the Psychology Department Psychological Services Center" based on the researcher's personal experience with the setting, a label that conveys virtually no information about the size of the setting, its funding, client flow, staff, or the range of diagnoses that are encountered. Such clinics, in fact, vary considerably—from small centers with few nonpaying clients who are almost entirely college students and who are seen by graduate students under the supervision of a single staff member to large centers with a large staff of full-time professionals, who themselves see a wide array of diagnostic problems from local communities, in addition to supervising such cases. But settings are often assessed more formally, as with the measures of setting environment developed by Moos (e.g., Moos, 1997) or with descriptors that are inferred from empirical data, as when profile analysis of the characteristics of nursing homes is used to identify different types of nursing homes (e.g., Shadish, Straw, McSweeny, Koller, & Bootzin, 1981).

Regarding treatments, many areas have well-developed traditions of assessing the characteristics of treatments they administer. In laboratory social psychology experiments by Festinger (e.g., 1953) on cognitive dissonance, for example, detailed scripts were prepared to ensure that the prototypical features of cognitive dissonance were included in the study operations; then those scripts were meticulously rehearsed; and finally manipulation checks were used to see whether the participants perceived the study operations to reflect the constructs that were intended. These measurements increase our confidence that the treatment construct was, in fact, delivered. They are, however, difficult to do for complex social programs such as psychotherapy or whole-school reform. In psychotherapy experiments, for example, primary experimenters usually provide only simple labels about the kind of therapy performed (e.g., behavioral, systemic, psychodynamic). Sometimes these labels are accompanied by one- or two-page descriptions of what was done in therapy, and some quantitative measurements such as the number of sessions are usually provided. More sophisticated systems for measuring therapy content are the exception rather than the rule (e.g., Hill, O'Grady, & Elkin, 1992), in part because of their expense and in part because of a paucity of consensually accepted measures of most therapies.

Construct mislabelings often have serious implications for either theory or practice. For example, some persons who score low on intelligence tests have been given labels such as "retarded," though it turned out that their low performance may have been due to language barriers or to insufficient exposure to those aspects of U.S. culture referenced in intelligence tests. The impact on them for school placement and the stigmatization were often enormous. Similarly, the move on the part of some psychotherapy researchers to call a narrow subset of treatments "empirically supported psychological therapies" (Chambless & Hollon, 1998; Kendall, 1998) implies to both researchers and funders that other psychological therapies are not empirically supported, despite several decades of psychotherapy experiments that confirm their effectiveness. When these mislabelings occur in a description of an experiment, they may lead the reader to err in how they apply experimental results to their theory or practice. Indeed, this is one reason that qualitative researchers so much value the "thick description" of study instances (Emerson, 1981; Geertz, 1973; Ryle, 1971)—so that readers of a study can rely more on their own "naturalistic generalizations" than on one researcher's labels (Stake & Trumbull, 1982). We entirely support this aspiration, at least within the limits of reporting conventions that usually apply to experiments; and so we also support the addition of qualitative methodologies to experiments to provide this capacity.

These examples make clear that assessments of study particulars need not be done using formal multi-item scales—though the information obtained would often be better if such scales were used. Rather, assessments include any method for generating data about sampling particulars. They would include archival records, such as patient charts in psychiatric hospitals in which data on diagnosis and symptoms are often recorded by hand or U.S. Census Bureau records in which respondents indicated their racial and ethnic identities by checking a box. They would include qualitative observations, sometimes formal ones such as participant observation or unstructured interviews conducted by a trained anthropologist but often simply the report of the research team who, say, describe a setting as a "poverty neighborhood" based on their personal observations of it as they drive to and from work each day. Assessments may even include some experimental manipulations that are designed to shed light on the nature of study operations, as when a treatment is compared with a placebo control to clarify the extent to which treatment *is* a placebo.

Of course, the attention paid to construct validity in experiments has historically been uneven across persons, treatments, observations, and settings. Concern with construct representations of settings has probably been a low priority, except for researchers interested in the role of environment and culture. Similarly, in most applied experimental research, greater care may go into the construct validity of outcomes, for unless the experimenter uses a measure of recidivism or of employment or of academic achievement that most competent language community members find reasonable, the research is likely to be seen as irrelevant. In basic research, greater attention may be paid to construct validity of the cause so that its link to theory is strong. Such differentiation of priorities is partly functional and may well have evolved to meet needs in a given research field; but it is probably also partly accidental. If so, increased attention to construct validity across persons and settings would probably be beneficial.

The preceding discussion treated persons, treatments, settings, and outcomes separately. But as we mentioned in Chapter 1, construct labels are appropriately applied to relationships among the elements of a study, as well. Labeling the causal relationship between treatment and outcome is a frequent construct validity concern, as when we categorize certain treatments for cancer as cytotoxic or cytostatic to refer to whether they kill tumor cells directly or delay tumor growth by modulating tumor environment. Some other labels have taken on consensual meanings that include more than one feature; the label Medicare in the United States, for example, is nearly universally understood to refer both to the intervention (health care) and to the persons targeted (the elderly).

Threats to Construct Validity

Threats to construct validity (Table 3.1) concern the match between study operations and the constructs used to describe those operations. Sometimes the problem is the explication of constructs, and sometimes it is the sampling or measurement design. A study's operations might not incorporate all the characteristics of the relevant construct (construct underrepresentation), or they may contain extraneous construct content. The threats that follow are specific versions of these more general errors, versions that research or experience have shown tend to occur frequently. The first five threats clearly apply to persons, settings, treatments, and outcomes. The remaining threats primarily concern construct validity of outcomes and especially treatments, mostly carried forward by us from Cook and

TABLE 3.1 Threats to Construct Validity: Reasons Why Inferences About the Constructs That Characterize Study Operations May Be Incorrect

1. *Inadequate Explication of Constructs:* Failure to adequately explicate a construct may lead to incorrect inferences about the relationship between operation and construct.
2. *Construct Confounding:* Operations usually involve more than one construct, and failure to describe all the constructs may result in incomplete construct inferences.
3. *Mono-Operation Bias:* Any one operationalization of a construct both underrepresents the construct of interest and measures irrelevant constructs, complicating inference.
4. *Mono-Method Bias:* When all operationalizations use the same method (e.g., self-report), that method is part of the construct actually studied.
5. *Confounding Constructs with Levels of Constructs:* Inferences about the constructs that best represent study operations may fail to describe the limited levels of the construct that were actually studied.
6. *Treatment Sensitive Factorial Structure:* The structure of a measure may change as a result of treatment, change that may be hidden if the same scoring is always used.
7. *Reactive Self-Report Changes:* Self-reports can be affected by participant motivation to be in a treatment condition, motivation that can change after assignment is made.
8. *Reactivity to the Experimental Situation:* Participant responses reflect not just treatments and measures but also participants' perceptions of the experimental situation, and those perceptions are part of the treatment construct actually tested.
9. *Experimenter Expectancies:* The experimenter can influence participant responses by conveying expectations about desirable responses, and those expectations are part of the treatment construct as actually tested.
10. *Novelty and Disruption Effects:* Participants may respond unusually well to a novel innovation or unusually poorly to one that disrupts their routine, a response that must then be included as part of the treatment construct description.
11. *Compensatory Equalization:* When treatment provides desirable goods or services, administrators, staff, or constituents may provide compensatory goods or services to those not receiving treatment, and this action must then be included as part of the treatment construct description.
12. *Compensatory Rivalry:* Participants not receiving treatment may be motivated to show they can do as well as those receiving treatment, and this compensatory rivalry must then be included as part of the treatment construct description.
13. *Resentful Demoralization:* Participants not receiving a desirable treatment may be so resentful or demoralized that they may respond more negatively than otherwise, and this resentful demoralization must then be included as part of the treatment construct description.
14. *Treatment Diffusion:* Participants may receive services from a condition to which they were not assigned, making construct descriptions of both conditions more difficult.

Campbell's (1979) list. We could have added a host of new threats particular to the construct validity of persons and settings. For example, Table 4.3 in the next chapter lists threats to validity that have been identified by epidemiologists for case-control studies. The threats in that list under the heading "Specifying and selecting the study sample" are particularly relevant to construct validity of persons (i.e., 2d, e, h, k, l, m, q, s, t, u, v) and settings (i.e., 2a, b, c, j). We do not add them here to keep the length of this list tractable. Conceptually, these biases always occur as one of the first five threats listed in Table 3.1; but specific instances of them in Table 4.3 often shed light on common errors that we make in describing people and settings in health contexts.

Inadequate Explication of Constructs

A mismatch between operations and constructs can arise from inadequate analysis of a construct under study. For instance, many definitions of aggression require both intent to harm others and a harmful result. This is to distinguish between (1) the black eye one boy gives another as they collide coming around a blind bend, (2) the black eye that one boy gives another to get his candy (instrumental aggression) or to harm him (noninstrumental aggression), and (3) the verbal threat by one child to another that he will give him a black eye unless the other boy gives him the candy. If both intent and physical harm are part of the definition, only (2) is an instance of aggression. A precise explication of constructs permits tailoring the study instances to whichever definitions emerge from the explication and allows future readers to critique the operations of past studies. When several definitions are reasonable, resources and the extent to which one definition is preferred in the relevant language community play an important role in shaping the research.

Poststudy criticism of construct explications is always called for, however careful the initial explication, because results themselves sometimes suggest the need to reformulate the construct. For example, many researchers have studied the deterrent effects of jail sentences on drunk drivers compared with less severe sanctions such as monetary fines. After many studies showed that jail time did not reduce instances of recidivism, researchers began to question whether jail is experienced as "more severe" than fines (e.g., Martin, Annan, & Forst, 1993). Notice that the finding of no effect is not at issue here (that is an internal validity question), only whether that finding is best characterized as comparing more severe with less severe treatments.

Mark (2000) suggests that researchers make four common errors in explicating constructs: (1) the construct may be identified at too general a level, for example, calling the treatment in a study psychotherapy even though its characteristics make it better described as research psychotherapy (Weisz, Weiss & Donenberg, 1992); (2) the construct may be identified at too specific a level, such as arguing that the levels of unhappiness characteristic of mental patients in nursing homes are really characteristic of mental patients in *any* poverty setting

(Shadish, Silber, Orwin, & Bootzin, 1985); (3) the wrong construct may be identified, as in the case of immigrants to the United States who are labeled retarded because of low scores on intelligence tests when the meaning of their test scores might be better described as lack of familiarity with U.S. language and culture; and (4) a study operation that really reflects two or more constructs may be described using only one construct; for instance, outcome measures that are typically referred to by the names of the traits they measure should also be named for the methods used to measure them (e.g., self-reports of depression). As these examples illustrate, each of these errors occurs in characterizing all four study features—persons, settings, treatments, and outcomes.

Construct Confounding

The operations in an experiment are rarely pure representations of constructs. Consider the example given at the start of this chapter about a study of persons called "unemployed." The researcher may have applied that label as the best representation of the persons actually studied—those whose family income has been below the poverty level for 6 months before the experiment begins or who participate in government welfare or food stamp programs. However, it may also have been the case that these men were disproportionately African-American and victims of racial prejudice. These latter characteristics were not part of the intended construct of the unemployed but were nonetheless confounded with it in the study operations.

Mono-Operation Bias

Many experiments use only one **operationalization** of each construct. Because single operations both underrepresent constructs and may contain irrelevancies, construct validity will be lower in single-operation research than in research in which each construct is multiply operationalized. It is usually inexpensive to use several measures of a given outcome, and this procedure tends to be most prevalent in social science research. Multiple kinds of units and occasionally many different times can be used, too. But most experiments often have only one or two manipulations of an intervention per study and only one setting, because multisite studies are expensive; and increasing the total number of treatments can entail very large sample sizes (or sizes that are too small within each cell in a study with a fixed total sample size). Still, there is no substitute for deliberately varying several exemplars of a treatment. Hence, if one were studying the effects of communicator expertise, one might use, say, three fictitious sources: a distinguished male professor from a well-known university, a distinguished female scientist from a prestigious research center, and a famous science journalist from Germany. The variance due to source differences can then be examined to see if the sources differentially affected responses. If they did, the assumption that communicator expertise is a single construct might be worth revisiting. But even if sample size does not permit analyzing results by each of these sources, the data can still be combined from all three. Then the investigator can test if expertise is effective despite whatever sources of heterogeneity are contained within the three particular operations.

Monomethod Bias

Having more than one operational representation of a construct is helpful, but if all treatments are presented to respondents in the same way, the method itself may influence results. The same is true if all the outcome measures use the same means of recording responses, if all the descriptions of settings rely on an interview with a manager, or if all the person characteristics are taken from hospital charts. Thus, in the previous hypothetical example, if the respondents had been presented with written statements from all the experts, it would be more accurate to label the treatment as *experts presented in writing,* to make clearer that we do not know if the results would hold with experts who are seen or heard. Similarly, attitude scales are often presented to respondents without much thought to (1) using methods of recording other than paper and pencil or (2) varying whether statements are positively or negatively worded. Yet in the first case, different results might occur for physiological measures or for observer ratings, and in the second case, response biases might be fostered when all items are worded in one direction.

Confounding Constructs with Levels of Constructs

Sometimes an experimenter will draw a general conclusion about constructs that fails to recognize that only some levels of each facet of that construct were actually studied and that the results might have been different if different levels were studied (Cooper & Richardson, 1986). In treatment-control comparisons, for example, the treatment may be implemented at such low levels that no effects are observed, leading to an incorrect characterization of the study as showing that treatment had no effect when the correct characterization is that treatment-implemented-at-low-level had no effect. One way to address this threat is to use several levels of treatment. This confounding can be even more complex when comparing two treatments that are operationalized in procedurally nonequivalent ways. The researcher might erroneously conclude that Treatment A works better than Treatment B when the conclusion should have been that Treatment-A-at-Level-1 works better than Treatment-B-at-Level-0. Similar confounding occurs for persons, outcomes, and settings, for example, when restricted person characteristics (e.g., restricted age) or setting characteristics (e.g., using only public schools) were used but this fact was not made clear in the report of the study.

Treatment-Sensitive Factorial Structure

When discussing internal validity previously, we mentioned instrumentation changes that occur even in the absence of treatment. However, instrumentation changes can sometimes occur because of treatment, as when those exposed to an educational treatment learn to see a test in a different way from those not so exposed. For instance, those not getting treatment might respond to an attitude test about people of another race on a largely uniform basis that yields a one-factor test of racial prejudice. Those exposed to treatment might make responses with a more complex factor structure (e.g., "I don't engage in physical harassment or verbal denigration in conversation, but I now see that racial jokes constitute a class of discrimination I did not previously appreciate"). This changed factor structure is itself part of the outcome of the treatment, but few researchers look for different factor structures over groups as an outcome. When all items are summed to a total for both groups, such a summation could mischaracterize the construct being measured, assuming it to be comparable across groups.

Reactive Self-Report Changes

Aiken and West (1990) describe related measurement problems with self-report observations by which both the factorial structure and the level of responses can be affected by whether a person is or is not accepted into the treatment or control group—even before they receive treatment. For example, applicants wanting treatment may make themselves look either more needy or more meritorious (depending on which one they think will get them access to their preferred condition). Once assignment is made, this motivation may end for those who receive treatment but continue for those who do not. Posttest differences then reflect both symptom changes and differential motivation, but the researcher is likely to mistakenly characterize the outcome as only symptom changes. In a similar vein, Bracht and Glass (1968) suggest that posttest (as opposed to pretest) sensitization can occur if the posttest sensitizes participants to the previous intervention they received and so prompts a response that would otherwise not have occurred. Remedies include using external (not self-report) measures that may be less reactive (Webb et al., 1966, 1981); techniques that encourage accurate responding, such as the bogus pipeline, in which participants are monitored by a physiological device they are (falsely) told can detect the correct answer (Jones & Sigall, 1971; Roese & Jamieson, 1993); preventing pretest scores from being available to those allocating treatment; and using explicit reference groups or behavioral criteria to anchor responding.

Reactivity to the Experimental Situation

Humans actively interpret the situations they enter, including experimental treatment conditions, so that the meaning of the molar treatment package includes those reactions. This reactivity takes many forms.[3] Rosenzweig (1933) suggested

[3] See Rosnow and Rosenthal (1997) for a far more extended treatment of this and the next threat, including an analysis of ethical issues and informed consent raised by these threats.

that research participants might try to guess what the experimenter is studying and then try to provide results the researcher wants to see. Orne (1959, 1962, 1969) showed that "demand characteristics" in the experimental situation might provide cues to the participant about expected behavior and that the participant might be motivated (e.g., by altruism or obedience) to comply. Reactivity includes placebo effects due to features of treatment not thought to be "active ingredients" (Shapiro & Shapiro, 1997; L. White, Tursky, & Schwartz, 1985). In drug research, for example, the mere act of being given a pill may cause improvement even if the pill contains only sugar. Rosenberg (1969) provided evidence that respondents are apprehensive about being evaluated by persons who are experts in the outcome and so may respond in ways they think will be seen as competent and psychologically healthy.

Rosenthal and Rosnow (1991) suggest many ways to reduce these problems, including many of those discussed previously with reactive self-report changes, but also by (1) making the dependent variable less obvious by measuring it outside the experimental setting, (2) measuring the outcome at a point much later in time, (3) avoiding pretests that provide cues about expected outcomes, (4) using the Solomon Four-Group Design to assess the presence of the problem, (5) standardizing or reducing experimenter interactions with participants, (6) using masking procedures that prevent participants and experimenters from knowing hypotheses,[4] (7) using deception when ethical by providing false hypotheses, (8) using quasi-control participants who are told about procedures and asked how they think they should respond, (9) finding a preexperimental way of satisfying the participant's desire to please the experimenter that satiates their motivation, and (10) making the conditions less threatening to reduce evaluation apprehension, including ensuring anonymity and confidentiality. These solutions are at best partial because it is impossible to prevent respondents from generating their own treatment-related hypotheses and because in field settings it is often impossible, impractical, or unethical to do some of them.

Experimenter Expectancies

A similar class of problems was suggested by Rosenthal (1956): that the *experimenter's* expectancies are also a part of the molar treatment package and can influence outcomes. Rosenthal first took note of the problem in clinical psychology in his own dissertation on the experimental induction of defense mechanisms. He developed the idea extensively in laboratory research, especially in social psychology. But it has also been demonstrated in field research. In education, for example, the problem includes the Pygmalion effect, whereby teachers' expectancies

[4] These procedures were called "blinding" in the past, as with double-blind designs; but we follow the recommendation of the American Psychological Association's (1994) Publication Manual in referring to masking rather than blinding.

about student achievements become self-fulfilling prophecies (Rosenthal, 1973a, 1973b). Those parts of the placebo effect from the previous threat that are induced by experimenter expectancies—such as a nurse telling a patient that a pill will help, even if the pill is an inert placebo—fall under this category as well. To reduce the problem, Rosenthal and Rosnow (1991) suggest (1) using more experimenters, especially if their expectancies can be manipulated or studied, (2) observing experimenters to detect and reduce expectancy-inducing behavior, (3) using masking procedures in which those who administer treatments do not know the hypotheses, (4) minimizing contacts between experimenter and participant, and (5) using control groups to assess the presence of these problems, such as placebo controls.

Novelty and Disruption Effects

Bracht and Glass (1968) suggested that when an innovation is introduced, it can breed excitement, energy, and enthusiasm that contribute to success, especially if little innovation previously occurred.[5] After many years of innovation, however, introducing another one may not elicit welcoming reactions, making treatment less effective. Conversely, introducing an innovation may also be quite disruptive, especially if it impedes implementation of current effective services. The innovation may then be less effective. Novelty and disruption are both part of the molar treatment package.

Compensatory Equalization

When treatment provides desirable goods or services, administrators, staff, or constituents may resist the focused inequality that results (Stevens, 1994).[6] For example, Schumacher and colleagues (1994) describe a study in which usual day care for homeless persons with substance abuse problems was compared with an enhanced day treatment condition. Service providers complained about the inequity and provided some enhanced services to clients receiving usual care. Thus the planned contrast broke down. Equalization can also involve taking benefits away from treatment recipients rather than adding them for control group members. In one study,

[5] One instance of this threat is frequently called the "Hawthorne effect" after studies at Western Electric Company's Hawthorne site (Roethlisberger & Dickson, 1939). In an early interpretation of this research, it was thought that participants responded to the attention being given to them by increasing their productivity, whatever the treatment was. This interpretation has been called into question (e.g., Adair, 1973; Bramel & Friend, 1981; Gillespie, 1988); but the label "Hawthorne effect" is likely to continue to be used to describe it.

[6] Cook and Campbell's (1979) previous discussion of this threat and the next three (resentful demoralization, compensatory rivalry, diffusion) may have misled some readers (e.g., Conrad & Conrad, 1994) into thinking that they occur only with random assignment. To the contrary, they result from a comparison process that can occur in any study in which participants are aware of discrepancies between what they received and what they might have received. Such comparison processes occur in quasi-experiments and are not even limited to research studies (see J. Z. Shapiro, 1984, for an example in a regression discontinuity design).

lawyers in a district attorney's office thought the treatment condition was too favorable to defendants and so refused to plea bargain with them at all (compensatory deprivation), as the treatment required them to do (Wallace, 1987). Such focused inequities may explain some administrators' reluctance to employ random assignment when they believe their constituencies want one treatment more than another. To assess this problem, interviews with administrators, staff, and participants are invaluable.

Compensatory Rivalry

Public assignment of units to experimental and control conditions may cause social competition, whereby the control group tries to show it can do as well as the treatment group despite not getting treatment benefits. Saretsky (1972) called this a "John Henry effect" after the steel driver who, when he knew his output was to be compared with that of a steam drill, worked so hard that he outperformed the drill and died of overexertion. Saretsky gave the example of an education experiment in which the success of treatment-group performance contractors (commercial contractors paid according to the size of learning gains made by students) would threaten the job security of control teachers who might be replaced by those contractors. Hence teachers in the control groups may have performed much better than usual to avoid this possibility. Saretsky (1972), Fetterman (1982), and Walther and Ross (1982) describe other examples. Qualitative methods such as unstructured interviews and direct observation can help discover such effects. Saretsky (1972) tried to detect the effects by comparing performance in current control classes to the performance in the same classes in the years before the experiment began.

Resentful Demoralization

Conversely, members of a group receiving a less desirable treatment or no treatment can be resentful and demoralized, changing their responses to the outcome measures (Bishop & Hill, 1971; Hand & Slocum, 1972; J. Z. Shapiro, 1984; Walther & Ross, 1982). Fetterman (1982) describes an evaluation of an educational program that solicited unemployed high school dropouts to give them a second chance at a career orientation and a high school diploma. Although the design called for assigning only one fourth of applicants to the control group so as to maximize participation, those assigned to the control group were often profoundly demoralized. Many had low academic confidence and had to muster up courage just to take one more chance, a chance that may really have been their last chance rather than a second chance. Resentful demoralization is not always this serious, but the example highlights the ethical problems it can cause. Of course, it is wrong to think that participant reactions are uniform. Lam, Hartwell, and Jekel (1994) show that those denied treatment report diverse reactions. Finally, Schumacher et al. (1994) show how resentful demoralization can occur in a group assigned to a *more* desirable treatment—client expectations for enhanced services were raised but then dashed when funds were cut and community resistance to proposed housing arrangements emerged. Reactivity problems can occur not just in reaction to other groups but also to one's own hopes for the future.

Treatment Diffusion

Sometimes the participants in one condition receive some or all of the treatment in the other condition. For example, in Florida's Trade Welfare for Work experiment, about one fourth of all control group participants crossed over to receive the job-training treatment (D. Greenberg & Shroder, 1997). Although these crossovers were discovered by the researchers, participants who cross over often do it surreptitiously for fear that the researcher would stop the diffusion, so the researcher is frequently unaware of it. The problem is most acute in cases in which experimental and control units are in physical proximity or can communicate. For example, if Massachusetts is used as a control group to study the effects of changes in a New York abortion law, the true effects of the law would be obscured if those from Massachusetts went freely to New York for abortions. Diffusion can occur when both conditions are exposed to the same treatment providers, as in a study comparing behavior therapy with eclectic psychotherapy. The same therapists administered both treatments, and one therapist used extensive behavioral techniques in the eclectic condition (Kazdin, 1992). Preventing diffusion is best achieved by minimizing common influences over conditions (e.g., using different therapists for each condition) and by isolating participants in each condition from those in other conditions (e.g., using geographically separate units). When this is not practical, measurement of treatment implementation in both groups helps, for a small or nonexistent experimental contrast on implementation measures suggests that diffusion may have occurred (see Chapter 10).

Construct Validity, Preexperimental Tailoring, and Postexperimental Specification

The process of assessing and understanding constructs is never fully done. The preceding treatment of construct validity emphasizes that, before the experiment begins, the researcher should critically (1) think through how constructs should be defined, (2) differentiate them from cognate constructs, and (3) decide how to index each construct of interest. We might call this the domain of intended application. Then we emphasized (4) the need to use multiple operations to index each construct when possible (e.g., multiple measures, manipulations, settings, and units) and when no single way is clearly best. We also indicated (5) the need to ensure that each of the multiple operations reflects multiple methods so that single-method confounds (e.g., self-report biases) can be better assessed.

After the data have been collected and provisionally analyzed, researchers may reconsider the extent to which the initially conceptualized construct has or has not been achieved (the domain of achieved application), perhaps because the planned operations were not implemented as intended or because evidence suggests that constructs other than the intended ones may better represent what the study actually did. Some postexperimental respecification of constructs is almost inevitable, particularly in programs of research. Imagine an experiment intended to compare more credible with less credible communicators in which a difference on the outcome measure is found. If a reliable measure of communicator credibility suggests that a communicator was not perceived to be more credible in one experimental group than in another, the investigator is forced to use whatever means are available to specify what might have caused the observed effects if credibility did not. Or suppose that a manipulation affected two reliably measured exemplars of a particular construct but not three other reliable measures of the same construct. R. Feldman's (1968) experiment in Boston, Athens, and Paris used five measures of cooperation (the construct as conceived at the start of the study) to test whether compatriots receive greater cooperation than foreigners. The measures were: giving street directions; doing a favor by mailing a lost letter; giving back money that one could easily, but falsely, claim as one's own; giving correct change when one did not have to; and charging the correct amount to passengers in taxis. The data suggested that giving street directions and mailing the lost letter were differently related to the experimental manipulations than were forgoing chances to cheat in ways that would be to one's advantage. This forced Feldman to specify two kinds of cooperation (low-cost favors versus forgoing one's own financial advantage). However, the process of hypothesizing constructs and testing how well operations fit these constructs is similar both before the research begins and after the data are received.

Once a study has been completed, disagreements about how well a given study represents various constructs are common, with critics frequently leveling the charge that different constructs were sampled or operationalized from those the researcher claims was the case. Because construct validity entails socially creating *and recreating* the meanings of research operations, lasting resolutions are rare, and constructs are often revisited. Fortunately, these disagreements about the composition of constructs and about the best way to measure them make for better inferences about constructs because they can be successfully tested, not only across overlapping operational representations of the same definition but also across different (but overlapping) definitions of the same construct. For example, various language communities disagree about whether to include intent to harm as part of the construct of aggression. It is only when we have learned that such intent makes little difference to actual study outcomes that we can safely omit it from our description of the concept of aggression. Disagreements about construct definitions are potentially of great utility, therefore.

EXTERNAL VALIDITY

External validity concerns inferences about the extent to which a causal relationship holds over variations in persons, settings, treatments, and outcomes. For example, the Transitional Employment Training Demonstration experiment randomly assigned 18- to 40-year-old adults with mental retardation to either a control condition receiving usual services or a treatment condition that received job training along with unsubsidized and potentially permanent jobs (Greenberg & Shroder, 1997). Results showed that the treatment improved both job placement and earnings. Yet the researchers noted serious remaining questions about the external validity of these effects. For example, their own data suggested that results were larger for participants with higher IQs and that participants with IQs less than 40 showed little or no gain; and their between-site analyses showed that success rates depended greatly on the kind of job in which the site tried to place the participant. The researchers also raised other external validity questions that their data did not allow them to explore. For example, the program was implemented in 12 sites in the United States, but no sites were in the South. In addition, only 5% of those who were sent invitation letters volunteered to participate; and of these, two thirds were screened out because they did not meet study eligibility criteria that included lack of severe emotional problems and likelihood of benefiting from treatment. Whether results would also be found in more severely disturbed, nonvolunteer retarded adults remains at issue. Further, the researchers noted that successful program participants were more adventuresome and willing to move from the well-established and comparatively safe confines of traditional sheltered employment into the real world of employment; they questioned whether less adventuresome retarded adults would show the same benefits.

As this example shows, external validity questions can be about whether a causal relationship holds (1) over variations in persons, settings, treatments, and outcomes that *were* in the experiment and (2) for persons, settings, treatments, and outcomes that *were not* in the experiment. Targets of generalization can be quite diverse:

- *Narrow to Broad:* For instance, from the persons, settings, treatments, and outcomes in an experiment to a larger population, as when a policymaker asks if the findings from the income maintenance experiments in New Jersey, Seattle, and Denver would generalize to the U.S. population if adopted as national policy.
- *Broad to Narrow:* From the experimental sample to a smaller group or even to a single person, as when an advanced breast cancer patient asks whether a newly-developed treatment that improves survival in general would improve her survival in particular, given her pathology, her clinical stage, and her prior treatments.
- *At a Similar Level:* From the experimental sample to another sample at about the same level of aggregation, as when a state governor considers adapting a new welfare reform based on experimental findings supporting that reform in a nearby state of similar size.

- *To a Similar or Different Kind:* In all three of the preceding cases, the targets of generalization might be similar to the experimental samples (e.g., from male job applicants in Seattle to male job applicants in the United States) or very different (e.g., from African American males in New Jersey to Hispanic females in Houston).
- *Random Sample to Population Members:* In those rare cases with random sampling, a generalization can be made from the random sample to other members of the population from which the sample was drawn.

Cronbach and his colleagues (Cronbach et al., 1980; Cronbach, 1982) argue that most external validity questions are about persons, settings, treatments, and outcomes that *were not* studied in the experiment—because they arise only after a study is done, too late to include the instances in question in the study. Some scientists reject this version of the external validity question (except when random sampling is used). They argue that scientists should be held responsible only for answering the questions they pose and study, not questions that others might pose later about conditions of application that might be different from the original ones. They argue that inferences to as-yet-unstudied applications are no business of science until they can be answered by reanalyses of an existing study or by a new study.

On this disagreement, we side with Cronbach. Inferences from completed studies to as-yet-unstudied applications are necessary to both science and society. During the last two decades of the 20th century, for example, researchers at the U.S. General Accounting Office's Program Evaluation and Methodology Division frequently advised Congress about policy based on reviews of past studies that overlap only partially with the exact application that Congress has in mind (e.g., Droitcour, 1997). In a very real sense, in fact, the essence of creative science is to move a program of research forward by incremental extensions of both theory and experiment into untested realms that the scientist believes are likely to have fruitful yields given past knowledge (e.g., McGuire, 1997). Usually, such extrapolations are justified because they are incremental variations in some rather than all study features, making these extrapolations to things not yet studied more plausible. For example, questions may arise about whether the effects of a worksite smoking prevention program that was studied in the private sector would generalize to public sector settings. Even though the public sector setting may never have been studied before, it is still a work site, the treatment and observations are likely to remain substantially the same, and the people studied tend to share many key characteristics, such as being smokers. External validity questions about what would happen if *all* features of a study were different are possible but are so rare in practice that we cannot even construct a plausible example.

On the other hand, it is also wrong to limit external validity to questions about as-yet-unstudied instances. Campbell and Stanley (1963) made no such distinction in their original formulation of external validity as asking the question, "to what populations, settings, treatment variables, and measurement variables can this effect be generalized?" (p. 5). Indeed, one of the goals of their theory was to point out "numerous ways in which experiments can be made more valid externally" (p. 17). For example, they said that external validity was enhanced in single studies "if in the initial experiment we have demonstrated the phenomenon over a wide variety of conditions" (p. 19) and also enhanced by inducing "maximum similarity of experiments to the conditions of application" (p. 18). This goal of designing experiments to yield inferences that are "more valid externally" is not a novel concept. To the contrary, most experiments already test whether treatment effects hold over several different outcomes. Many also report whether the effect holds over different kinds of persons, although power is reduced as a sample is subdivided. Tests of effect stability over variations in treatments are limited to studies having multiple treatments, but these do occur regularly in the scientific literature (e.g., Wampold et al., 1997). And tests of how well causal relationships hold over settings also occur regularly, for example, in education (Raudenbush & Willms, 1995) and in large multisite medical and public health trials (Ioannidis et al., 1999).

Yet there are clear limits to this strategy. Few investigators are omniscient enough to anticipate all the conditions that might affect a causal relationship. Even if they were that omniscient, a full solution requires the experiment to include a fully heterogeneous range of units, treatments, observations, and settings. Diversifying outcomes is usually feasible. But in using multiple sites *or* many operationalizations of treatments *or* tests of causal relationships broken down by various participant characteristics, each becomes increasingly difficult; and doing all of them at once is impossibly expensive and logistically complex. Even if an experiment had the requisite diversity, detecting such interactions is more difficult than detecting treatment main effects. Although power to detect interactions can be increased by using certain designs (e.g., West, Aiken, & Todd, 1993), these designs must be implemented before the study starts, which makes their use irrelevant to the majority of questions about external validity that arise after a study is completed. Moreover, researchers often have an excellent reason *not* to diversify all these characteristics—after all, extraneous variation in settings and respondents is a threat to statistical conclusion validity. So when heterogeneous sampling is done because interactions are expected, the total sample size must be increased to obtain adequate power. This, too, costs money that could be used to improve other design characteristics. In a world of limited resources, designing studies to anticipate external validity questions will often conflict with other design priorities that may require precedence.

Sometimes, when the original study included the pertinent variable but did not analyze or report it, then the original investigator or others (in the latter case, called **secondary analysis**; Kiecolt & Nathan, 1990) can reanalyze data from the experiment to see what happens to the causal relationship as the variable in question is varied. For example, a study on the effects of a weight-loss program may have found it to be effective in a sample composed of both men and women. Later, a question may arise about whether the results would have held separately for both men and women. If the original data can be accessed, and if they were coded and stored in such a way that the analysis is possible, the question can be addressed by reanalyzing the data to test this interaction.

Usually, however, the original data set is either no longer available or does not contain the required data. In such cases, reviews of published results from many studies on the same question are often excellent sources for answering external validity questions. As Campbell and Stanley (1963) noted, we usually "learn how far we can generalize an internally valid finding only piece by piece through trial and error" (p. 19), typically over multiple studies that contain different kinds of persons, settings, treatments, and outcomes. Scientists do this by conducting programs of research during their research careers, a time-consuming process that gives maximum control over the particular generalizations at issue. Scientists also do this by combining their own work with that of other scientists, combining basic and applied research or laboratory and field studies, as Dwyer and Flesch-Janys (1995) did in their review of the effects of Agent Orange in Vietnam. Finally, scientists do this by conducting quantitative reviews of many experiments that addressed a common question. Such **meta-analysis** is more feasible than secondary analysis because it does not require the original data. However, meta-analysis has problems of its own, such as poor quality of reporting or statistical analysis in some studies. Chapter 13 of this book discusses all these methods.

Threats to External Validity

Estimates of the extent to which a causal relationship holds over variations in persons, settings, treatments, and outcomes are conceptually similar to tests of statistical interactions. If an interaction exists between, say, an educational treatment and the social class of children, then we cannot say that the same result holds across social classes. We know that it does not, for the significant interaction shows that the effect size is different in different social classes. Consequently, we have chosen to list threats to external validity in terms of interactions (Table 3.2) of the causal relationship (including mediators of that relationship) with (1) units, (2) treatments, (3) outcomes, and (4) settings.

However, our use of the word *interaction* in naming these threats is not intended to limit them to statistical interactions. Rather, it is the concept behind the interaction that is important—the search for ways in which a causal relationship might or might not change over persons, settings, treatments, and outcomes. If that question can be answered using an interaction that can be quantified and tested statistically, well and good. But the inability to do so should not stop the search for these threats. For example, in the case of generalizations to persons, settings, treatments, and outcomes that were not studied, no statistical test of interactions is possible. But this does not stop researchers from generating plausible hypotheses about likely interactions, sometimes based on professional experience and sometimes on related studies, with which to criticize the generalizability of

TABLE 3.2 Threats to External Validity: Reasons Why Inferences About How Study Results Would Hold Over Variations in Persons, Settings, Treatments, and Outcomes May Be Incorrect

1. *Interaction of the Causal Relationship with Units:* An effect found with certain kinds of units might not hold if other kinds of units had been studied.
2. *Interaction of the Causal Relationship Over Treatment Variations:* An effect found with one treatment variation might not hold with other variations of that treatment, or when that treatment is combined with other treatments, or when only part of that treatment is used.
3. *Interaction of the Causal Relationship with Outcomes:* An effect found on one kind of outcome observation may not hold if other outcome observations were used.
4. *Interactions of the Causal Relationship with Settings:* An effect found in one kind of setting may not hold if other kinds of settings were to be used.
5. *Context-Dependent Mediation:* An explanatory mediator of a causal relationship in one context may not mediate in another context.

experimental results and around which to design new studies. Nor should we be slaves to the statistical significance of interactions. Nonsignificant interactions may reflect low power, yet the result may still be of sufficient practical importance to be grounds for further research. Conversely, significant interactions may be demonstrably trivial for practice or theory. At issue, then, is not just the statistical significance of interactions but also their practical and theoretical significance; not just their demonstration in a data set but also their potential fruitfulness in generating compelling lines of research about the limits of causal relationships.

Interaction of Causal Relationship with Units

In which units does a cause-effect relationship hold? For example, common belief in the 1980s in the United States was that health research was disproportionately conducted on white males—to the point where the quip became, "Even the rats were white males," because the most commonly used rats were white in color and because to facilitate homogeneity only male rats were studied.[7] Researchers became concerned that effects observed with human white males might not hold equally well for females and for more diverse ethnic groups, so the U.S. National

[7] Regarding gender, this problem may not have been as prevalent as feared. Meinart, Gilpin, Unalp, and Dawson (2000) reviewed 724 clinical trials appearing between 1966 and 1998 in *Annals of Internal Medicine, British Medical Journal, Journal of the American Medical Association, Lancet,* and *New England Journal of Medicine.* They found in the U.S. journals that 55.2% of those trials contained both males and females, 12.2% contained males only, 11.2% females only, and 21.4% did not specify gender. Over all journals, 355,624 males and 550,743 females were included in these trials.

Institutes of Health (National Institutes of Health, 1994) launched formal initiatives to ensure that such variability is systematically examined in the future (Hohmann & Parron, 1996). Even when participants in an experiment belong to the target class of interest (e.g., African American females), those who are successfully recruited into an experiment may differ systematically from those who are not. They may be volunteers, exhibitionists, hypochondriacs, scientific do-gooders, those who need the proffered cash, those who need course credit, those who are desperate for help, or those who have nothing else to do. In the Arkansas Work Program experiment, for example, the program intentionally selected the most job-ready applicants to treat, and such "creaming" may result in effect estimates that are higher than those that would have been obtained for less job-ready applicants (Greenberg & Shroder, 1997). Similarly, when the unit is an aggregate such as a school, the volunteering organizations may be the most progressive, proud, or self-confident. For example, Campbell (1956), although working with the Office of Naval Research, could not get access to destroyer crews and had to use high-morale submarine crews. Can we generalize from such situations to those in which morale is lower?

Interaction of Causal Relationship Over Treatment Variations

Here, the size or direction of a causal relationship varies over different treatment variations. For example, reducing class size may work well when it is accompanied by substantial new funding to build new classrooms and hire skilled teachers, but it may work poorly if that funding is lacking, so that the new small classes are taught in temporary trailers by inexperienced teachers. Similarly, because of the limited duration of most experimental treatments, people may react differently than they would if the treatment were extended. Thus, in the New Jersey Income Maintenance Experiment, respondents reacted to an income that was guaranteed to them for 3 years only. Because of suspicion that the respondents would react differently if the treatment lasted longer, the later Seattle-Denver Income Maintenance Experiment contained some families whose benefits were guaranteed for 20 years, more like a permanent program (Orr, 1999). Similarly, the effects in a small-scale experimental test might be quite different from those in a full-scale implementation of the same treatment (Garfinkel, Manski, & Michalopoulos, 1992; Manski & Garfinkel, 1992). For example, this could happen if a social intervention is intended to cause changes in community attitudes and norms that could occur only when the intervention is widely implemented. In such cases, social experiments that are implemented on a smaller scale than that of the intended policy implementation might not cause these community changes. Finally, this threat also includes interaction effects that occur when treatments are administered jointly. Drug interaction effects are a well-known example. A drug may have a very positive effect by itself, but when used in combination with other drugs may be either deadly (the interaction of Viagra with certain blood pressure medications) or totally ineffective (the interaction of some antibiotics with dairy products). Conversely, a combination of drugs to treat AIDS may dramatically reduce death, but each drug by itself might be ineffective.

Interaction of Causal Relationship with Outcomes

Can a cause-effect relationship be generalized over different outcomes? In cancer research, for example, treatments vary in effectiveness depending on whether the outcome is quality of life, 5-year metastasis-free survival, or overall survival; yet only the latter is what laypersons understand as a "cure." Similarly, when social science results are presented to audiences, it is very common to hear comments such as: "Yes, I accept that the youth job-training program increases the likelihood of being employed immediately after graduation. But what does it do to adaptive job skills such as punctuality or the ability to follow orders?" Answers to such questions give a fuller picture of a treatment's total impact. Sometimes treatments will have a positive effect on one outcome, no effect on a second, and a negative effect on a third. In the New Jersey Income Maintenance Experiment, for example, income maintenance payments reduced the number of hours worked by wives in experimental families, had no effect on home ownership or major appliance purchases, and increased the likelihood that teenagers in experimental families would complete high school (Kershaw & Fair, 1977; Watts & Rees, 1976). Fortunately, this is the easiest study feature to vary. Consultation with **stakeholders** prior to study design is an excellent method for ensuring that likely questions about generalizability over outcomes are anticipated in the study design.

Interaction of Causal Relationship with Settings

In which settings does a cause-effect relationship hold? For example, Kazdin (1992) described a program for drug abusers that was effective in rural areas but did not work in urban areas, perhaps because drugs are more easily available in the latter settings. In principle, answers to such questions can be obtained by varying settings and analyzing for a causal relationship within each. But this is often costly, so that such options are rarely feasible. Sometimes, though, a single large site (e.g., a university) has some subsettings (e.g., different departments) that vary naturally along dimensions that might affect outcome, allowing some study of generalizability. Large multisite studies also have the capacity to address such issues (Turpin & Sinacore, 1991), and they are doing an increasingly sophisticated job of exploring the reasons why sites differ (Raudenbush & Willms, 1991).

Context-Dependent Mediation

Causal explanation is one of the five principles of causal generalization in the grounded theory we outlined in Chapter 1. Though we discuss this principle in more detail in Chapter 12, one part of explanation is identification of mediating processes. The idea is that studies of causal mediation identify the essential processes that must occur in order to transfer an effect. However, even if a correct mediator is identified in one context, that variable may not mediate the effect in another context. For example, a study of the effects of a new health care insurance program in nonprofit hospitals might show that the program reduces costs through a reduction in the number of middle management positions. But this explanation might not generalize to for-profit hospitals in which, even if the cost reduction does occur, it may occur through reduction in patient services instead. In this example, the contextual change is settings, but it could also be a change in the persons studied or in the nature of the treatment or the outcome variables used. Context dependencies in any of these are also interactions—in this case an interaction of the mediator in the causal relationship with whatever feature of the context was varied. When such mediator variables can be identified and studied over multiple contexts, their consistency as mediators can be tested using multigroup structural equation models.

Constancy of Effect Size Versus Constancy of Causal Direction

We have phrased threats to external validity as interactions. How large must these interactions be to threaten generalization? Does just a tiny change in effect size count as a failure to generalize? These questions are important statistically because a study with high power can detect even small variations in effect sizes over levels of a potential moderator. They are important philosophically because many theorists believe that the world is full of interactions by its very nature, so that statistical main effects will rarely describe the world with perfect accuracy (Mackie, 1974). And they are important practically because some scientists claim that complex statistical interactions are the norm, including Cronbach and Snow (1977) in education, Magnusson (2000) in developmental science, and McGuire (1984) in social psychology. It is entirely possible, then, that if robustness were specified as *constancy of effect sizes,* few causal relationships in the social world would be generalizable.

However, we believe that generalization often is appropriately conceptualized as *constancy of causal direction,* that the sign of the causal relationship is constant across levels of a moderator. Several factors argue for this. First, casual examination of many meta-analyses convinces us that, for at least some topics in which treatments are compared with control groups, causal signs often tend to be similar across individual studies even when the effect sizes vary considerably (e.g., Shadish, 1992a). Second, in the social policy world it is difficult to shape legislation or regulations to suit local contingencies. Instead, the same plan has to be promulgated across an entire nation or state to avoid focused inequities between individual places or groups. Policymakers hope for positive effects overall, despite the inevitable variability in effect sizes from site to site or person to person or over different kinds of outcomes or different ways of delivering the treatment. Their fear is of different causal signs over these variations. Third, substantive theories are usually built around causal relationships whose occurrence is particularly dependable, not just those that are obviously novel. The former reduces the risk of theorizing about unstable phenomena—an unfortunate commonplace in much of today's social science! Fourth, the very nature of scientific theory is that it reduces complex phenomena to simpler terms, and minor fluctuations in effect size are often irrelevant to basic theoretical points. Because defining robustness in terms of constant effect sizes loses all these advantages, we favor a looser criterion based on the stability of causal signs, especially when research that some might call applied is involved.

Nonetheless, we would not abandon constancy of effect size entirely, for sometimes small differences in effect size have large practical or theoretical importance. An example is the case in which the outcome of interest is a harm, such as death. For instance, if the addition of an angiogenesis inhibitor to chemotherapy increases life expectancy in prostate cancer patients by only 6 months but the cost of the drug is low and it has no significant side effects, then many patients and their physicians would want that addition because of the value they place on having even just a little more time to live. Such judgments take into account individual differences in the value placed on small differences in effects, estimates of the contextual costs and benefits of the intervention, and knowledge of possible side effects of treatment. Again, judgments about the external validity of a causal relationship cannot be reduced to statistical terms.

Random Sampling and External Validity

We have not put much emphasis on random sampling for external validity, primarily because it is so rarely feasible in experiments. When it is feasible, however, we strongly recommend it, for just as random assignment simplifies internal validity inferences, so random sampling simplifies external validity inferences (assuming little or no attrition, as with random assignment). For example, if an experimenter randomly samples persons before randomly assigning them to conditions, then random sampling guarantees—within the limits of sampling error—that the average causal relationship observed in the sample will be the same as (1) the average causal relationship that would have been observed in any other random sample of persons of the same size from the same population and (2) the average causal relationship that would have been observed across *all* other persons in that population who were not in the original random sample. That is, random sampling eliminates possible interactions between the causal relationship and the class of persons who are studied versus the class of persons who are not studied within the same population. We cite examples of such experiments in Chapter 11, though they are rare. Further, suppose the researcher also tests the interaction of treatment with a characteristic of persons (e.g., gender). Random sampling also guarantees that interaction will be the same in the groups defined in (1) and (2)—although power decreases as samples are subdivided. So, although we argue in Chapters 1 and 11 that random sampling has major practical limitations when combined with experiments, its benefits for external validity are so great that it should be used on those rare occasions when it is feasible.

These benefits hold for random samples of settings, too. For example, Puma, Burstein, Merrell, and Silverstein (1990) randomly sampled food stamp agencies in one randomized experiment about the Food Stamp Employment and Training Program. But random samples of settings are even more rare in experiments than are random samples of persons. Although defined populations of settings are fairly common—for example, Head Start Centers, mental health centers, or hospitals—the rarity of random sampling from these populations is probably due to the logistical costs of successfully randomly sampling from them, costs that must be added to the already high costs of multisite experiments.

Finally, these benefits also would hold for treatments and outcomes. But lists of treatments (e.g., Steiner & Gingrich, 2000) and outcomes (e.g., American Psychiatric Association, 2000) are rare, and efforts to defend random sampling from them are probably nonexistent. In the former case, this rarity exists because the motivation to experiment in any given study stems from questions about the effects of a particular treatment, and in the latter case it exists because most researchers probably believe diversity in outcome measures is better achieved by more deliberate methods such as the following.

Purposive Sampling and External Validity

Purposive sampling of heterogeneous instances is much more frequently used in single experiments than is random sampling; that is, persons, settings, treatments, or outcomes are deliberately chosen to be diverse on variables that are presumed to be important to the causal relationship. For instance, if there is reason to be concerned that gender might moderate an effect, then both males and females are deliberately included. Doing so has two benefits for external validity. Most obviously, it allows tests of the interaction between the causal relationship and gender in the study data. If an interaction is detected, this is prima facie evidence of limited external validity. However, sometimes sample sizes are so small that responsible tests of interactions cannot be done, and in any case there will be many potential moderators that the experimenter does not think to test. In these cases, heterogeneous sampling still has the benefit of demonstrating that a main effect for treatment occurs *despite* the heterogeneity in the sample. Of course, random sampling demonstrates this even more effectively, for it makes the sample heterogeneous on every possible moderator variable; but deliberately heterogeneous sampling makes up for its weakness by being practical.

The same benefits ensue from purposive sampling of heterogeneous settings, so that it is common in multisite research to ensure some diversity in including, say, both public and private schools or both nonprofit and proprietary hospitals. Purposive sampling of heterogeneous outcome measures is so common in most areas of field experimentation that its value for exploring the generalizability of the effect is taken for granted, though there is surprisingly little theory trying to explain or predict such variability (e.g., Shadish & Sweeney, 1991). Purposive sampling of heterogeneous treatments in single experiments is again probably nonexistent, for the same reasons for which random sampling of treatments is not done. However, over a program of research or over a set of studies conducted by many different researchers, heterogeneity is frequently high for persons, settings, treatments, and outcomes. This is one reason that our grounded theory of causal generalization relies so heavily on methods for multiple studies.

MORE ON RELATIONSHIPS, TRADEOFFS, AND PRIORITIES

At the end of Chapter 2, we discussed the relationship between internal and statistical conclusion validity. We now extend that discussion to other relationships between validity types and to priorities and tradeoffs among them.

The Relationship Between Construct Validity and External Validity

Construct validity and external validity are related to each other in two ways. First, both are generalizations. Consequently, the grounded theory of generalization that we briefly described in Chapter 1 and that we extend significantly in Chapters 11 through 13 helps enhance both kinds of validities. Second, valid knowledge of the constructs that are involved in a study can shed light on external validity questions, especially if a well-developed theory exists that describes how various constructs and instances are related to each other. Medicine, for example, has well-developed theories for categorizing certain therapies (say, the class of drugs we call chemotherapies for cancer) and for knowing how these therapies affect patients (how they affect blood tests and survival and what their side effects are). Consequently, when a new drug meets the criteria for being called a chemotherapy, we can predict much of its likely performance before actually testing it (e.g., we can say it is likely to cause hair loss and nausea and to increase survival in patients with low tumor burdens but not advanced cases). This knowledge makes the design of new experiments easier by narrowing the scope of pertinent patients and outcomes, and it makes extrapolations about treatment effects likely to be more accurate. But once we move from these cases to most topics of field experimentation in this book, such well-developed theories are mostly lacking. In these common cases, knowledge of construct validity provides only weak evidence about external validity. We provide some examples of how this occurs in Chapters 11 through 13.

However, construct and external validity are different from each other in more ways than they are similar. First, they differ in the kinds of inferences being made. The inference of construct validity is, by definition, always a *construct* that is applied to study instances. For external validity generalizations, the inference concerns whether the size or direction of a causal relationship changes over persons, treatments, settings, or outcomes. A challenge to construct validity might be that we have mischaracterized the settings in a health care study as private sector hospitals and that it would have been more accurate to call them private nonprofit hospitals to distinguish them from the for-profit hospitals that were not in the study. In raising this challenge, the size or direction of the causal relationship need never be mentioned.

Second, external validity generalizations cannot be divorced from the causal relationship under study, but questions about construct validity can be. This point is most clear in the phrasing of threats to external validity, which are always interactions of the *causal relationship* with some other real or potential persons, treatments, settings, or outcomes. There is no external validity threat about, say, the interaction of persons and settings without reference to the causal relationship. It is not that such interactions could not happen—it is well known, for example, that the number of persons with different psychiatric diagnoses that one finds in state mental hospitals is quite different from the number generally found in the outpatient offices of a private sector clinical psychologist. We can even raise construct validity questions about all these labels (e.g., did we properly label the setting as a state mental hospital? Or might it have been better characterized as a state psychiatric long-term-care facility to distinguish it from those state-run facilities that treat only short-term cases?). But because this particular interaction did not involve a causal relationship, it cannot be about external validity.

Of course, in practice we use abstract labels when we raise external validity questions. In the real world of science, no one would say, "I think this causal relationship holds for units on List A but not for units on List B." Rather, they might say, "I think that gene therapies for cancer are likely to work for patients with low tumor burden rather than with high tumor burden." But the use of construct labels in this latter sentence does not make external validity the same as, or even dependent on, construct validity. The parallel with internal validity is instructive here. No one in the real world of science ever talks about whether A caused B. Rather, they always talk about descriptive causal relationships in terms of constructs, say, that gene therapy increased 5-year survival rates. Yet we have phrased internal validity as concerning whether A caused B *without* construct labels in order to highlight the fact that the logical issues involved in validating a descriptive causal inference (i.e., whether cause precedes effect, whether alternative causes can be ruled out, and so forth) are orthogonal to the accuracy of those construct labels. The same point holds for external validity—the logical issues involved in knowing whether a causal relationship holds over variations in persons, settings, treatments, and outcomes are orthogonal to those involved in naming the constructs.

Third, external and construct validity differ in that we may be wrong about one and right about the other. Imagine two sets of units for which we have well-justified construct labels, say, males versus females or U.S. cities versus Canadian cities or self-report measures versus observer ratings. In these cases, the construct validity of those labels is not at issue. Imagine further that we have done an experiment with one of the two sets, say, using only self-report measures. The fact that we correctly know the label for the other set that we did not use (observer ratings) rarely makes it easier for us to answer the external validity question of whether the causal effect on self-reported outcomes would be the same as for observer-rated outcomes (the exception being those rare cases in which strong theory exists to help make the prediction). And the converse is also true: that I may have the labels for these two sets of units incorrect, but if I have done the same experiment with both sets of units, I still can provide helpful answers to external validity questions about whether the effect holds over the two kinds of outcomes despite using the wrong labels for them.

Finally, external and construct validity differ in the methods emphasized to improve them. Construct validity relies more on clear construct explication and on good assessment of study particulars, so that the match between construct and particulars can be judged. External validity relies more on tests of changes in the size and direction of a causal relationship. Of course, those tests cannot be done without some assessments; but that is also true of statistical conclusion and internal validity, both of which depend in practice on having assessments to work with.

The Relationship Between Internal Validity and Construct Validity

Both internal and construct validity share in common the notion of confounds. The relationship between internal validity and construct validity is best illustrated by the four threats listed under internal validity in Cook and Campbell (1979) that are now listed under construct validity: resentful demoralization, compensatory equalization, compensatory rivalry, and treatment diffusion. The problem of whether these threats should count under internal or construct validity hinges on exactly what kinds of confounds they are. Internal validity confounds are forces that could have occurred in the absence of the treatment and could have caused some or all of the outcome observed. By contrast, these four threats would not have occurred had a treatment not been introduced; indeed, they occurred because the treatment was introduced, and so are part of the treatment condition (or perhaps more exactly, part of the treatment contrast). They threaten construct validity to the extent that they are usually not part of the intended conceptual structure of the treatment, and so are often omitted from the description of the treatment construct.

Tradeoffs and Priorities

In the last two chapters, we have presented a daunting list of threats to the validity of generalized causal inferences. This might lead the reader to wonder if any single experiment can successfully avoid all of them. The answer is no. We cannot reasonably expect one study to deal with all of them simultaneously, primarily because of logistical and practical tradeoffs among them that we describe in this section. Rather, the threats to validity are heuristic devices that are intended to raise consciousness about priorities and tradeoffs, not to be a source of skepticism or despair. Some are much more important than others in terms of both prevalence and consequences for quality of inference, and experience helps the researcher to identify those that are more prevalent and important for any given context. It *is* more realistic to expect a program of research to deal with most or all of these threats over time. Knowledge growth is more cumulative than episodic, both with experiments and with other types of research. However, we do not mean all this to say that single experiments are useless or all equally full of uncertainty in the results. A good experiment does not have to deal with all threats but only with the subset of threats that a particular field considers most serious at the time. Nor is dealing with threats the only mark of a good experiment; for example, the best experiments influence a field by testing truly novel ideas (Eysenck & Eysenck, 1983; Harré, 1981).

In a world of limited resources, researchers always make tradeoffs among validity types in any single study. For example, if a researcher increases sample size in order to improve statistical conclusion validity, he or she is reducing resources that could be used to prevent treatment-correlated attrition and so improve internal validity. Similarly, random assignment can help greatly in improving internal validity, but the organizations willing to tolerate this are probably less representative than organizations willing to tolerate passive measurement, so external validity may be compromised. Also, increasing the construct validity of effects by operationalizing each of them in multiple ways is likely to increase the response burden and so cause attrition from the experiment; or, if the measurement budget is fixed, then increasing the number of measures may lower reliability for individual measures that must then be shorter.

Such countervailing relationships suggest how crucial it is in planning any experiment to be explicit about the priority ordering among validity types. Unnecessary tradeoffs between one kind of validity and another have to be avoided, and the loss entailed by necessary tradeoffs has to be estimated and minimized. Scholars differ in their estimate of which tradeoffs are more desirable. Cronbach (1982) maintains that timely, representative, but less rigorous studies can lead to reasonable causal inferences that have greater external validity, even if the studies are nonexperimental. Campbell and Boruch (1975), on the other hand, maintain that causal inference is problematic outside of experiments because many threats to internal validity remain unexamined or must be ruled out by fiat rather than through direct design or measurement. This is an example of the major and most discussed tradeoff—that between internal and external validity.

Internal Validity: A Sine Qua Non?

Noting that internal validity and external validity often conflict in any given experiment, Campbell and Stanley (1963) said that "*internal validity* is the *sine qua non*" (p. 5).[8] This one statement gave internal validity priority for a generation of field experimenters. Eventually, Cronbach took issue with this priority, claiming that internal validity is "trivial, past-tense, and local" (1982, p. 137), whereas external validity is more important because it is forward looking and asks general questions. Because Cronbach was not alone in his concerns about the original validity typology, we discuss here the priorities among internal validity and other validities, particularly external validity.

Campbell and Stanley's (1963) assertion that internal validity is the sine qua non of experimentation is one of the most quoted lines in research methodology. It appeared in a book on experimental and quasi-experimental design, and the text makes clear that the remark was meant to apply *only* to experiments, not to other forms of research:

> *Internal validity* is the basic minimum without which any experiment is uninterpretable: Did in fact the experimental treatments make a difference in this specific experimental instance? *External validity* asks the question of *generalizability:* To what populations, settings, treatment variables, and measurement variables can this effect be generalized? Both types of criteria are obviously important, even though they are frequently at odds in that features increasing one may jeopardize the other. While *internal validity* is the *sine qua non,* and while the question of *external validity,* like the question of inductive inference, is never completely answerable, the selection of designs strong in both types of validity is obviously our ideal. This is particularly the case for research on teaching, in which generalization to applied settings of known character is the desideratum. (Campbell & Stanley, 1963, p. 5)

Thus Campbell and Stanley claimed that internal validity was necessary for experimental and quasi-experimental designs probing causal hypotheses, not for research generally. Moreover, the final sentence of this quote is almost always overlooked. Yet it states that external validity is a *desideratum* (purpose, objective, requirement, aim, goal) in educational research. This is nearly as strong a claim as the sine qua non claim about internal validity.

As Cook and Campbell (1979) further clarified, the sine qua non statement is, to a certain degree, a tautology:

> There is also a circular justification for the primacy of internal validity that pertains in any book dealing with experiments. The unique purpose of experiments is to provide stronger tests of *causal* hypotheses than is permitted by other forms of research, most of which were developed for other purposes. For instance, surveys were developed to describe population attitudes and reported behaviors while participant observation

[8] *Sine qua non* is Latin for "without which not" and describes something that is essential or necessary. So this phrase describes internal validity as necessary.

> methods were developed to describe and generate new hypotheses about ongoing behaviors *in situ*. Given that the unique original purpose of experiments is cause-related, internal validity has to assume a special importance in experimentation since it is concerned with how confident one can be that an observed relationship between variables is *causal* or that the absence of a relationship implies *no cause*. (p. 84)

Despite all these disclaimers, many readers still misinterpret our position on internal validity. To discourage such misinterpretation, let us be clear: *Internal validity is not the sine qua non of all research. It does have a special (but not inviolate) place in cause-probing research, and especially in experimental research, by encouraging critical thinking about descriptive causal claims.* Next we examine some issues that must be examined before knowing exactly how high a priority internal validity should be.

Is Descriptive Causation a Priority?

Internal validity can have high priority only if a researcher is self-consciously interested in a descriptive causal question from among the many competing questions on a topic that might be asked. Such competing questions could be about how the problem is formulated, what needs the treatment might address, how well a treatment is implemented, how best to measure something, how mediating causal processes should be understood, how meanings should be attached to findings, and how costs and fiscal benefits should be measured. Experiments rarely provide helpful information about these questions, for which other methods are to be preferred. Even when descriptive causation is a high priority, these other questions might also need to be answered, all within the same resource constraints. Then a method such as a survey might be preferred because it has a wider **bandwidth**[9] that permits answering a broader array of questions even if the causal question is answered less well than it would be with an experiment (Cronbach, 1982). The decision to prioritize on descriptive causal questions or some alternative goes far beyond the scope of this book (Shadish, Cook, & Leviton, 1991). Our presumption is that the researcher has already justified such a question before he or she begins work within the experimental framework being elaborated in this book.

Can Nonexperimental Methods Give a Satisfactory Answer?

Even if a descriptive causal inference has been well justified as a high priority, experimental methods are still not the only choice. Descriptive causal questions can be studied nonexperimentally. This happens with correlational path analysis in sociology (e.g., Wright, 1921, 1934), with case-control studies in epidemiology (e.g.,

[9] Cronbach's analogy is to radios that can have high bandwidth or high fidelity, there being a tradeoff between the two. **Bandwidth** means a method can answer many questions but with less accuracy, and fidelity describes methods that answer one or a few questions but with more accuracy.

Ahlbom & Norell, 1990), or with qualitative methods such as case studies (e.g., Campbell, 1975). The decision to investigate a descriptive causal question using such methods depends on many factors. Partly these reflect disciplinary traditions that developed for either good or poor reasons. Some phenomena are simply not amenable to the manipulation that experimental work requires, and at other times manipulation may be undesirable for ethical reasons or for fear of changing the phenomenon being studied in undesirable ways. Sometimes the cause of interest is not yet sufficiently clear, so that interest is more in exploring a range of possible causes than in zeroing in on one or two of them. Sometimes the investment of time and resources that experiments may require is premature, perhaps because insufficient pilot work has been done to develop a treatment in terms of its theoretical fidelity and practical implementability, because crucial aspects of experimental procedures such as outcome measurement are underdeveloped, or because results are needed more quickly than an experiment can provide. Premature experimental work is a common research sin.

However, the nature of nonexperimental methods can often prevent them from making internal validity the highest priority. The reason is that experimental methods match the requirements of causal reasoning more closely than do other methods, particularly in ensuring that cause precedes effect, that there is a credible source of counterfactual inference, and that the number of plausible alternative explanations is reduced. In their favor, however, the data generally used with nonexperimental causal methods often entail more representative samples of constructs than in an experiment and a broader sampling scheme that facilitates external validity. So nonexperimental methods will usually be less able to facilitate internal validity but equally or more able to promote external or construct validity. But these tendencies are not universal. Nonexperimental methods can sometimes yield descriptive causal inferences that are fully as plausible as those yielded by experiments, as in some epidemiological studies. As we said at the start of Chapter 2, validity is an attribute of knowledge claims, not methods. Internal validity depends on meeting the demands of causal reasoning rather than on using a particular method. No method, including the experiment, guarantees an internally valid causal inference, even if the experiment is often superior.

The Weak and Strong Senses of Sine Qua Non

However, suppose the researcher has worked through all these matters and has decided to use an experiment to study a descriptive causal inference. Then internal validity can be a sine qua non in two senses. The weak sense is the tautological one from Campbell and Stanley (1963): "*internal validity* is the basic minimum without which any experiment is uninterpretable" (p. 5). That is, to do an experiment and have no interest in internal validity is an oxymoron. Doing an experiment makes sense only if the researcher has an interest in a descriptive causal question, and to have this interest without a concomitant interest in the validity of the causal answer seems hard to justify.

The strong sense in which internal validity can have priority occurs when the experimenter can exercise choice within an experiment about how much priority to give to each validity type. Unfortunately, any attempt to answer this question is complicated by the fact that we have no accepted measures of the amount of each kind of validity, and so it is difficult to tell how much of each validity is present. One option would be to use methodological indices, for example, claiming that randomized studies with low attrition yield inferences that are likely to be higher in internal validity. But such an index fails to measure the internal validity of other cause-probing studies. Another option would be to use measures based on the number of identified threats to validity that still remain to be ruled out. But the conceptual obstacles to such measures are daunting; and even if it were possible to construct them for all the validity types, we can think of no way to put them on the common metric that would be needed for making comparative priorities.

A feasible option is to use the amount of resources devoted to a particular validity type as an indirect index of its priority. After all, it is possible to reduce the resources given, say, to fostering internal validity and to redistribute them to fostering some other validity type. For example, a researcher might take resources that would otherwise be devoted to random assignment, to measuring selection bias, or to reducing attrition and use them either (1) to study a larger number of units (in order to facilitate statistical conclusion validity), (2) to implement several quasi-experiments on existing treatments at a larger number of representatively sampled sites (in order to facilitate external validity), or (3) to increase the quality of outcome measurement (in order to facilitate construct validity). Such resource allocations effectively reduce the priority of internal validity.

These allocation decisions vary as a function of many variables. One is the basic versus applied research distinction. Basic researchers have high interest in construct validity because of the key role that constructs play in theory construction and testing. Applied researchers tend to have more interest in external validity because of the particular value that accrues to knowledge about the reach of a causal relationship in applied contexts. For example, Festinger's (e.g., 1953) basic social psychology experiments were justly famous for the care they put into ensuring that the variable being manipulated was indeed cognitive dissonance. Similarly, regarding units, Piagetian developmental psychologists often devote extra resources to assessing whether children are at preoperational or concrete operational stages of development. By contrast, the construct validity of settings tends to be of less concern in basic research because few theories specify crucial target settings. Finally, external validity is frequently of the lowest interest in basic research. Much basic psychological research is conducted using college sophomores for the greater statistical power that comes through having large numbers of homogeneous respondent populations. The tradeoff is defended by the hope that the results achieved with such students will be general because they tap into general psychological processes—an assumption that needs frequent empirical testing. However, assuming (as we are) that these examples occurred in the context of an experiment, it is still unlikely that the basic researcher would let the resources given to internal validity fall below a minimally acceptable level.

By contrast, much applied experimentation has different priorities. Applied experiments are often concerned with testing whether a particular problem is alleviated by an intervention, so many readers are concerned with the construct validity of effects. Consider, for example, debates about which cost-of-living adjustment based on the Consumer Price Index (CPI) most accurately reflects the actual rise in living costs—or indeed, whether the CPI should be considered a cost-of-living measure at all. Similarly, psychotherapy researchers have debated whether traditional therapy outcome measures accurately reflect the notion of clinically significant improvement among therapy clients (Jacobson, Follette, & Revenstorf, 1984). Applied research also has great stake in plausible generalization to the specific external validity targets in which the applied community is interested. Weisz, Weiss, and Donenberg (1992), for example, suggested that most psychotherapy experiments were done with units, treatments, observations, and settings that are so far removed from those used in clinical practice as to jeopardize external validity inferences about how well psychotherapy works in those contexts.

It is clear from these examples that decisions about the relative priority of different validities in a given experiment cannot be made in a vacuum. They must take into account the status of knowledge in the relevant research literature generally. For example, in the "phase model" of cancer research at the National Institutes of Health (Greenwald & Cullen, 1984), causal inferences about treatment effects are always an issue, but at different phases different validity types have priority. Early on, the search for possibly effective treatments tolerates weaker experimental designs and allows for many false positives so as not to overlook a potentially effective treatment. As more knowledge accrues, internal validity gets higher priority to sort out those treatments that really do work under at least some ideal circumstances (efficacy studies). By the last phase of research, external validity is the priority, especially exploring how well the treatment works under conditions of actual application (effectiveness studies).

Relatively few programs of research are this systematic. However, one might view the four validity types as a loose guide to programmatic experimentation, instructing the researcher to iterate back and forth among them as comparative weaknesses in generalized causal knowledge of one kind or another become apparent. For example, many researchers start a program of research by noticing an interesting relationship between two variables (e.g., McGuire, 1997). They may do further studies to confirm the size and dependability of the relationship (statistical conclusion validity), then study whether the relationship is causal (internal validity), then try to characterize it more precisely (construct validity) and to specify its boundaries (external validity). Sometimes, the phenomenon that piques an experimenter's curiosity already has considerable external validity; for instance, the covariation between smoking and lung cancer across different kinds of people in different settings and at different times led to a program of research designed to determine if the relationship was causal, to characterize its size and dependability, and then to explain it. Other times, the construct validity of the variables has already been subject to much attention, but the question of a causal relationship between them suddenly attracts notice. For instance, the construct validity of both race and intelligence had already been extensively studied when a controversy arose in the 1990s over the possibility of a causal relationship between them (Devlin, 1997; Herrnstein & Murray, 1994). Programs of experimental research can start at many different points, with existing knowledge lending strength to different kinds of inferences and with need to repair knowledge weaknesses of many different kinds. Across a program of research, all validity types are a high priority. By the end of a program of research, each validity type should have had its turn in the spotlight.

SUMMARY

In Chapters 2 and 3, we have explicated the theory of validity that drives the rest of this book. It is a heavily pragmatic theory, rooted as much or more in the needs and experiences of experimental practice as in any particular philosophy of science. Chapter 2 presented a validity typology consisting of statistical conclusion validity, internal validity, construct validity, and external validity that retains the central ideas of Campbell and Stanley (1963) and Cook and Campbell (1979) but that does so in slightly expanded terms that extend the logic of generalizations to more parts of the experiment. With a few minor exceptions, the threats to validity outlined in previous volumes remain largely unchanged in this book.

However, the presentation to this point has been abstract, as any such theory must partly be. If the theory is to retain the pragmatic utility that it achieved in the past, we have to show how this theory is used to design and criticize cause-probing studies. We begin doing so in the next chapter, in which we start with the simplest quasi-experimental designs that have sometimes been used for investigating causal relationships, showing how each can be analyzed in terms of threats to validity and how those threats can be better diagnosed or sometimes reduced in plausibility by improving those designs in various ways. Each subsequent chapter in the book presents a new class of designs, each of which is in turn subject to a similar validity analysis—quasi-experimental designs with comparison groups and pretests, interrupted time series designs, regression discontinuity designs, and randomized designs. In all these chapters, the focus is primarily but not exclusively on internal validity. Finally, following the presentation of these designs, the emphasis is reversed as the book moves to a discussion of methods and designs for improving construct and external validity.

8

Randomized Experiments: Rationale, Designs, and Conditions Conducive to Doing Them

Ran·dom (răn′dəm): [From at random, by chance, at great speed, from Middle English *randon,* speed, violence, from Old French from *randir,* to *run,* of Germanic origin.] adj. 1. Having no specific pattern, purpose, or objective: *random movements; a random choice.* See Synonyms at chance. 2. Statistics. Of or relating to the same or equal chances or probability of occurrence for each member of a group.

DOES EARLY preschool intervention with disadvantaged children improve their later life? The Perry Preschool Program experiment, begun in 1962, studied this question with 128 low-income African-American children who were randomly assigned to receive either a structured preschool program or no treatment. Ninety-five percent of participants were followed to age 27, and it was found that treatment group participants did significantly better than controls in employment, high school graduation, arrest records, home ownership, welfare receipt, and earnings (Schweinhart, Barnes, & Weikart, 1993), although early IQ and academic aptitude gains were not maintained into early adulthood. Along with other experimental evidence on the effects of preschool interventions (e.g., Olds et al., 1997; Olds et al., 1998), these results helped marshal continued political support and funding for programs such as Head Start in the United States. In this chapter, we present the basic logic and design of randomized experiments such as this one, and we analyze the conditions under which it is less difficult to implement them outside the laboratory.

In the natural sciences, scientists introduce an intervention under circumstances in which no other variables are confounded with its introduction. They then look to see how things change—for instance, whether an increase in heat affects the pressure of a gas. To study this, a scientist might place the gas in a fixed enclosure, measure the pressure, heat the gas, and then measure the pressure again to see if it has changed. The gas is placed in the enclosure to isolate it from anything else that would affect the pressure inside. But even in this simple example, the intervention is still a molar treatment package that is difficult to explicate fully. The enclosure is made of a certain material, the heat comes from a certain kind of burner, the humidity is at a certain level, and so forth. Full control and full isolation of the "intended" treatment are difficult, even in the natural sciences.

In much social research, more formidable control problems make successful experimentation even more difficult. For example, it is impossible to isolate a person from her family in order to "remove" the influences of family. Even in agricultural tests of a new seed, the plot on which those seeds are planted cannot be isolated from its drainage or soil. So many scientists rely on an approach to experimental control that is different from physical isolation—random assignment. The randomized experiment has its primary systematic roots in the agricultural work of statistician R. A. Fisher (1925, 1926, 1935; see Cowles, 1989, for a history of Fisher's work). Randomization was sometimes used earlier (e.g., Dehue, 2000; Gosnell, 1927; Hacking, 1988; Hrobjartsson, Gotzche, & Gluud, 1998; McCall, 1923; Peirce & Jastrow, 1884; Richet, 1884; Stigler, 1986). But Fisher explicated the statistical rationale and analyses that tie causal inference to the physical randomization of units to conditions in an experiment (Fisher, 1999).

THE THEORY OF RANDOM ASSIGNMENT

Random assignment reduces the plausibility of alternative explanations for observed effects. In this, it is like other design features such as pretests, cohorts, or nonequivalent dependent variables. But random assignment is distinguished from those features by one very special characteristic shared only with the regression discontinuity design: it can yield unbiased estimates of the average treatment effect (Rosenbaum, 1995a).[1] Moreover, it does this with greater efficiency than the

[1] Three observations about the phrase "unbiased estimates of average treatment effects" are worth noting. First, some statisticians would prefer to describe the advantage of randomization as yielding a consistent estimator (one that converges on its population parameter as sample size increases), especially because we never have the infinite number of samples suggested by the theory of expectations discussed shortly in this chapter. We use the term *unbiased* in this book primarily because it will be more intuitively understood by nonstatistical readers and because it fits better with the qualitative logic of bias control that undergirds our validity typology. Second, in a random sampling model, sample means are always unbiased estimates of population means, so differences between sample means are always unbiased estimates of differences between population means. The latter estimates can be obtained without using random assignment. But such estimates are not the same as unbiased estimates of treatment effects. It is the latter that random assignment facilitates; hence its ability to facilitate the causal inference that we refer to with the shorthand phrase "unbiased estimates of treatment effects." Third, the phrase correctly refers to the average effect over units in the study, as distinguished from the effects on each unit in the study, which is not tested in a randomized experiment.

regression discontinuity design in a greater diversity of applications. Because unbiased and efficient causal inference is a goal of experimental research, it is crucial that researchers understand what random assignment is and how it works.

What Is Random Assignment?

Random assignment is achieved by any procedure that assigns units to conditions based only on chance, in which each unit has a nonzero probability of being assigned to a condition. A well-known random assignment procedure is a coin toss. On any given toss, a fair coin has a known (50%) chance of coming up heads. In an experiment with two conditions, if heads comes up for any unit, then that unit goes into the treatment condition; but if tails comes up, then it becomes a control unit. Another random assignment procedure is the roll of a fair die that has the numbers 1 through 6 on its sides. Any number from 1 to 6 has a known (1/6) chance of coming up, but exactly which number comes up on a roll is entirely up to chance. Later we recommend more formal randomization procedures, such as the use of tables of random numbers. But coin tosses and dice rolls are well-known and intuitively plausible introductions to randomization.

Random *assignment* is not random *sampling.* We draw random samples of units from a population by chance in public opinion polls when we ask random samples of people about their opinions. Random sampling ensures that answers from the sample approximate what we would have gotten had we asked everyone in the population. Random assignment, by contrast, facilitates causal inference by making samples randomly similar to *each other,* whereas random sampling makes a sample similar to *a population.* The two procedures share the idea of "randomness," but the purposes of this randomness are quite different.

Why Randomization Works

The literature contains several complementary statistical and conceptual explanations for why and how random assignment facilitates causal inference:

- It ensures that alternative causes are not confounded with a unit's treatment condition.
- It reduces the plausibility of threats to validity by distributing them randomly over conditions.
- It equates groups on the expected value of all variables at pretest, measured or not.
- It allows the researcher to know and model the selection process correctly.
- It allows computation of a valid estimate of error variance that is also orthogonal to treatment.

These seemingly different explanations are actually closely related. None of them by itself completely captures what random assignment does, but each sheds light on part of the explanation.

Random Assignment and Threats to Internal Validity

If treatment groups could be equated before treatment, and if they were different after treatment, then pretest selection differences could not be a cause of observed posttest differences. Given equal groups at pretest, the control group posttest serves as a source of counterfactual inference for the treatment group posttest, within limits we elaborate later. Note that the logic of causal inference is at work here. The temporal structure of the experiment ensures that cause precedes effect. Whether cause covaries with effect is easily checked in the data within known probabilities. The remaining task is to show that most alternative explanations of the cause-effect relationship are implausible. The randomized experiment does so by distributing these threats randomly over conditions. So treatment units will tend to have the same average characteristics as those not receiving treatment. The only systematic difference between conditions is treatment.

For example, consider a study of the effects of psychotherapy on stress. Stress has many alternative causes, such as illness, marital conflict, job loss, arguments with colleagues, and the death of a parent. Even positive events, such as getting a new job or getting married, cause stress. The experimenter must ensure that none of these alternative causes is confounded with receiving psychotherapy, because then one could not tell whether it was psychotherapy or one of the confounds that caused any differences at posttest. Random assignment ensures that every client who receives psychotherapy is equally likely as every client in the control group to have experienced, say, a new job or a recent divorce. Random assignment does not prevent these alternative causes (e.g., divorce) from occurring; nor does it isolate the units from the occurrence of such events. People in a randomized experiment still get divorces and new jobs. Random assignment simply ensures that such events are no more likely to happen to treatment clients than to control clients. As a result, if psychotherapy clients report less stress than control clients at posttest, the cause of that difference is unlikely to be that one group had more new jobs or divorces, because such stressors are equally likely in both groups. The only systematic difference left to explain the result is the treatment.

The only internal validity threat that randomization prevents from occurring is selection bias, which it rules out by definition, because selection bias implies that a systematically biased method was used for selecting units into groups but chance can have no such systematic bias. As for the other internal validity threats, randomization does not prevent units from maturing or regressing; nor does it prevent events other than treatment from occurring after the study begins (i.e., history). Pretests can still cause a testing effect, and changes in instrumentation can still occur. Random assignment simply reduces the likelihood that these threats are confounded with treatment.

Equating Groups on Expectation

In statistics, the preceeding explanation is often summarized by saying that random assignment equates groups on expectation at pretest. What does this mean? First, it does not mean that random assignment equates units on *observed* pretest scores. Howard, Krause, and Orlinsky (1986) remind us that when a deck of 52 playing cards is well shuffled, some players will still be dealt a better set of cards than others. This is called the luck of the draw by card players (and sampling error by statisticians). In card games, we do not expect every player to receive equally good cards for each hand, but we do expect the cards to be equal in the long run over many hands. All this is true of the randomized experiment. In any given experiment, observed pretest means will differ due to luck of the draw when some conditions are dealt a better set of participants than others. But we can expect that participants will be equal over conditions in the long run over many randomized experiments.

Technically, then, random assignment equates groups on the *expectation* of group means at pretest—that is, on the mean of the distribution of all possible sample means resulting from all possible random assignments of units to conditions. Imagine that a researcher randomly assigned units to treatment and control conditions in one study and then computed a sample mean on some variable for both conditions. These two means will almost certainly be different due to sampling error—the luck of the draw. But suppose the researcher repeated this process a second time, recorded the result, and continued to do this a very large number of times. At the end, the researcher would have a distribution of means for the treatment group over the samplings achieved and also one for the control group. Some of the treatment group means would be larger than others; the same would be true for the control group. But the average of all the means for the treatment group would be the same as the average of all the means for the control group. Thus the expectation to which the definition of random assignment is linked involves the mean of all possible means, not the particular means achieved in a single study.

When random differences do exist in observed pretest means, those differences will influence the results of the study. For example, if clients assigned to psychotherapy start off more depressed than those assigned to the control group despite random assignment, and if psychotherapy reduces depression, posttest depression scores might still be equal in both treatment and control groups because of the pretest group differences. Posttest differences between treatment and control groups then might suggest no treatment effect when treatment did, in fact, have an effect that was masked by sampling error in random assignment. More generally, the results of any individual randomized experiment will differ somewhat from the population effects by virtue of these chance pretest differences. Thus summaries of results from multiple randomized experiments on the same topic (as in psychotherapy meta-analysis) can yield more accurate estimates of treatment effects than any individual study. Even so, we still say that the estimate from an individual study is unbiased. Unbiased simply means that any differences between the observed effects and the population effect are the result of chance; it does not mean that the results of the individual study are identical to the "true" population effect.

The preceding explanation uses pretest means to illustrate how randomization works. However, this is merely a teaching device, and the use of actual measured pretests is irrelevant to the logic. Randomization equates groups on expectations of *every variable before treatment, whether observed or not.* In practice, of course, pretests are very useful because they allow better diagnosis of and adjustment for attrition, they facilitate the use of statistical techniques that increase statistical power, and they can be used to examine whether treatment is equally effective at different levels of the pretest.

Additional Statistical Explanations of How Random Assignment Works

Randomization ensures that confounding variables are unlikely to be *correlated* with the treatment condition a unit receives. That is, whether a coin toss comes up heads or tails is unrelated to whether you are divorced, nervous, old, male, or anything else. Consequently, we can predict that the pretest correlation between treatment assignment and potential confounding variables should not be significantly different from zero.

This zero correlation is very useful statistically. To understand this requires a digression into how to estimate treatment effects in linear models. Let us distinguish between the *study* and the *analysis* of the study. In a *study* of the effects of psychotherapy, stress is the dependent variable (Y_i), psychotherapy is the independent variable (Z_i), and potential confounds are contained in an error term (e_i). In the *analysis* of that study, the effects of treatment are estimated from the linear model:

$$Y_i = \mu + \hat{\beta} Z_i + e_i \tag{8.1}$$

where μ is a constant, $\hat{\beta}$ is a regression coefficient, and the subscript i ranges from 1 to n, where n is the number of units in the study. Thus Y_i is the score of the ith unit on a measure of stress, Z_i is scored as 1 if the unit is in psychotherapy and 0 if not, and e_i consists of all potential confounding variables. In the analysis, if $\hat{\beta}$ is significantly different from zero, then psychotherapy had a significant effect on stress, and $\hat{\beta}$ measures the magnitude and direction of that effect.

For all this to work properly, however, the model that is specified in the analysis must match the reality of the study. Failure to achieve this match is called **specification error**—an incorrect specification of the model presumed to give rise to the data. Specifically, the statistical techniques used to estimate models such as equation (8.1) choose values of $\hat{\beta}$ so that correlations between the resulting errors and the predictor variables are zero (Reichardt & Gollob, 1986). The statistics do this whether or not that correlation really was zero in the study. Fortunately, random assignment assures that the correlation in the study will be zero for reasons outlined in the previous section; so the study matches the analysis. However, in nonrandomized studies, many confounds are probably correlated with receipt of treatment, but the computer program still chooses $\hat{\beta}$ so that the error is minimally correlated with the predictors in the data analysis, yielding a mismatch between the study and the analysis. The result is an incorrect estimate of treatment effects.[2]

A related way of thinking about the benefit of randomization is that it provides a valid estimate of error variance (e.g., Keppel, 1991; R. Kirk, 1982). Two possible causes of total variation in outcome (i.e., of how much people differ from each other in stress levels) exist—variation caused by treatment conditions (e.g., whether the person received psychotherapy) and variation caused by other factors (e.g., all the other causes of stress). Random assignment allows us to separate out these two sources of variability. Error variation is estimated as the amount of variation among units within each condition. For example, for those clients who were assigned to psychotherapy, variation in whether or not they received psychotherapy cannot contribute to their different stress levels because there was no such variation—they all got psychotherapy. So any variance in outcome among people randomly assigned to psychotherapy must be caused only by confounds. The average of each of these computed error terms from within each condition serves as our best estimate of error. This error term is the baseline against which we see if differences *between* treatment conditions exceed the differences that normally occur among units as a function of all the other causes of the outcome.

Summary

Random assignment facilitates causal inference in many ways—by equating groups before treatment begins, by making alternative explanations implausible, by creating error terms that are uncorrelated with treatment variables, and by allowing valid estimates of error terms. These are interrelated explanations. For example, groups that are equated before treatment begins allow fewer alternative explanations if differences later emerge, and uncorrelated errors are necessary to estimate the size of the error term. But randomization is not the only way to accomplish these things. Alternative explanations can sometimes be made implausible through logical means, as is typically the aim with quasi-experimentation; and uncorrelated errors can be created with other forms of controlled assignment to conditions, as with the regression discontinuity design. But randomization is the only design feature that accomplishes all of these goals at once, and it does so more reliably and with better known properties than any alternatives.

[2] One way to think about the selection bias models in Chapter 5 is that they try to make the error terms orthogonal to the predictors in a statistically acceptable way, but this is hard, so they often fail; and one way to think about the regression discontinuity design is that it is able to make this correlation zero for reasons outlined in the Appendix to Chapter 7.

Random Assignment and Units of Randomization

We have frequently used the word "unit" to describe whatever or whoever is assigned to experimental conditions. A unit is simply "an opportunity to apply or withhold the treatment" (Rosenbaum, 1995a, p. 17).

Kinds of Units

In much field experimentation, the units being assigned to conditions are people—clients in psychotherapy, patients in cancer trials, or students in educational studies. But units can be other kinds of entities (Boruch & Foley, 2000). R. A. Fisher (1925) assigned plots of land randomly to different levels of fertilizer or different strains of seed. In psychological and medical research, animals are often randomly assigned to conditions. Researchers in the New Jersey Negative Income Tax experiment (Rees, 1974) randomly assigned families to conditions. Gosnell (1927) randomly assigned neighborhoods to conditions. Edgington (1987) discussed single-participant designs in which treatment times were randomly assigned. Schools have been randomly assigned (Cook et al., 1998; Cook, Hunt & Murphy, 2000). Nor is randomization useful just in the social sciences. Wilson (1952) describes a study in which the steel plates used in gauges were randomized prior to testing different explosives, so that variations in the strength of the plates would not be systematically associated with any one explosive. The possibilities are endless.

Higher Order Units

Units such as families, work sites, classrooms, psychotherapy groups, hospital wards, neighborhoods, or communities are aggregates of individual units such as family members, employees, students, clients, patients, neighbors, or residents. Studies of the effects of treatments on such higher order units are common, and a literature has developed specific to experiments on higher order units (e.g., Donner & Klar, 2000; Gail, Mark, Carroll, Green, & Pee, 1996; Moerbeek, van Breukelen, & Berger, 2000; Murray, 1998; Sorensen, Emmons, Hunt, & Johnston, 1998). For example, the National Home Health Agency Prospective Payment Demonstration experiment assigned 142 home health agencies to different Medicare payment options to see how use of care was affected (Goldberg, 1997); the San Diego Nursing Home Incentive Reimbursement Experiment assigned 36 nursing homes to different Medicare reimbursement options (Jones & Meiners, 1986); the Tennessee Class Size Experiment randomly assigned 347 classes to large or small numbers of students (Finn & Achilles, 1990); and Kelly et al. (1997) randomly assigned eight cities to two conditions to study an HIV prevention intervention. The higher order unit need not be a naturally occurring entity such as a work site or a neighborhood. The researcher can create the higher order unit solely for the research, as in the case of a stop-smoking program that is administered in small groups so that participants can benefit from mutual support. Nor is it necessary that the individual units know or interact with each other. For instance, when physicians' practices are randomized to conditions, the physician's practice is a higher order unit even though the majority of the physician's patients will never meet. Finally, sometimes a treatment cannot be restricted to particular individuals by its very nature. For example, when a radio-based driving safety campaign is broadcast over a listening area, the entire area receives treatment, even if only some individual drivers are formally included in the research (Reicken et al., 1974).

There are often good practical and scientific reasons to use aggregate units. In a factory experiment, it may not be practical to isolate each worker and give him or her a unique treatment, for resentful demoralization or diffusion of treatment might result. Similarly, in the first evaluation of "Plaza Sesamo," Diaz-Guerro and Holtzmann (1974) randomly assigned some individual children in Mexican day care centers to watch "Plaza Sesamo" in small groups. They were in a special room with two adult monitors who focused attention on the show. At the same time, other children watched cartoons in larger groups in the regular room with no special monitors. Because treating classmates in these different ways may have led to a focused inequity, it would have been desirable if the experimenters' resources had permitted them to assign entire classes to treatments.

The research question also determines at which level of aggregation units should be randomized. If effects on individuals are at issue, the individual should be the unit, if possible. But if school or neighborhood phenomena are involved or if the intervention is necessarily performed on an aggregate, then the unit of randomization should not be at a lower level of aggregation.[3] Thus, if one is investigating whether frequent police car patrols deter crime in a neighborhood, different amounts of patrolling should be assigned to neighborhoods and not, say, to blocks within neighborhoods.

In aggregate units, the individual units within aggregates may no longer be independent of each other because they are exposed to common influences besides treatment. For example, students within classrooms talk to each other, have the same teacher, and may all receive treatment at the same time of day. These dependencies lead to what used to be called the **unit of analysis** problem (Koepke & Flay, 1989) but what is more recently discussed as multilevel models or hierarchical linear models. Because this book focuses on design rather than analysis, we do not treat the analytic issues in detail (Feldman, McKinlay, & Niknian, 1996; Gail et al., 1996; Green et al., 1995; Murray, 1998; Murray et al., 1994; Murray, Moskowitz, & Dent, 1996). But from a design perspective, using higher order units raises several issues.

[3] The nesting of participants in higher order units can still pose problems even when individuals are assigned to treatment. For example, if individual cancer patients who each have multiple tumors are randomly assigned to treatment but treatment is administered separately to each tumor and tumor response is observed separately for each tumor, those responses are not independent (Sargent, Sloan, & Cha, 1999).

Studies that use higher order units frequently have fewer such units available to randomize. Consider the limiting case in which students in one classroom are given the treatment and those in a second classroom serve as controls. Treatment conditions are then totally confounded with classrooms, making it impossible to tell if performance differences at posttest are due to differences in treatment or in classroom characteristics, such as the charisma of the teacher, the mix of students, or the physical conditions of the class. When more than one, but still few, higher order units are assigned to conditions, randomization may result in very different means, variances, and sample sizes across conditions. Such cases are surprisingly common in the literature (e.g., Simpson, Klar, & Dohner, 1995); but they incur substantial problems for internal and statistical conclusion validity (Varnell, Murray, & Baker, in press). Such problems occur most often with studies of schools and communities, because it is expensive to add new sites. Random assignment of higher order units from within blocks or strata can reduce such problems. For example, McKay, Sinisterra, McKay, Gomez, and Lloreda (1978) studied the effects of five levels of a program of nutrition, health care, and education on the cognitive ability of chronically undernourished children in Cali, Colombia. They divided Cali into 20 relatively homogeneous neighborhood sectors. Then they rank-ordered sectors on a standardized combination of pretreatment screening scores and randomly assigned those sectors to the five conditions from blocks of five. The Kansas City Preventive Patrol experiment followed a similar procedure in its study of whether the visible presence of police patrols deterred crime (Kelling, Pate, Dieckman, & Brown, 1976). The researchers placed 15 patrol districts into blocks of three that were homogenous on demographic characteristics; they then randomly assigned districts from these blocks into the three experimental conditions.

Planning proper sample size and analysis of designs with higher order units is more complex than usual because individual units are not independent within aggregate units (Bock, 1989; Bryk & Raudenbush, 1992; Bryk, Raudenbush, & Congdon, 1996; H. Goldstein, 1987; Raudenbush, 1997; Snijders & Bosker, 1999). Given the same number of individual units, power is almost always lower in designs with higher order units than in those with individual units; and special power analyses must be used.[4] Moreover, power is improved more by increasing the number of aggregate units (e.g., adding more classrooms) than by increasing the number of individuals within units (e.g., adding more students within classrooms). Indeed, at a certain point the latter can rapidly become wasteful of resources without improving power at all, depending on the size of the dependencies within cluster (as measured by the intraclass correlation).

[4] See Donner (1992); Donner and Klar (1994); Feldman et al. (1996); Gail, Byar, Pechacek, and Corle (1992); Gail et al. (1996); Hannan and Murray (1996); Koepsell et al. (1991); Murray (1998); Murray and Hannan (1990); Murray, Hannan, and Baker (1996); Raudenbush (1997); and Raudenbush and Liu (2000). Both Orr (1999) and Raudenbush and Liu (2000) address tradeoffs between power and the cost of adding more participants within and between treatment sites.

Often resources will prevent the researcher from including the number of higher order units that the power analyses suggest is required to conduct a sensitive statistical analysis. In such situations, it helps to treat the study as if it were a quasi-experiment, adding features such as switching replications or double pretests to facilitate causal inferences. Shadish, Cook, and Houts (1986) discuss this strategy and provide illustrations. For example, in the Cali study, McKay et al. (1978) staggered the introduction of treatment across the five treatment groups, so that some received treatment for the full length of the study but others received treatment progressively later. All had a common final posttest time. Their demonstration that effects started concurrent with implementation of treatment in each group helped to bolster the study's interpretability despite its use of only four higher order units per condition. Finally, measurement of the characteristics of higher order units helps diagnose the extent to which those characteristics are confounded with treatment.

Researchers sometimes create an unnecessary unit of analysis problem when, in order to save the extra costs and logistical complexity of treating participants individually, they administer to a group a treatment that could have been administered to individuals. By doing this, the researcher may thus create dependencies among participants within groups. For example, suppose a treatment for insomnia is administered to 50 participants in 10 groups of 5 people each; and suppose further that the treatment could have been administered individually in the sense that it does not involve transindividual theoretical components such as mutual interpersonal support. Nonetheless, group members are now exposed to many common influences. For example, some of them might become romantically involved, with possible consequences for their sleep patterns! These group influences may vary from group to group and so affect outcome differentially. So researchers should administer the treatment to individual units if the research question makes this possible; if not, then group membership should be taken into account in the analysis.

The Limited Reach of Random Assignment

Though random assignment is usually better than other design features for inferring that an observed difference between treatment and control groups is due to some cause, its applicability is often limited. Random assignment is useful only if a researcher has already decided that a local molar causal inference is of most interest. Such inferences are a common goal in social research, but they are not the only goal. Yet random assignment is conceptually irrelevant to all other research goals. Further, random assignment is just one part of experimental design, and experimental design is only part of an overall research design. Experimental design involves the scheduling of observations, the choice of treatments and comparisons, the selection of observations and measures, the determination of who should be the respondents, and the manner of assigning units to treatments. Random assignment deals with only the last of these issues, so to assign at random does not guarantee a useful experimental or research design.

Thus, if a randomized experiment is conducted with units that do not correspond to the population of theoretical or policy interest, the usefulness of the research is weakened even if the quality of the causal inference is high. Rossi and Lyall (1976) criticized the New Jersey Negative Income Tax Experiment because the respondents were working poor, but most guaranteed incomes in a national scheme would go to the jobless poor. Similarly, Cook et al. (1975) criticized Ball and Bogatz (1970) for manipulating levels of social encouragement to view "Sesame Street," thus confounding viewing with encouragement. Larson (1976) criticized the Kansas City Patrol Experiment because the amount of police patrolling that was achieved in the high-patrolling condition was not even as high as the average in New York City and because the contrast between high- and low-patrol areas in Kansas City was reduced due to police squad cars crossing atypically often over the low-patrol areas with their lights flashing and sirens screaming. These are all useful criticisms of details from social experiments, though none is a criticism of random assignment itself. Such criticisms have implications for the desirability of random assignment only to the extent that implementing such assignment caused the problems to emerge. This is rarely the case.

SOME DESIGNS USED WITH RANDOM ASSIGNMENT

This section reviews many variants of randomized experimental designs (see Table 8.1; for other variations, see Fleiss, 1986; Keppel, 1991; Kirk, 1982; Winer, Brown, & Michels, 1991). The designs we present are the most commonly used in field research, providing the basic building blocks from which more complex designs can be constructed. This section uses the same design notation as in earlier chapters, except that the letter *R* indicates that the group on that line was formed by random assignment. We place *R* at the start of each line, although random assignment could occur either before or after a pretest, and the placement of *R* would vary accordingly.

The Basic Design

The basic randomized experiment requires at least two conditions, random assignment of units to conditions, and posttest assessment of units. Structurally, it can be represented as:

$$\begin{array}{ccc} R & X & O \\ R & & O \end{array}$$

TABLE 8.1 Schematic Diagrams of Randomized Designs

The Basic Randomized Design Comparing Treatment to Control					
R		X	O		
R			O		
The Basic Randomized Design Comparing Two Treatments					
R		X_A	O		
R		X_B	O		
The Basic Randomized Design Comparing Two Treatments and a Control					
R		X_A	O		
R		X_B	O		
R			O		
The Pretest-Posttest Control Group Design					
R	O	X	O		
R	O		O		
The Alternative-Treatments Design with Pretest					
R	O	X_A	O		
R	O	X_B	O		
Multiple Treatments and Controls with Pretest					
R	O	X_A	O		
R	O	X_B	O		
R	O	O			
Factorial Design					
R		X_{A1B1}	O		
R		X_{A1B2}	O		
R		X_{A2B1}	O		
R		X_{A2B2}	O		
Longitudinal Design					
R	$O \ldots O$	X	O	$O \ldots O$	
R	$O \ldots O$		O	$O \ldots O$	
A Crossover Design					
R	O	X_A	O	X_B	O
R	O	X_B	O	X_A	O

Note: For simplicity, we place the *R* (to indicate random assignment) at the front of the schematic diagram; however, assignment sometimes occurs before and sometimes after the pretest, so that the placement of *R* could be varied accordingly.

A good example of the use of this design with a single treatment and a control group is the test of the Salk polio vaccine in 1954. More than 400,000 children were randomly assigned to receive either the vaccine or a placebo (Meier, 1972).

A key issue is the nature of the control condition. Selection of a particular kind of control group depends on what one wants to control. For example, a no-treatment control condition tests the effects of a molar treatment package, including all its active and passive, important and trivial components. However, when interest is in the effects of a part of that package, the control should include everything but that part. In drug studies, for example, the researcher often wants to separate out the effects of the pharmaceutically active ingredients in the drugs from the effects of the rest of the package—things such as swallowing a pill or having contact with medical personnel. A placebo control does this, with medical personnel providing patients with, say, an inert pill in a manner that includes all the extraneous conditions except the active ingredients (Beecher, 1955).

Many types of control groups exist, for example, no-treatment controls, dose-response controls, wait-list controls, expectancy controls, or attention-only controls (Borkovec & Nau, 1972; Garber & Hollon, 1991; International Conference on Harmonization, 1999; Jacobson & Baucom, 1977; Kazdin & Wilcoxon, 1976; O'Leary & Borkovec, 1978; Orne, 1962; Seligman, 1969; Shapiro & Shapiro, 1997). The variations are limited only by the researcher's imagination. But in all cases the question is always, "Control for what?" For instance, Rossi and Lyall (1976, 1978) criticized the New Jersey Negative Income Tax in part on this basis—that the control group differed not only in failure to receive the treatment of interest but also in receiving far fewer and less intrusive administrative experiences than the treatment group.

Two Variants on the Basic Design

One variation compares two treatments by substituting X_A and X_B for the X and blank space in the previous diagram:

$$\begin{array}{ccc} R & X_A & O \\ R & X_B & O \end{array}$$

If X_A is an innovative treatment, for example, X_B is often a "gold-standard" treatment of known efficacy. The causal question is then, "What is the effect of the innovation compared with what would have happened if units had received the standard treatment?" This design works well if the standard treatment has a known track record against no-treatment controls. But if not, and if those receiving X_A are not different from those receiving X_B at posttest, the researcher cannot know if both treatments were equally effective or equally ineffective. In that case, a control group helps:

$$\begin{array}{ccc} R & X_A & O \\ R & X_B & O \\ R & & O \end{array}$$

This design was used in Boston to study the effects of an experimental housing project designed to improve the kinds of neighborhoods in which poor families lived (Katz, Kling, & Liebman, 1997; Orr, 1999). Poverty families in Treatment A received housing vouchers good for use only in low-poverty areas so that if they moved, they would move to better neighborhoods; those in Treatment B received vouchers for use anywhere, including in high-poverty areas; and those in the control group did not receive any vouchers at all.

Risks to This Design Due to Lack of Pretest

Omitting a pretest is a virtue whenever pretesting is expected to have an unwanted sensitization effect; and it is a necessity when a pretest cannot be gathered (as in some studies of cognitive development in infants), is seriously impractical (as with expensive and time-consuming interviews of patients by physicians), or is known to be a constant (as in studies of mortality in which all patients are alive at the start). Otherwise, the absence of a pretest is usually risky if there is any likelihood of attrition from the study; in fact, some observers cite the need for a pretest as one of the most important lessons to emerge from the last 20 years of social experiments (Haveman, 1987). Attrition occurs often in field experiments, leaving the researcher with the need to examine whether (1) those who dropped out of the study were different from those who remained and, especially, (2) if those who dropped out of one condition were different from those who dropped out of the other condition(s). Pretreatment information, preferably on the same dependent variable used at posttest, helps enormously in answering such questions.

Of course, attrition is not inevitable in field experiments. In medical trials of surgical procedures that have immediate outcomes, the treatments happen too quickly to allow much attrition; and patient follow-up care is often thorough enough and medical records good enough that posttests and follow-up observations on patients are available. An example is Taylor et al.'s (1978) study of short-term mortality rates among 50 heart attack patients randomly assigned to receive either manual or mechanical chest compression during cardiopulmonary resuscitation. The intervention was started and finished within the space of an hour; the heart attack patients could not very well get up and leave the hospital; and the dependent variable was quickly and easily gathered. A second situation conducive to minimal attrition is one in which the outcome is a matter of mandatory public record. For instance, in both the LIFE (Living Insurance for Ex-Offenders) experiment and the TARP (Transitional Aid for Released Prisoners) experiment (Rossi, Berk, & Lenihan, 1980), the main dependent variable was arrests, about which records were available for all participants from public sources. In general, however, attrition from conditions will occur in most field experiments, and pretests are vital to the methods we outline in Chapter 10 for dealing with attrition.

The Pretest-Posttest Control Group Design

Consequently, adding pretests to the basic randomized design is highly recommended:

$$\begin{array}{cccc} R & O & X & O \\ R & O & & O \end{array}$$

Or, if random assignment occurred after pretest,

$$\begin{array}{cccc} O & R & X & O \\ O & R & & O \end{array}$$

This is probably the most commonly used randomized field experiment. Its special advantage is its increased ability to cope with attrition as a threat to internal validity, in ways we outline in Chapter 10. A secondary advantage, however, is that it allows certain statistical analyses that increase power to reject the null hypothesis (Maxwell & Delaney, 1990). S. E. Maxwell (1994) says that allocating 75% of assessment to posttest and 25% to pretest is often a good choice to maximize power with this design. Maxwell, Cole, Arvey, and Salas (1991) discuss tradeoffs between ANCOVA using a pretest as a covariate and repeated-measures ANOVA with a longer posttest as a method for increasing power.

Although the researcher should try to make the pretest be identical to the outcome measures at posttest, this need not be the case. In research on child development, for example, tests for 8-year-old children must often be substantially different in content than those for 3-year-old children. If the pretest and posttest assess the same unidimensional construct, logistic test theory can sometimes be used to calibrate tests if they contain some common content (Lord, 1980), as McKay et al. (1978) did in the Cali study of changes in cognitive ability in 300 children between the ages of 30 and 84 months.

Alternative-Treatments Design with Pretest

The addition of pretests is also recommended when different substantive treatments are compared:

$$\begin{array}{cccc} R & O & X_A & O \\ R & O & X_B & O \end{array}$$

If posttests reveal no differences between groups, the researcher can examine pretest and posttest scores to learn whether both groups improved or if neither did.[5] This design is particularly useful when ethical concerns mitigate against comparing treatment

[5] Here and elsewhere, we do not mean to imply that change scores would be desirable as measures of that improvement. ANCOVA will usually be much more powerful, and concerns about linearity and homogeneity of regression are at least as important for change scores as for ANCOVA.

with a control condition, for example, in medical research in which all patients must be treated. It is also useful when some treatment is the acknowledged gold standard against which all other treatments must measure up. Comparisons with this standard treatment have particularly practical implications for later decision-making.

Multiple Treatments and Controls with Pretest

The randomized experiment with pretests can involve a control group and multiple treatment groups:

$$\begin{array}{cccc} R & O & X_A & O \\ R & O & X_B & O \\ R & O & & O \end{array}$$

H. S. Bloom's (1990) study of reemployment services for displaced workers used this design. More than 2,000 eligible unemployed workers were assigned randomly to job-search assistance, job-search assistance plus occupational training, or no treatment. Note that the first treatment included only one part of the treatment in the second condition, giving some insight into which parts contributed most to outcome. This is sometimes called a **dismantling study,** though the dismantling was only partial because the study lacked an occupational-training-only condition. Clearly, resources and often logistics prevent the researcher from examining too many parts, for each part requires a large number of participants in order to test it well. And not all parts will be worth examining, particularly if some of the individual parts are unlikely to be implemented in policy or practice.

This design can be extended to include more than two alternative treatments or more than one control condition. An example is the National Institute of Mental Health Treatment of Depression Collaborative Research Program (NIMH-TDCRP; Elkin, Parloff, Hadley, & Autry, 1985; Elkin et al., 1989; Imber et al., 1990). In this study, 250 depressed patients were randomly assigned to receive cognitive behavior therapy, interpersonal psychotherapy, antidepressant chemotherapy (imipramine) plus clinical management, or a placebo pill plus clinical management.

This design is also used to vary the independent variable in a series of increasing levels (sometimes called *parametric* or *dose-response* studies). For example, the Housing Allowance Demand Experiment randomly assigned families to receive housing subsidies equal to 0%, 20%, 30%, 40%, 50%, or 60% of their rent (Friedman & Weinberg, 1983). The Health Insurance Experiment randomly assigned families to insurance plans that required them to pay 0%, 25%, 50%, or 95% of the first $1,000 of covered services (Newhouse, 1993). The more levels of treatment are administered, the finer the assessment can be of the functional form of dosage effects. A wide range of treatment levels also allows the study to detect effects that might otherwise be missed if only two levels of a treatment that are not powerful enough to have an effect are varied. The Cali, Colombia, study (McKay et al., 1978), for example, administered a combined educational, nutritional, and medical treatment in four increasing dosage levels—990 hours, 2,070 hours, 3,130 hours, and 4,170 hours. At the smallest dosage—which itself took nearly a full year to implement and which might well be the maximum dosage many authors might consider—the effects were nearly undetectable, but McKay et al. (1978) still found effects because they included this wide array of higher dosages.

Factorial Designs

These designs use two or more independent variables (called *factors*), each with at least two levels (Figure 8.1). For example, one might want to compare 1 hour of tutoring (Factor A, Level 1) with 4 hours of tutoring (Factor A, Level 2) per week and also compare tutoring done by a peer (Factor B, Level 1) with that done by an adult (Factor B, Level 2). If the treatments are factorially combined, four groups or cells are created: 1 hour of tutoring from a peer (Cell A1B1), 1 hour from an adult (A1B2), 4 hours from a peer (A2B1), or 4 hours from an adult (A2B2). This is often described as a 2 × 2 ("two by two") factorial design written in the notation used in this book as:

$$\begin{array}{ccc} R & X_{A1B1} & O \\ R & X_{A1B2} & O \\ R & X_{A2B1} & O \\ R & X_{A2B2} & O \end{array}$$

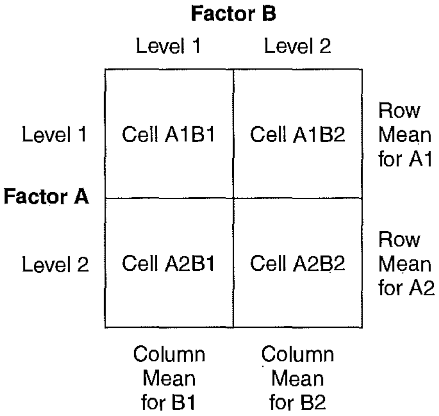
This logic extends to designs with more than two factors. If we add a third factor in which the tutor is or is not trained in effective tutoring methods (Factor C, Levels 1 and 2), we have a 2 × 2 × 2 design with 8 possible cells. The levels of the factors can include control conditions, for example, by adding a no-tutoring condition to Factor A. This increases the number of levels of A so that we have a 3 × 2 × 2 design with 12 cells. This notation generalizes to more factors and more levels in similar fashion.

Factorial designs have three major advantages:

- They often require fewer units.
- They allow testing combinations of treatments more easily.
- They allow testing interactions.

First, they often allow smaller sample sizes than would otherwise be needed.[6] An experiment to test for differences between peer versus adult tutoring might require

[6] Two exceptions to this rule are (1) detecting interactions of special substantive interest may require larger sample sizes because power to detect interactions is usually lower than power to detect main effects; and (2) if the outcome is a low base rate event (e.g., death from pneumonia during the course of a brief clinical trial) and if the treatments in both factors reduce death, their combined effect may reduce the number of outcome events to the point that more participants are needed in the factorial design than if just one treatment were tested.



		Factor B		
		Level 1	Level 2	
Factor A	Level 1	Cell A1B1	Cell A1B2	Row Mean for A1
	Level 2	Cell A2B1	Cell A2B2	Row Mean for A2
		Column Mean for B1	Column Mean for B2	

FIGURE 8.1 Factorial Design Terms and Notation

50 participants per condition, as might a second experiment to test for differences between 1 versus 4 hours of tutoring—a total of 200 participants. In a factorial design, fewer than 200 participants may be needed (the exact number would have to be determined by a power analysis) because each participant does double duty, being exposed to both treatments simultaneously.

Second, factorial designs allow the investigator to test whether a combination of treatments is more effective than one treatment. Suppose that an investigator runs both an experiment in which participants are assigned to either aspirin or placebo to see if aspirin reduces migraine headaches and a second experiment to test biofeedback versus placebo for the same outcome. These two experiments provide no information about the effects of aspirin and biofeedback applied jointly. The factorial design provides information about the effects of aspirin only, biofeedback only, aspirin plus biofeedback, or no treatment.

Third, factorial experiments test interactions among factors (Abelson, 1996; D. Meyer, 1991; Petty, Fabrigar, Wegener, & Priester, 1996; Rosnow & Rosenthal, 1989, 1996). Treatments produce *main effects;* for example, the main effect of aspirin relative to a placebo pill is to reduce headaches. Main effects are *average* effects that may be misleading if, for example, some kinds of headaches respond well to aspirin but others do not. Interactions occur when treatment effects are not constant but rather vary over levels of other factors, for example, if aspirin reduces tension headaches a lot but migraine headaches very little. Here the treatment (aspirin) interacts with a moderator variable (type of headache), the word *moderator* describing a second factor that interacts with (moderates the effect of) a treatment. The same general logic extends to designs with three or more factors, though higher order interactions are more difficult to interpret.

Interactions are often more difficult to detect than are main effects (Aiken & West, 1991; Chaplin, 1991, 1997; Cronbach & Snow, 1977; Fleiss, 1986), so large sample sizes with appropriate power analyses are essential whenever interactions are an important focus.[7] Indeed, some authors have argued that predicted interactions are sufficiently important in advancing scientific theory to warrant testing them at a larger than usual Type I error rate (Meehl, 1978; Platt, 1964; Smith & Sechrest, 1991; Snow, 1991). If testing a predicted interaction is at issue, deliberate oversampling of observations that are extreme on the interacting variables provides a more powerful (and still unbiased) test of the interaction, although it gives a poorer estimate of the total variance accounted for by the predictors. The test for the interaction could be done using an unweighted sample, and the test for total variance could be done using a sample that is weighted to reflect the population of interest (McClelland & Judd, 1993). This is a special case of optimal design theory (e.g., A. Atkinson, 1985) that can help select treatment levels and combinations that maximize the power of the design to detect parameters that may be of particular policy or theoretical interest.

When using a factorial design, the researcher need not actually assign units to all possible combinations of factors, though empty cells can reduce power. It might waste resources to test treatment combinations that are of no theoretical interest or unlikely to be implemented in policy. The New Jersey Negative Income Tax (NJNIT) experiment, for example, studied proposals for dealing with poverty and welfare reform (Kershaw & Fair, 1976)—specifically, the joint effects of two independent variables: the guarantee level and the tax rate. Guarantee level was an amount of money paid to poor families or individuals if they had no other income; it was defined as 50%, 75%, 100%, or 125% of the poverty level. The tax rate was the rate at which that guaranteed income was reduced as a family's other income rises: 30%, 50%, or 70%. So the design was a 4 × 3 factorial experiment that could assign participants to 12 different cells. However, the investigators assigned the 725 participants only to the eight cells that were not too costly and that were considered politically feasible for eventual policy implementation. The empty cells can complicate data analysis, but the flexibility of this option often outweighs the complications. This design is an example of a fractional factorial design that allows estimates of some higher order interaction terms even when the full factorial design is not implemented (Anderson & McLean, 1984; Box, Hunter, & Hunter, 1978; West, Aiken, & Todd, 1993).

Nested and Crossed Designs

In a crossed design, each level of each factor is exposed to (crossed with) all levels of all other factors. For example, in an educational experiment, if some students in each classroom are exposed to treatment and some to control, then the

[7] Interactions are ordinal (when you graph the cell means, the resulting lines do not cross) or disordinal (they do cross) (Maxwell & Delaney, 1990). Tests of ordinal interactions frequently have lower power than main effects, but tests of disordinal interactions are often more powerful than the test of either main effect. Rosnow and Rosenthal (1989) explain which lines should and should not cross when interactions are present.

treatment factor is crossed with classroom. In a **nested design,** some levels of one factor are not exposed to all levels of the other factors. For example, when some classrooms receive the treatment but not the control condition, classrooms are nested within treatment conditions. Crossed designs yield unconfounded statistical tests of all main effects and interactions, but nested designs may not. The distinction between nested and crossed designs is particularly relevant in the presence of higher order units (e.g., schools, hospitals, work sites). Often the researcher will nest treatments within these units to minimize the chances of diffusion and communication of treatment within higher order units. The dilemma is that the crossed design yields separate statistical estimates of the effects of higher order units, treatment conditions, and their interaction; but crossing increases problems such as diffusion of treatment. Each researcher will have to review the specifics of this tradeoff as it applies to the experiment at hand before deciding whether nesting or crossing is to be preferred.

A Disadvantage of Factorial Designs

Factorial designs are common in laboratory research and in highly controlled settings, such as those of some medical research. They are more difficult to implement in many field settings. They require close control over the combination of treatments given to each unit, but such control is difficult as more factors or more levels are included—especially if each cell has different eligibility criteria, as in pharmaceutical studies in which rules for who can receive which drug combinations can be complex. In addition, much field research is conducted to assess the policy implications of a proposed innovation. Yet the ability of policymakers to legislate or regulate interactions is low, given traditions of local control and professional discretion in the delivery of services and given difficulties in ensuring that social interventions are implemented as intended (Pressman & Wildavsky, 1984; Rossi & Wright, 1984). Policymakers are often more interested in generalized inferences about which treatments work than in the highly specific and localized inferences about the effects of particular combinations of particular levels of particular factors in a particular setting that factorial designs sometimes provide.

Longitudinal Designs

Longitudinal designs add multiple observations taken before, during, or after treatment, the number and timing of which are determined by the hypotheses under study:

R	*O* . . . *O*	*X*	*O*	*O* . . . *O*
R	*O* . . . *O*		*O*	*O* . . . *O*

These designs closely resemble the time-series studies in Chapter 6, but they have far fewer pre- and posttest observations. Longitudinal designs allow examination of how effects change over time, allow use of growth curve models of individual differences in response to treatment, and are frequently more powerful than designs with fewer observations over time, especially if five or more waves of measurement are used (Maxwell, 1998). So especially when sample sizes are small, adding pretests and posttests can improve power.

Longitudinal randomized experiments with multiple pretests are rare. Bloom (1990), for example, randomly assigned displaced workers to three treatment or control conditions designed to help them get jobs. He reported quarterly earnings at four pretests and four posttests. The pretests showed that participants experienced an acute drop in earnings during the one or two quarters immediately preceding assignment to conditions, perhaps reflecting a short-term job loss by workers who move rapidly into and out of the labor market. So regression effects might cause some improvement in all groups even if treatments did not work. Indeed, control participants did improve, though not as much as treatment participants.

The use of multiple posttests is more common. For example, the Cambridge-Somerville Youth Study began in 1939 when 650 adolescent boys were randomly assigned from blocked pairs either to a counseling program or to no treatment (W. McCord & McCord, 1959; Powers & Witmer, 1951), with a follow-up taken 37 years later in 1976 (J. McCord, 1978). A current study in a health maintenance organization aims to follow patients for their entire lives (Hillis et al., 1998). Here the multiple posttests explore whether treatment gains are maintained or changed over time. This is especially important if the primary outcome can be measured only many years later—for example, children's eventual educational and occupational achievement after participating in Head Start, mortality rates from AIDS among gay men after exposure to a program teaching safe sex, or lifetime earned income among Job Corps trainees. Sometimes longitudinal studies follow different outcomes simultaneously over time to explore the validity of a hypothesized causal chain of effects—for example, that a treatment to help children of lower socioeconomic status to rise out of poverty will first improve their aspirations, which will affect expectations, which will affect achievements in grammar school, which will help them successfully complete high school and college, which will finally lead to a better paying or higher status job as an adult. Here the timing of the observations follows the hypothesized schedule of events to be observed in order to explore if and at what point the chain breaks down.

Practical problems plague longitudinal designs. First, attrition rises with longer follow-up periods as, for example, participants move to unknown locations or simply tire of the research. Still, we are impressed with what tireless follow-up procedures can achieve with most populations (Ribisl et al., 1996). Second, some long-term outcomes, such as lifetime earned income, are nearly impossible to assess given current technology and limited access to such relevant data sources as the Internal Revenue Service or the Social Security Administration (Boruch & Cecil, 1979). Third, it is not always ethical to withhold treatments from participants for long periods of time, and the use of longitudinal observations on no-treatment or wait-list control-group participants is rare because such participants often simply obtain treatment elsewhere. An example of all these problems is provided by Snyder and Wills (1989; Snyder, Wills, & Grady-Fletcher, 1991), who randomly assigned 79 distressed couples to receive behavioral marital therapy ($N = 29$), insight-oriented marital therapy ($N = 30$), or a wait-list control group ($N = 20$). At 6-month and 4-year follow-ups they assessed outcomes only on the two treatment group conditions because control participants had already begun dropping out of the study despite the agreed-upon 3-month waiting period between pretest and posttest. Despite participant death, medical problems, and relocation out of state, Snyder and Wills (1989) were able to gather 4-year follow-up data on 55 of the 59 treatment couples—a remarkably high retention rate, although still a loss of participants such as one nearly always experiences in longitudinal research. Finally, a 4-year follow-up is far longer than most psychotherapy outcome studies use, but it is still far short of such long-term outcomes as distress levels over the life of the marriage or lifetime divorce rates. Even exemplary longitudinal studies such as this experience such problems.

Crossover Designs

Imagine an experiment in which some participants are randomly assigned to receive either Treatment A or B, after which they receive a posttest. In a crossover design, after that posttest the participants cross over to receive the treatment they did not previously get, and they take another posttest after that second treatment is over. In our design notation this crossover design is written:

$$\begin{array}{cccccc} R & O & X_A & O & X_B & O \\ R & O & X_B & O & X_A & O \end{array}$$

Sometimes the interval between treatments is extended so that the effects of the first treatment can dissipate before the second treatment begins.

This design is often used in medical research, as in cases in which several drugs are given to participants in a within-participants design and the crossover is used to counterbalance and assess order effects.[8] It is also used to gather even more causal information from a study that would otherwise stop after the first posttests were administered. In either use, the crossover design is most practical when the treatments promise short-term relief (otherwise carryover effects will occur), when the treatments work quickly (otherwise the experiment will take too long), and when participants are willing and able to continue through both treatments even if the first treatment fixes the problem. If analysis finds an interaction between treatments and

[8] The crossover design is a variation of a more general class called Latin squares (Cochran & Cox, 1957; Fisher & Yates, 1953; Fleiss, 1986; R. Kirk, 1982; Pocock, 1983; Rosenthal & Rosnow, 1991; Winer et al., 1991). Latin squares are widely used to counterbalance treatments in within-participants factors and to estimate effects in very large factorial designs in which all possible combinations of conditions cannot be administered.

TABLE 8.2 Ten Situations Conducive to Randomized Experiments

1. When demand outstrips supply
2. When an innovation cannot be delivered to all units at once
3. When experimental units can be temporally isolated
4. When experimental units are spatially separated or interunit communication is low
5. When change is mandated and solutions are acknowledged to be unknown
6. When a tie can be broken or ambiguity about need can be resolved
7. When some persons express no preference among alternatives
8. When you can create your own organization
9. When you have control over experimental units
10. When lotteries are expected

order, then the effect of the second round of treatments cannot be interpreted without taking order effects into account, although the first round of treatment is still just as interpretable as would have been the case without the crossover.

CONDITIONS MOST CONDUCIVE TO RANDOM ASSIGNMENT

This section (and Table 8.2) explicates the situations that increase the probability of successfully doing a randomized field experiment.

When Demand Outstrips Supply

When demand for service outstrips supply, randomization can be a credible rationale for distributing service fairly. For example, Dunford (1990) describes an experiment on the effects of a summer youth employment program. Initially, program personnel objected to randomly assigning some youths to jobs and others not. However, they also recognized that far fewer jobs were available than there were applicants, and they eventually agreed that random allocation of those jobs was fair. They later reported that the obviously unbiased nature of randomization helped them to show a vocal group of critics that entry into the program discriminated neither for nor against minority youth. Similarly, the Omnibus Budget Reconciliation Act of 1981 allowed states to experiment with novel approaches to welfare reform. Many states wanted to do so, but few states could afford to implement programs that could be given to all welfare recipients; random assignment was again accepted as a fair mechanism for distributing services in one experiment

(Gueron, 1985). Finally, the Milwaukee Parental Choice Program tested the use of school vouchers by random selection of participants when there were more applicants to a particular school and grade than could be accommodated (Rouse, 1998).

When demand exceeds supply, applicants originally assigned to the comparison condition sometimes reapply for the treatment. Experimenters need to be clear about whether they will have this right, and if they do, whether reapplicants will have priority over new applicants. Sometimes the right to reapply cannot be denied on ethical or regulatory grounds, as in the case of a distressed psychotherapy client assigned to a wait-list who becomes severely symptomatic or that of welfare recipients who have a regulatory right to reapply to a job-training program. It is crucial to negotiate support from everyone in the experiment about dealing with reapplicants, for dissenters can thwart those arrangements (Conrad, 1994). For example, the Rockefeller Foundation's Minority Female Single Parent (MFSP) program could not afford to provide services to all eligible candidates right away, so randomization was proposed as an ethically appropriate way to distribute services among the many eligible women (Boruch, 1997). However, some local program managers disagreed and spread their resources more thinly over a large number of women rather than limit the number of women served. Ultimately, if a large proportion of rejected applicants is likely to reapply and be accepted into treatment, the feasibility of a randomized experiment is questionable. If the proportion of successful reapplicants is likely to be small, methods that we discuss in Chapter 10 for dealing with treatment implementation problems may be useful.

When an Innovation Cannot Be Delivered to All Units at Once

Often it is physically or financially impossible to introduce an innovation simultaneously to all units. Such situations arise in education as curricula are slowly changed, as new teaching devices filter down through the schools in a system, or as computers are introduced or new training schemes are implemented. In these situations, the experiment can deliberately introduce the innovation in stages, with some units receiving it before others on a random basis. This provides an experimental and control comparison until the point at which the controls get their turn for treatment. It is even better if it can be done using the switching-replications design feature described for previous quasi-experimental designs, but with replications now randomly assigned.

When Experimental Units Can Be Temporally Isolated: The Equivalent-Time-Samples Design

Although we typically think of randomly assigning people, schools, communities, or cities to conditions, we can also randomly assign times to conditions (Hahn, 1984). Campbell and Stanley (1963) called this an "Equivalent Time Samples Design" to highlight that randomization equates the time periods in which the treatment was present to those in which it was absent. Edgington (1987) provides several examples of single-participant designs in which treatments were presented and removed randomly over time—a comparison of three drugs for narcolepsy, of a medication with a placebo for an intestinal disorder, and of the effects of artificial food colorings with placebo on the behavior of hyperactive children. The effect must be of short duration so that it can decrease in magnitude when treatment is withdrawn; and the effect must continue to respond to repeated exposure to treatment so that it can increase when treatment is readministered.

But the principle applies to more than just single-participant designs. It can be used when there are naturally occurring rotations of groups and each group is isolated from the others in time. Thus, when 24 groups of persons came for sequential 2-week stays at a pastoral counseling center, Mase (1971) randomly assigned those groups to one of two kinds of sensitivity training, twelve groups receiving each kind. In this example, the creation of simultaneous treatment and control conditions might have led to diffusion of treatment or other reactive threats to validity, but the equivalent-time-samples design avoided such problems. Note, however, that participants are now nested within time samples in the same way they could be nested within some aggregate such as a school or a neighborhood; the analysis should take this into account.

When Experimental Units Are Spatially Separated or Interunit Communication Is Low

When units are geographically separated and have minimal contact with one another, or when they can be made this way, those units can be randomly assigned. This often occurs in organizations that have many branches, for example, supermarkets, units in the armed forces, university alumni, schools within school districts, wards within hospitals, residential units of religious orders, branches of health clubs in large cities, and dealerships that sell automobiles, appliances, and the like. However, spatial isolation does not guarantee minimal contact, so care should be taken to check that this is indeed the case.

For example, an experiment in Peru studied the effects of providing gynecological and family planning services to clients of 42 geographically separated community clinics (Population Council, 1986). Clinics were assigned randomly to receive one, two, or four physician visits per month. The geographical separation of clinics meant that women tended to visit the same clinic over time, so little diffusion of treatment was likely. If some diffusion was possible (e.g., if women regularly visited two clinics very close to each other), the researchers could have blocked clinics by geographic area and assigned areas rather than individual clinics. Similarly, Perng (1985) randomly assigned people to six different methods that the Internal Revenue Service was considering for collecting delinquent income tax returns. Most people were separated geographically. But even if they had been in close physical proximity to each other, by law the very fact that their tax return was part of a study was confidential, and people are generally reluctant to discuss their income tax returns; so it was unlikely that communication between people in different conditions would occur.

These experiments had an additional strength; both took advantage of the natural appearance of the interventions to randomize treatments unobtrusively. After all, patients expect that physicians will visit clinics and are not likely to notice minor variations in the number of times those visits occur. Those receiving delinquent tax letters from the IRS are rarely familiar enough with specific IRS procedures to know that any variation on normal routine was occurring. Unobtrusiveness is a worthy goal to strive for, except when treatment is deliberately designed to stand out from what respondents expect.

When Change Is Mandated and Solutions Are Acknowledged to Be Unknown

Sometimes, all concerned parties agree that an undesirable situation needs changing, but it is not clear which changes we should make despite passionate advocacy of certain alternatives by interested parties. If administrative, political, and economic conditions allow, trying out several alternative changes in a formal experiment is more likely to win acceptance. An example is the Minneapolis Spouse Abuse Experiment (Berk, Smyth, & Sherman, 1988). Spouse abuse is a serious felony that can lead to the murder of the spouse, and so police officers who are called to such a crime must take some action. But concerned parties disagreed about whether that action should be to do on-the-spot counseling between the two spouses, to require the offender to leave the premises for 8 hours, or to arrest the offender. An administrator who had an attitude favoring experimentation in finding a solution allowed the implementation of a randomized experiment to test which of these three options worked best. Similarly, a randomized experiment to treat severely mentally ill patients with either standard care or a radically different form of community care could be done in part because all parties acknowledged that they were unsure which treatment worked best for these patients (Test & Burke, 1985).

Though such planned variation studies promise important results, each variation may not define its goals *exclusively* in terms of the same target problem. In the Minneapolis Spouse Abuse Experiment, this potential disagreement was not a problem because most parties agreed that the end point of interest was a decrease in postintervention repeat violence. However, disagreement may be more likely if participants are assigned to projects with different management, staff, and funders than to projects in which all variations are implemented by the same people. Nor will the directors of the various projects always agree which measures should be used to measure those things they are trying in common to change.

When a Tie Can Be Broken or Ambiguity About Need Can Be Resolved

Assignment of people to conditions on the basis of need or merit is often a more compelling rule to program managers, staff, and recipients than is randomization. Such considerations are one justification for the regression discontinuity design. However, the need or merit of some people is often ambiguous. In those cases, the ambiguity can sometimes be resolved by randomly assigning people of ambiguous need to conditions, perhaps in combination with the regression discontinuity design. Similarly, Lipsey, Cordray, and Berger (1981) used random assignment to resolve ambiguity in their evaluation of a juvenile delinquency diversion program. In a quasi-experimental design, police officers used their best judgment as to whether an arrested juvenile needed to be counseled and released, referred to probation, or diverted to a more intensive social service project that provided counseling, remedial education, recreation, and substance abuse services. However, when the officer was unsure which assignment was most needed and also judged that either counseling and release or diversion would be appropriate, the officer randomized juveniles to one of these two conditions.

In such tie-breaking experiments, generalization is restricted to persons scoring in the area of ambiguous need, the group about which we know least as far as effective treatment. However, if an organization specializes in treating the best, the worst, or the full range of participants, its officials may well object that evaluating their performance with "ambiguous" participants is insensitive to what they really do. Fortunately, it may be possible to link a tie-breaking experiment with some form of interpretable quasi-experiment, as Lipsey et al. (1981) did, to satisfy these objections.

When Some Persons Express No Preference Among Alternatives

Even if ethics or public relations require that people be allowed to choose which option they will receive, persons who express no preference from among the options can be assigned by chance. For example, Valins and Baum (1973) wanted to study some effects of physical environment on university freshmen who entered one of two kinds of living quarters that differed in the number of persons a resident was likely to meet each day. The authors restricted the study to the 30% of freshmen who expressed no preference for either kind of living quarters. College authorities assigned this 30% to living units on a haphazard basis; but it would presumably have been easy to do the assignment randomly. Of course, limiting the experiment to persons who have no preference does make generalization beyond such persons more problematic. If the full range of decisive and no-preference respondents is of interest, the randomized experiment with the no-preference respondents could be conducted along with the best possible quasi-experiment with the decisive respondents. Then the results of the studies can be compared, with the weakness of one study being the strength of the other. Where the results coincide, a global overall inference is easier.

When You Can Create Your Own Organization

Random assignment is an accepted part of the organizational culture of laboratory experimentation, but most field experiments are conducted in organizational cultures in which randomization is mostly foreign. Yet sometimes researchers can create their own organizations in which they can make the practice of randomization a more usual norm. For example, university psychology departments often set up a psychological services center to facilitate the training of graduate students in clinical psychology and to allow department faculty members to exert more experimental control than would typically be possible in most clinics (Beutler & Crago, 1991). In such centers, researchers can better control not just randomization but also such features as treatment standardization, measurement, and case selection. Freestanding research institutes and centers focused on particular problems frequently allow similar levels of broad control. The California Smokers' Helpline, for example, provides free smoking cessation help to smokers in that state who call the helpline (Zhu, 1999). Randomizing callers to treatment and control was not feasible. All callers received a treatment mailing with instructions to call back when they were ready to start treatment. Those who did not call back were then randomized into two groups: no further action or proactive callback from the treatment staff to begin treatment. In principle, this procedure could be used to randomly subdivide the nonresponders in any quasi-experimental treatment group into treatment and control—for example, those who request psychotherapy but fail to show for appointments, those who are given prescriptions but fail to fill them, those who are accepted to a job training program but fail to attend, and so forth. Finally, researchers can sometimes set up organizations just to control randomization, as is often done in multisite medical trials in which a central clearinghouse controlled by the researcher is created to do randomization. The National Institute of Mental Health (NIMH) Collaborative Depression Project (Collins & Elkin, 1985) used this clearinghouse method to control randomization.

When You Have Control over Experimental Units

Being able to establish one's own organization or randomization clearinghouse is rare. Most field researchers are guests in someone else's organization, and they derive many of their possibilities for control from their powerful hosts. An example comes from an evaluation of solutions to the "peak load" problem by utility companies (Aigner & Hausman, 1980). Electricity usage varies by time of day, and the utility company must have enough capacity to meet peak demand even if that capacity is largely unused at other times. Building that capacity is expensive. So utility companies wanted to know whether charging more for electricity during peak demand periods would reduce demand and so reduce the need to build more capacity. Experimenters were able to randomly assign households to higher peak demand rates versus standard rates because their hosts completely controlled electricity supply to the affected households and were interested in getting an answer to this question with experimental methods.

Randomization is also more likely whenever major funders insist on it. For example, both the National Institute on Drug Abuse and the National Institute on Alcohol and Alcohol Abuse have offered funding for innovative service provision contingent upon evaluation of those services by rigorous experimental designs, both paid for by the grant (Coyle, Boruch & Turner, 1991). The NIMH Collaborative Depression project used the same approach (Boruch & Wothke, 1985). However, especially when funder and fundee have a long-term relationship, the use of the purse strings for control can lead to tension. Lam, Hartwell, and Jekel (1994), for example, noted the "contentious codependence" (p. 56) that developed between Yale University and the city of New Haven, in which Yale is located, due to the fact that Yale researchers frequently offer social services to the city that it might not otherwise be able to afford but with a research string attached. There is a thin line between contentious codependence and oblique coercion, and it is even self-defeating to conduct a randomized experiment in a way that directly or indirectly demeans hosts or respondents. After all, the motivation for hosts and participants to volunteer tomorrow may well be related to how we treat them in experiments today.

When Lotteries Are Expected

Lotteries are sometimes used as a socially accepted means of distributing resources. Examples include a lottery used to assign female students to dormitories at Stanford (Siegel & Siegel, 1957), a lottery to choose among applicants to a newly developed "magnet" school (Zigulich, 1977), and the 1970 draft lottery in the United States (Notz, Staw, & Cook, 1971). In the latter case, Hearst, Newman, and Hulley (1986) asked whether being randomly assigned an eligible draft number elevated mortality and found that it did do so. Angrist et al. (1996a) confirmed this finding, with the average causal effect on mortality of being randomly assigned an eligible draft number equal to less than one tenth of one percent. In these cases, the motivation for randomization was not to do research but rather to capitalize on the perception that randomization is an unbiased way of distributing a resource. These social uses of randomization create a natural randomized experiment that the investigator can exploit. Unfortunately, formal social lotteries do not occur frequently, so they cannot be relied upon as a means of creating probabilistically equivalent groups very often.

WHEN RANDOM ASSIGNMENT IS NOT FEASIBLE OR DESIRABLE

Even when interest exists in whether a treatment is effective, some circumstances mitigate against using a randomized experiment to answer the question. First, randomized experiments may not be desirable when quick answers are needed. Typically, several years pass between the conception of a major field experiment and the availability of results—particularly if the treatment requires time (as with long-term psychotherapy) and if medium- to long-term outcomes are of interest (as with lifetime earnings). In the New Jersey Negative Income Tax Experiment, for example, "the four years of the operating phase were sandwiched between 44 months of planning and design and 16 months of data analysis" (Haveman, 1987, p. 180)—8 years total. So, if information is needed rapidly, alternatives to randomized experiments may be better. For example, the Program Evaluation and Methodology Division (PEMD) of the U.S. General Accounting Office (GAO) frequently fielded questions from legislators who wanted answers quickly about pending decisions. Some of those questions involved the effects of programs or policies. A delay of a few years might delay the decision too long—indeed, the question may no longer be of policy interest, and the legislator who asked the question may no longer be serving. Consequently, PEMD rarely used randomized experiments, relying instead on combinations of quasi-experiments, surveys, and reviews of existing literature about the effects of related policies (Chan & Tumin, 1997; Datta, 1997; Droitcour, 1997). Such procedures may be weaker for inferring cause than a new randomized experiment, because even when the literature contains randomized experiments, they are rarely on the exact question of legislative interest. But GAO's methods are almost always more timely than those of a new randomized experiment and often of reasonable accuracy.

Second, randomized experiments provide a precise answer about whether a treatment worked (Cronbach et al., 1980). But the need for great precision may be low in many cases. For example, when much high-quality prior information exists about the treatment, a review of existing literature may be a better use of resources than would be a new randomized trial. When a causal question is of secondary interest to a noncausal question, such as whether services are being provided as intended, program monitoring procedures may be better. When an effect is so large and dramatic that no one doubts it resulted from the treatment, as with the dramatic effects of screening for PKU on PKU-based retardation among children, investing in an additional randomized experiment may be superfluous.

Third, randomized experiments can rarely be designed to answer certain kinds of questions. It is not possible to assign persons at random to variables that cannot be manipulated, such as age or race, or to manipulate events that occurred in the past, such as the effects of the death of President John F. Kennedy or of the Great Depression in the 1930s. It is unethical to assign persons at random to many manipulable events that cause significant harm, such as to cigarette smoking or to having a spinal cord injury.

Fourth, before conducting an experiment, a good deal of preliminary conceptual or empirical work must be done. The Federal Judicial Center (1981) recommends that, before an experiment is conducted, it should be demonstrated that the present conditions need improvement, that the proposed improvement is of unclear value, that only an experiment could provide the necessary data to clarify the question, that the results of the experiment would be used to change the practice or policy, and that the rights of individuals would be protected in the experiment. Similarly, the National Cancer Institute's five-phase model of testing a potential cancer control method suggests that, before a randomized experiment is conducted, the existing scientific literature should be identified and synthesized to see if an empirically supportable and testable hypothesis can be generated; pilot tests should be done to investigate the feasibility or acceptability of an intervention; studies assessing participation and adherence in the population should be conducted; data collection forms should be developed and validated; and quasi-experimentally controlled studies should be used to provide preliminary evidence about treatment effects (Greenwald & Cullen, 1984). Premature experimentation can be a great waste of resources—indeed, it can undermine potentially promising interventions for which there has not yet been time to develop recruitment procedures, identify and fix implementation problems, and serve the clientele long enough to make a difference.

DISCUSSION

The randomized experiment is often the preferred method for obtaining a precise and statistically unbiased estimate of the effects of an intervention. It involves fewer assumptions than other methods, the validity of those assumptions is usually easier to check against the data and the procedures used, and it requires less prior knowledge about such matters as selection processes and unit characteristics than do quasi-experiments, causal modeling, and selection bias models. Given all these strengths, it is easy to forget the many practical problems that can arise in implementing randomized experiments.

One practical problem concerns the feasibility and desirability of experimenting in particular cases. Some experimental manipulations are not ethical, as in the case of a physician deciding that a certain class of patients must be given a certain treatment and so cannot be randomized, or of an experimental treatment producing positive or negative effects that are so large that it would be unethical to continue to study them. Other times, it is not acceptable to wait the years that a well-designed and implemented experiment can take. Still other times, legal problems arise, not only because ethical violations can become legal problems but also because the law is often involved in certain experimental situations, for instance, when it mandates experimental evaluations of a program; when participants are directly under legal scrutiny, as with prisoners; or when legal systems are themselves the target of study.

A second practical problem is that a sufficiently large number of people (units) may not exist who are both eligible and willing to receive the treatment if assigned to it at random. Many is the experiment that has failed on this count. Frequently, especially with researchers who have never run a large field experiment before, the number of eligible people is vastly overestimated, as is the ease with which they can be located. When they are located, they often refuse to participate. In the worst case, the result is the death of the experiment for lack of participants.

A third practical problem is that the randomization procedure is not always properly designed and implemented. Sometimes this problem occurs because the researcher does not understand what random assignment is and so substitutes a seemingly haphazard assignment procedure. Or the researcher may introduce ad hoc adjustments to a random assignment procedure that seems to be yielding groups that are unequal before treatment, all the while thinking that these procedures are random when they are not. Other times the researcher correctly designs random assignment procedures but fails to create or supervise the procedures for implementing random assignment, so the assignment is implemented improperly. Whenever randomization is incorrectly or incompletely implemented, its benefits may be thwarted.

A fourth practical problem is that the treatment assigned is not always the treatment received. Participants may fail to fully receive the treatment to which they are assigned or may not receive it at all, as in the case of patients assigned to drug therapy who fail to take the drug or take only part of it. They may cross over to another condition (in a design that does not call for a crossover), as in the case of participants in a control condition who reapply for treatment and are accepted. Diffusion of treatment may occur through such means as treatment-related communication between participants in different conditions. Here, too, the participant is now receiving some part of both conditions. In all these cases, the intended treatment contrast is thwarted. If so, although the inference that *assignment to condition caused outcome* is still clear, the construct validity of the treatment is not clear. Hence it can be useful to prevent these failures of treatment implementation or measure their occurrence in many experiments in which pure treatment contrasts are desired.

A fifth problem is attrition. The randomized experiment does not just aim to make groups equivalent before treatment begins; it also aims to make groups equivalent at posttest in all respects except for differences in treatment conditions. Differential attrition from conditions after initial random assignment can vitiate this latter aim. Such attrition occurs often in field experiments. So preventing attrition, coping with attrition, measuring attrition, and analyzing data with attrition all become crucial adjunct topics to the study of the randomized experiment.

This chapter, being mostly about the design and logic of randomized experiments, has skirted all these problems in the interests of presenting the simplest case and its variants. But the researcher needs to know about these problems because they bear on the decision whether to use the randomized experiment at all, and if the decision is to do so, then they bear on how well the experiment is implemented and subsequently interpreted. So we turn to these problems in more detail in the next two chapters.

Glossary

Alternative Hypothesis: Whatever alternative to the null hypothesis is being considered. (See also *Null Hypothesis, Null Hypothesis Significance Testing*)

Analogue Experiment: An experiment that manipulates a cause that is similar to another cause of interest in order to learn about the latter cause.

Assignment Variable: A variable or variables used to assign units to conditions.

Attrition: Loss of units; in randomized experiments, refers to loss that occurs after random assignment has taken place (also called *Mortality*).

Autocorrelation: The correlation of consecutive observations over time.

Balanced Sample: A purposive sample whose mean on a characteristic matches the population mean for that characteristic.

Bandwidth: The capacity of a method to provide data about many different kinds of questions, often at the cost of reduced precision in the answers.

Batch Randomization: Many or all units are available to be assigned to conditions at one time.

Between-Participants Design: Different units are studied in different conditions. (See also *Within-Participants Design*)

Bias: Systematic error in an estimate or an inference.

Blocking: The process of dividing units into groups with similar scores on a blocking variable, each group having the same number of units as the number of conditions. (See also *Matching, Stratifying*)

Carryover Effects: The effects of one treatment do not end prior to the administration of a second treatment, so that the effects observed in the second treatment include residual effects from the first.

Case-Control Study: A study that contrasts units with an outcome of interest to those without the outcome to identify retrospectively the predictors or causes of the outcome (also called *Case-Referent Study*).

Case-Referent Study: See *Case-Control Study*

Causal Description: Identifying that a causal relationship exists between A and B.

Causal Explanation: Explaining how A causes B.

Causal Generalization: Inferences that describe how well a causal relationship extends across or beyond the conditions that were studied.

Causal Model: A model of causal relationships, usually with mediators; sometimes refers to efforts to identify causes and effects in nonexperimental studies.

Cause: A variable that produces an effect or result.

Ceiling Effect: Responses on a variable closely approach the maximum possible response so that further increases are difficult to obtain. (See also *Floor Effect*)

Coherence Theory of Truth: An epistemological theory that says a claim is true if it belongs to a coherent set of claims.

Comparison Group: In an experiment, a group that is compared with a treatment group and that may receive either an alternate intervention or no intervention. (See also *Control Group, Placebo, Treatment Group*)

Compound Path: A path consisting of two or more direct paths connected together.

Confirmation: The strategy of showing that a hypothesis is correct or is supported by evidence.

Confound: An extraneous variable that covaries with the variable of interest.

Construct: A concept, model, or schematic idea.

Construct Validity: The degree to which inferences are warranted from the observed persons, settings, and cause-and-effect operations sampled within a study to the constructs that these samples represent.

Control Group: In an experiment, this term typically refers to a comparison group that does not receive a treatment but that may be assigned to a no-treatment condition, to a wait list for treatment, or sometimes to a placebo intervention group. (See also *Comparison Group, Placebo, Treatment Group*)

Convergent Validity: The idea that two measures of the same thing should correlate with each other. (See also *Discriminant Validity*)

Correlation: A measure of the strength of relationship between two variables.

Correlational Study: A study that observes relationships between variables. (See also *Nonexperimental Study, Observational Study, Quasi-Experiment*)

Correspondence Theory of Truth: An epistemological theory that says a knowledge claim is true if it corresponds to the world.

Counterbalancing: In within-participants designs, arranging the order of conditions to vary over units so that some units are given Treatment A first but others are given Treatment B first.

Counterfactual: The state of affairs that would have happened in the absence of the cause.

Critical Multiplism: The claim that no single method is bias free, so that the strategy should be to use multiple methods, each of which has a different bias.

Cross-Lagged Panel Design: A design in which a cause and an effect are both measured at Times 1 and 2 and the researcher looks to see if the relationship between the cause at Time 1 and the effect at Time 2 is stronger than the relationship between the effect at Time 1 and the cause at Time 2.

Crossed Designs: Designs in which all units are exposed to all conditions.

Debriefing: The process of informing research participants about a study after it is over.

Deflationism: An epistemological theory that says truth is a trivial linguistic device for assenting to propositions expressed by sentences too numerous, lengthy, or cumbersome to utter.

Dependent Variable: Often synonymous with *effect* or *outcome,* a variable with a value that varies in response to the independent variable.

Design Element: Something an experimenter can manipulate or control in an experiment to help address a threat to validity.

Direct Path: A causal path that directly connects two variables.

Discriminant Validity: The notion that a measure of A can be discriminated from a measure of B, when B is thought to be different from A; discriminant validity correlations should be lower than convergent validity correlations. (See also *Convergent Validity*)

Dismantling Study: A study that breaks down a treatment into its component parts to test the effectiveness of the parts.

Double-Blind Study: An experiment in which both the treatment provider and treatment recipient are unaware of which treatment or control condition is being administered, primarily used in medical clinical trials.

Effect Size: A measure of the magnitude of a relationship, specific instances of which include the standardized mean difference statistic, the odds ratio, the correlation coefficient, the rate difference, and the rate ratio.

Effectiveness: How well an intervention works when it is implemented under conditions of actual application. (See also *Efficacy*)

Efficacy: How well an intervention works when implemented under ideal conditions. (See also *Effectiveness*)

Endogenous Variable: A variable that is caused by other variables within the model.

Epistemology: Philosophy of the justifications for knowledge claims.

Ethnography: Unstructured exploratory investigation, usually of a small number of cases, of the meaning and functions of human action, reported primarily in narrative form.

Exogenous Variable: A variable that is not caused by other variables in the model.

Expectation: The mean of a statistic based on repeated samplings. (See also *Sampling Error*)

Experiment: To explore the effects of manipulating a variable.

External Validity: The validity of inferences about whether the causal relationship holds over variations in persons, settings, treatment variables, and measurement variables.

Falsification: To show that data are inconsistent with a theory or hypothesis.

Fatigue Effects: Participants tire over time, causing performance to deteriorate in later conditions or later assessments. (See also *Practice Effects, Testing Effects*)

Fidelity: The capacity of a method to provide precise answers about a narrow question, often at the cost of high bandwidth.

File Drawer Problem: The hypothesis that studies that were rejected because of reviewer prejudice against null findings are never published and so remain unavailable to future literature reviews, resulting in a systematic bias in the results of the review. (See also *Publication Bias*)

Floor Effect: Responses on a variable approach the minimum possible score so that further decreases are difficult to obtain. (See also *Ceiling Effect*)

Functional Form: The characteristics of the true relationship among variables, represented graphically by the shape of the relationship (e.g., is it a curve?) and represented statistically by a model that may include nonlinear terms (e.g., powers and interactions) or other transformations.

Heterogeneity of Irrelevancies: Identifying things that are irrelevant to the inference at issue and then making those irrelevancies heterogeneous so inferences are not confounded with the same irrelevancy or with different irrelevancies whose direction of bias is presumptively in the same direction.

Hidden Bias: Unobserved variables that may cause bias in treatment effect estimates. (See also *Omitted Variables*)

Implementation: The activities, both intended and unintended, that did and·did not occur as part of the treatment conditions. (See also *Process Model*)

Independent Variable: Often synonymous with *cause* or *treatment,* a variable that purports to be independent of other influences. Some authors advocate a more limited usage whereby a variable is independent only if the methodology isolates the variable from other influences. (See also *Dependent Variable*)

Indirect Path: A path between two variables that requires going through a third variable to make the connection.

Informed Consent: The process of giving research participants the information they need to make an informed choice about whether to participate in a study given its risks and benefits.

Instrumental Variable: A variable or set of variables (or more generally, an estimation technique) that is correlated with outcome only through an effect on other variables.

Intent-to-Treat Analysis: An analysis of a randomized experiment in which units are analyzed in the condition to which they were assigned, regardless of whether they actually received the treatment in that condition.

Interaction: In experiments, when the effects of treatment vary over levels of another variable. (See also *Moderator*)

Internal Validity: The validity of inferences about whether the relationship between two variables is causal.

Interrupted Time-Series Design: A design in which a string of consecutive observations is interrupted by the imposition of a treatment to see if the slope or intercept of the series changes as a result of the intervention.

Inus Condition: From philosopher J. L. Mackie (1984), the idea that a cause is an *i*nsufficient but *n*on-redundant part of an *u*nnecessary but *s*ufficient condition for bringing about an effect.

Latent Variable: A variable that is not directly observed but is inferred or estimated from observed variables. (See also *Observed Variable*)

Local Molar Causal Validity: Alternative phrase for internal validity suggested by Donald Campbell (1986) as more clearly indicating the nature of internal validity.

Logic of Causation: To infer a causal relationship, the requirements that cause precedes effect, that cause covaries with effect, that alternative explanations can be ruled out, and that knowledge is available of what would have happened in the absence of the cause. (See also *Counterfactual*)

Lowess Smoother: A *lo*cally *we*ighted *s*catterplot *s*moother in which the result is a regression-fitted value for a local regression on a sample of observations in the vicinity of a selected horizontal axis point, done for many such points.

Matching: Sometimes synonymous with blocking, sometimes more specific to imply blocks in which units are exactly equal (rather than just similar) on a matching variable. (See also *Blocking, Stratifying*)

Measurement Attrition: Failure to obtain measures on units (whether or not they are treated).

Mediator: A third variable that comes between a cause and effect and that transmits the causal influence from the cause to the effect. (See also *Molecular Causation*)

Meta-Analysis: A set of quantitative methods for synthesizing research studies on the same topic (also called *Research Synthesis*).

Moderator: In an experiment, a variable that influences the effects of treatment. (See also *Interaction*)

Modus Operandi (M.O.): A method for inferring the cause of an observed effect by matching the pattern of observed effects to the patterns usually left by known causes (analogous to detective work investigating whether clues left at a crime match the modus operandi of known criminals).

Molar Causation: An interest in the overall causal relationship between a treatment package and its effects, in which both may consist of multiple parts.

Molecular Causation: An interest in knowing which parts of a treatment package are more or less responsible for which parts of the effects through which mediational processes. (See also *Mediator*)

Mortality: See *Attrition*

Multiple Operationalism: The notion that all the operations used to index a construct are relevant to the construct of interest but that, across the set of operations, there will be heterogeneity in conceptually irrelevant features.

Natural Experiment: Investigates the effects of a naturally occurring event, sometimes limited to events that are not manipulable, such as earthquakes, and sometimes used more generally.

Nested Designs: Designs in which units are exposed to some but not all conditions. (See also *Nesting, Unit of Analysis Problem*)

Nesting: When some units (e.g., students) are grouped together into aggregate units (e.g., classrooms), units are said to be nested within aggregates. (See also *Nested Designs, Unit of Analysis Problem*)

Nonequivalent Dependent Variable: A dependent variable that is predicted *not* to change because of the treatment but is expected to respond to some or all of the contextually important internal validity threats in the same way as the target outcome.

Nonexperimental Study: Any study that is not an experiment. (See also *Correlational Study, Observational Study*)

Nonrecursive Model: In the structural equation modeling literature, a model that allows reciprocal causation, although some literatures use the term differently. (See also *Reciprocal Causation, Recursive Model*)

Null Hypothesis: The hypothesis being tested, traditionally that there is no relationship between variables. (See also *Alternative Hypothesis, Null Hypothesis Significance Testing*)

Null Hypothesis Significance Testing: The practice of testing the hypothesis that there is no effect [the nil hypothesis] at $\alpha = .05$ and then declaring that an effect exists only if $p < .05$. (See also *Alternative Hypothesis, Null Hypothesis*)

Observational Study: A study in which variables are observed rather than manipulated; used in some literatures to include quasi-experiments (see also *Correlational Study, Nonexperimental Study, Quasi-Experiment*)

Observed Variable: A variable that is directly measured in a study.

Odds Ratio: An effect size measure for the difference between groups on a dichotomous outcome.

Omitted Variables: Variables that are not in a model or an analysis that influence both the cause and the effect and so may cause bias. (See also *Hidden Bias*)

Ontology: Philosophy of the nature of reality.

Operationalization: Usually synonymous with operations but sometimes used in a restricted sense to imply the methods used to represent a construct. (See also *Operations*)

Operations: The actions actually done in a study to represent units, treatments, observations, settings, and times. (See also *Operationalization*)

Order Effects: The outcome of a study is affected by the order in which the treatments were presented.

Participant Observation: A form of observation in which the researcher takes on an established participant role in the context being studied.

Path Coefficient: A measure of the strength of relationship between two variables connected by a direct path.

Pattern Matching: The general concept of matching a pattern of evidence to the pattern predicted by theory or past research.

Placebo: An intervention that does not include the presumed active ingredients of treatment. (See also *Control Group, Treatment Group*)

Power: The probability of correctly rejecting a false null hypothesis; in an experiment, usually interpreted as the probability of finding an effect when an effect exists. (See also *Type II error*)

Practice Effects: Participants become better at something the more often they do it, a potential problem in within-participants designs in which repeated tests are given to the same participants. (See also *Fatigue Effects, Testing Effects*)

Pragmatic Theory of Truth: An epistemological theory that says a claim is true if it is useful to believe that claim.

Process Model: A model that portrays the sequence of events that occur in an intervention. (See also *Implementation*)

Propensity Score: A predicted probability of group membership based on observed predictors, usually obtained from a logistic regression.

Publication Bias: A prejudice on the part of manuscript reviewers against publishing studies that fail to reject the null hypothesis. (See also *File Drawer Problem*)

Purposive Sample: A method by which units are selected to be in a sample by a deliberate method that is not random. (See also *Balanced Sample*)

Purposive Sampling of Heterogeneous Instances: Selecting features of a study (units, treatments, observations, settings, times) that are heterogeneous on characteristics that might make a difference to the inference.

Purposive Sampling of Typical Instances: Selecting features of a study (units, treatments, observations, settings, times) that are similar to typical units in the population of interest, where *typical* may be defined as the mean, median, or mode of that population, determined either impressionistically or based on data about the population.

Quasi-Experiment: An experiment in which units are not randomly assigned to conditions. (See also *Correlational Study, Nonexperimental Study, Observational Study*)

Random Assignment: In an experiment, any procedure for assigning units to conditions based on chance, with every unit having a nonzero probability of being assigned to each condition. (See also *Randomized Experiment*)

Random Measurement Error: Chance factors that influence observed scores so that those scores do not measure the true variable of interest.

Random Sampling: Any procedure for selecting a sample of units from a larger group based on chance, frequently used in survey research to facilitate generalization from sample to population.

Random Selection: More general term that is sometimes used synonymously with either random sampling or random assignment in different contexts.

Randomized Experiment: An experiment in which units are randomly assigned to conditions. (See also *Random Assignment*)

Reciprocal Causation: When two variables cause each other. (See also *Nonrecursive Model, Recursive Model*)

Recursive Model: In the structural equation modeling literature, a model that does not allow reciprocal causation, although some literatures use the term differently. (See also *Nonrecursive Model, Reciprocal Causation*)

Regression Discontinuity: The regression line for a treatment group is discontinuous from the regression line for the control group.

Regression Discontinuity Design: An experiment in which units are assigned to conditions based on exceeding a cutoff on an assignment variable.

Reliability: Consistency.

Research Synthesis: See *Meta-Analysis*

Response Burden: The costs of adding additional measurement to a study in terms of respondent time, energy, and goodwill.

Risk Analysis: An analysis of the likely risks and benefits from a study, including the size of the risks, the likelihood of the risks, and who will suffer them.

Sampling Error: That part of the difference between a population parameter and its sample estimate that is due to the fact that only a sample of observations from the population are observed. (See also *Expectation*)

Secondary Analysis: Reanalysis of primary study data after the study is completed, usually done by someone other than the original authors.

Selection: (1) The process by which units are assigned to conditions. (2) A threat to internal validity in which systematic differences over conditions in respondent characteristics could also cause the observed effect.

Selection Bias: When selection results in differences in unit characteristics between conditions that may be related to outcome differences.

Selection Bias Model: A statistical model that attempts to adjust effect estimates for selection bias.

Self-Selection: When units decide the condition they will enter.

Simple Random Assignment: Random assignment with equal probability of assignment to each condition, without use of ancillary methods such as blocking, matching, or stratification.

Single-Case Designs: A time series done on one person, common in clinical research.

Specification Error: An incorrect specification of the model presumed to have given rise to the data.

Stakeholder: Persons or groups with a stake in a treatment or the study of that treatment.

Standardized Mean Difference Statistic: An effect size measure for continuous variables, computed as the difference between two means divided by the variability of that difference.

Statistical Conclusion Validity: The validity of inferences about covariation between two variables.

Stratifying: The process of creating homogeneons groups of units in which each group has more units than there are experimental conditions. (See also *Blocking, Matching*)

Step Function: A functional relationship between two variables in which the value of one variable suddenly and completely moves from one level to another.

Testing Effects: Effects due to repeated testing of participants over time. (See also *Fatigue Effects, Practice Effects*)

Threats to Validity: Reasons why an inference might be incorrect.

Treatment Adherence: Whether the participant uses the treatment as instructed.

Treatment Attrition: Failure of units to receive treatment (whether or not they are measured).

Treatment Delivery: Whether the treatment is provided by the experimenter to the participant.

Treatment Group: In an experiment, the group that receives the intervention of interest. (See also *Comparison Group, Control Group, Placebo*)

Treatment Receipt: Whether the participant actually receives the treatment that was provided.

Trickle Process Assignment: Units to be assigned are available slowly over time.

Type I Error: Incorrectly rejecting a true null hypothesis; in an experiment, this usually implies concluding that there is an effect when there really is no effect.

Type II Error: Failing to reject a false null hypothesis; in an experiment, this usually implies concluding that there is no effect when there really is an effect. (See also *Power*)

Unit: An opportunity to apply or withhold the treatment.

Unit of Analysis Problem: Units are nested within aggregates in a way that may violate the independence assumption of many statistics. (See also *Nested Designs, Nesting*)

Unreliability: See *Reliability*

utos: An acronym to indicate the study operations that were actually done, where u = units, t = treatment, o = observations, s = setting (from Cronbach, 1982).

UTOS (Pronounced "capital utos"): An acronym to indicate generalizing to the "domain about which [the] question is asked" (Cronbach, 1982, p. 79).

***UTOS (Pronounced "star utos"):** An acronym to indicate generalizing to "units, treatments, variables, and settings not directly observed" (Cronbach, 1982, p. 83).

Validity: The truth of, correctness of, or degree of support for an inference.

Within-Participants Designs: The same units are studied in different conditions. (See also *Between-Participants Design*)

References

Abadzi, H. (1984). Ability grouping effects on academic achievement and self-esteem in a southwestern school district. *Journal of Educational Research, 77,* 287–292.

Abadzi, H. (1985). Ability grouping effects on academic achievement and self-esteem: Who performs in the long run as expected? *Journal of Educational Research, 79,* 36–39.

Abelson, R. P. (1995). *Statistics as principled argument.* Hillsdale, NJ: Erlbaum.

Abelson, R. P. (1996). Vulnerability of contrast tests to simpler interpretations: An addendum to Rosnow and Rosenthal. *Psychological Science, 7,* 242–246.

Abelson, R. P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science, 8,* 12–15.

Achen, C. H. (1986). *The statistical analysis of quasi-experiments.* Berkeley: University of California Press.

Adair, J. G. (1973). The Hawthorne effect: A reconsideration of a methodological artifact. *Journal of Applied Psychology, 69,* 334–345.

Ahlbom, A., & Norell, S. (1990). *Introduction to modern epidemiology.* Chestnut Hill, MA: Epidemiology Resources.

Ahn, C.-K. (1983). A Monte Carlo comparison of statistical methods for estimating treatment effects in regression discontinuity design (Doctoral dissertation, Washington State University, 1983). *Dissertation Abstracts International, 44*(03), 733A.

Aigner, D. J., & Hausman, J. A. (1980). Correcting for truncation bias in the analysis of experiments in time-of-day pricing of electricity. *Bell Journal, 35,* 405.

Aiken, L. S., & West, S. G. (1990). Invalidity of true experiments: Self-report pretest biases. *Evaluation Review, 14,* 374–390.

Aiken, L. S., & West, S. G. (1991). *Testing and interpreting interactions in multiple regression.* Newbury Park, CA: Sage.

Aiken, L. S., West, S. G., Schwalm, D. E., Carroll, J. L., & Hsiung, S. (1998). Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation: Efficacy of a university-level remedial writing program. *Evaluation Review, 22,* 207–244.

Albermarle Paper Co. v. *Moody,* 442 U.S. 435 (1975).

Alexander, R. A., Barrett, G. V., Alliger, G. M., & Carson, K. P. (1986). Towards a general model of uon-random sampling and the impact on population correlations: Generalization of Berkson's Fallacy and restriction of range. *British Journal of Mathematical and Statistical Psychology, 39,* 90–115.

Allen, J. P., Philliber, S., Herrling, S., & Kuperminc, G. P. (1997). Preventing teen pregnancy and academic failure: Experimental evaluation of a developmentally based approach. *Child Development, 64,* 729–742.

Allison, D. B. (1995). When is it worth measuring a covariate in a randomized trial? *Journal of Consulting and Clinical Psychology, 63,* 339–343.

Allison, D. B., Allison, R. L., Faith, M. S., Paultre, F., & Pi-Sunyer, F. X. (1997). Power and money: Designing statistically powerful studies while minimizing financial costs. *Psychological Methods, 2,* 20–33.

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behavior Research and Therapy, 31,* 621–631.

Allison, P. D. (1987). Estimation of linear models with incomplete data. In C. Clogg (Ed.), *Sociological methodology* (pp. 71–103). San Francisco: Jossey-Bass.

Allison, P. D., & Hauser, R. M. (1991). Reducing bias in estimates of linear models by remeasurement of a random subsample. *Sociological Methods and Research, 19,* 466–492.

Alwin, D. F., & Tessler, R. C. (1985). Causal models, unobserved variables, and experimental data. In H. M. Blalock (Ed.), *Causal models in panel and experimental designs* (pp. 55–88). New York: Aldine.

Amato, P. R., & Keith, B. (1991). Parental divorce and the well-being of children: A meta-analysis. *Psychological Bulletin, 110,* 26–46.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Evaluation Association. (1995). Guiding principles for evaluators. In W. R. Shadish, D. L. Newman, M. A. Scheirer, & C. Wye (Eds.), *Guiding principles for evaluators* (pp. 19–26). San Francisco: Jossey-Bass.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

American Psychiatric Association. (2000). *Handbook of psychiatric measures.* Washington, DC: Author.

American Psychological Association (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51* (Supplement).

American Psychological Association. (1992). Ethical principles of psychologists and code of conduct. *American Psychologist, 47,* 1597–1611.

American Psychological Association. (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington DC: Author.

Anastasi, A. (1968). *Psychological testing* (3rd ed.). New York: Macmillan.

Anderman, C., Cheadle, A., Curry, S., Diehr, P., Shultz, L., & Wagner, E. (1995). Selection bias related to parental consent in school-based survey research. *Evaluation Review, 19,* 663–674.

Anderson, C. A., Lindsay, J. J., & Bushman, B. J. (1999). Research in the psychological laboratory: Truth or triviality? *Current Directions in Psychological Science, 8,* 3–10.

Anderson, J. G. (1987). Structural equation models in the social and behavioral sciences: Model building. *Child Development, 58,* 49–64.

Anderson, V. L., & McLean, R. A. (1984). *Applied factorial and fractional designs.* New York: Dekker.

Angrist, J. D., & Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association, 90,* 431–442.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996a). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91,* 444–455.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996b). Rejoinder. *Journal of the American Statistical Association, 91,* 468–472.

Anson, A., Cook, T. D., Habib, F., Grady, M. K., Haynes, N. & Comer, J. P. (1991). The Comer School Development Program: A theoretical analysis. *Journal of Urban Education, 26,* 56–82.

Arbuckle, J. J. (1997). *Amos users' guide, version 3.6.* Chicago: Small Waters Corporation.

Armitage, P. (1999). Data and safety monitoring in the Concorde and Alpha trials. *Controlled Clinical Trials, 20,* 207–228.

Aronson, D. (1998). Using sibling data to estimate the impact of neighborhoods on children's educational outcomes. *Journal of Human Resources, 33,* 915–956.

Ashenfelter, O. (1978). Estimating the effects of trainiug programs on earnings. *Review of Economics and Statistics, 60,* 47–57.

Ashenfelter, O., & Card, D. (1985). Using the longitudinal structure of earnings to estimate the effect of trainiug programs. *Review of Economics and Statistics, 67,* 648–660.

Ashenfelter, O., & Krueger, A. B. (1994). Estimates of the economic returns to schooling from a new sample of twins. *American Economic Review, 84,* 1157–1173.

Atkinson, A. C. (1985). An introduction to the optimum design of experiments. In A. C. Atkinson & S. E. Fienberg (Eds.), *A celebration of statistics: The ISI centenary volume* (pp. 465–473). New York: Springer-Verlag.

Atkinson, A. C., & Donev, A. N. (1992). *Optimum experimental designs.* Oxford, England: Clarendon Press.

Atkinson, D. R., Furlong, M. J., & Wampold, B. E. (1982). Statistical significance, reviewer evaluations, and the scientific process: Is there a (statistically) significant relationship? *Journal of Counseling Psychology, 29,* 189–194.

Atkinson, R. C. (1968). Computerized instruction and the learning process. *American Psychologist, 23,* 225–239.

Atwood, J. R., & Taylor, W. (1991). Regression discontinuity design: Alternative for nursing research. *Nursing Research, 40,* 312–315.

Babcock, J. L. (1998). Retrospective pretests: Conceptual and methodological issues (Doctoral dissertation, University of Arizona, 1997). *Dissertation Abstracts International, 58*(08), 4513B.

Bagozzi, R. P., & Warshaw, P. R. (1992). An examination of the etiology of the attitude-behavior relation for goal-directed behaviors. *Multivariate Behavioral Research, 27,* 601–634.

Baker, F., & Curbow, B. (1991). The case-control study in health program evalnation. *Evaluation and Program Planning, 14,* 263–272.

Baker, S. H., & Rodriguez, O. (1979). Random time quote selection: An alternative to random selection in experimental evaluation. In L. Sechrest, S. G. West, M. A. Phillips, R.

Redner, & W. Yeaton (Eds.), *Evaluation studies review annual* (Vol. 4, pp. 185–196). Beverly Hills, CA: Sage.

Balke, A., & Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association, 92,* 1171–1176.

Ball, S., & Bogatz, G. A. (1970). *The first year of Sesame Street: An evaluation.* Princeton, NJ: Educational Testing Service.

Ballart, X., & Riba, C. (1995). Impact of legislation requiring moped and motorbike riders to wear helmets. *Evaluation and Program Planning, 18,* 311–320.

Barlow, D. H., & Hersen, M. (1984). *Single case experimental designs: Strategies for studying behavior change* (2nd ed.). New York: Pergamon Press.

Barnard, J., Du, J., Hill, J. L., & Rubin, D. R. (1998). A broader template for analyzing broken randomized experiments. *Sociological Methods and Research, 27,* 285–317.

Barnes, B. (1974). *Scientific knowledge and sociological theory.* London: Routledge & Kegan Paul.

Barnow, B. S. (1987). The impact of CETA programs on earnings: A review of the literature. *Journal of Human Resources, 22,* 157–193.

Barnow, B. S., Cain, G. G., & Goldberger, A. S. (1980). Issues in the analysis of selectivity bias. In E. W. Stromsdorfer & G. Farkas (Eds.), *Evaluation studies review annual* (Vol. 5, pp. 43–59). Beverly Hills, CA: Sage.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173–1182.

Beach, M. L., & Meier, P. (1989). Choosing covariates in the analysis of clinical trials. *Controlled Clinical Trials, 10,* 161S–175S.

Beauchamp, T. L. (Ed.). (1974). *Philosophical problems of causation.* Encino, CA: Dickenson.

Bechtel, W. (1988). *Philosophy of science: An overview for cognitive science.* Hillsdale, NJ: Erlbaum.

Beck, A. T., Ward, C. H., Mendelsohn, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4,* 561–571.

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology, 41,* 257–278.

Becker, B. J. (1992). Models of science achievement: Forces affecting male and female performance in school science. In T. D. Cook, H. M. Cooper, D. S. Cordray, H. Hartmann, L. V. Hedges, R. J. Light, T. A. Louis, & F. Mosteller (Eds.), *Meta-analysis for explanation: A casebook* (pp. 209–281). New York: Russell Sage Foundation.

Becker, B. J. (1994). Combining significance levels. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 215–230). New York: Russell Sage Foundation.

Becker, B. J., & Schram, C. M. (1994). Examining explanatory models through research synthesis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 357–381). New York: Russell Sage Foundation.

Becker, H. S. (1958). Problems of inference and proof in participant observation. *American Sociological Review, 23,* 652–660.

Becker, H. S. (1979). Do photographs tell the truth? In T. D. Cook & C. S. Reichardt (Eds.), *Qualitative and quantitative methods in evaluation research* (pp. 99–117). London: Sage.

Becker, M. H. (1992). Theoretical models of adherence and strategies for improving adherence. In S. A. Shumaker, E. B. Schron, & J. K. Onkene (Eds.), *The handbook of health behavior change* (pp. 5–43). New York: Springer.

Beecher, H. (1955). The powerful placebo. *Journal of the American Medical Association, 159,* 1602–1606.

Beecher, H. (1966). Ethics and clinical research. *New England Journal of Medicine, 274,* 1354–1360.

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., Pitkin, R., Rennie, D., Schulz, K. F., Simel, D., & Stroup, D. F. (1996). Improving the quality of reporting of randomized controlled trials. *Journal of the American Medical Association, 276,* 637–639.

Begg, C. B. (1990). Suspended judgment: Significance tests of covariate imbalance in clinical trials. *Controlled Clinical Trials, 11,* 223–225.

Begg, C. B. (1994). Publication bias. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 399–409). New York: Russell Sage Foundation.

Begg, C. B. (2000). Ruminations on the intent-to-treat principle. *Controlled Clinical Trials, 21,* 241–243.

Bell, S. H., Orr, L. L., Blomquist, J. D., & Cain, G. G. (1995). *Program applicants as a comparison group in evaluating training programs.* Kalamazoo, MI: Upjohn Institute for Employment Research.

Bentler, P. M. (1987). Drug use and personality in adolescence and young adulthood: Structural models with nonnormal variables. *Child Development, 58,* 65–79.

Bentler, P. M. (1993). *EQS/Windows user's guide.* (Available from BMDP Statistical Software, Inc., 1440 Sepulveda Blvd., Suite 316, Los Angeles, CA 90025)

Bentler, P. M. (1995). *EQS: Structural equations program manual.* Encino, CA: *Multivariate Software.*

Bentler, P. M,. & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588–606.

Bentler, P. M., & Chou, C.-P. (1988). Practical issues in structural modeling. In J. S. Long (Ed.), *Common problems/proper solutions: Avoiding error in quantitative research* (pp. 161–192). Newbury Park, CA: Sage.

Bentler, P. M., & Chou, C.-P. (1990). Model search with TETRAD II and EQS. *Sociological Methods and Research, 19,* 67–79.

Bentler, P. M., & Speckart, G. (1981). Attitudes "cause" behaviors: A structural equation analysis. *Journal of Personality and Social Psychology, 40,* 226–238.

Bentler, P. M., & Woodward, J. A. (1978). A Head Start reevaluation: Positive effects are not yet demonstrable. *Evaluation Quarterly, 2,* 493–510.

Bentler, P. M., & Wu, E. J. C. (1995). *EQS/Windows user's guide.* (Available from Multivariate Software, Inc., 4924 Balboa Blvd., #368, Encino, CA 91316)

Berg, A. T., & Vickrey, B. G. (1994). Outcomes research. *Science, 264,* 757–758.

Berg, B. L. (1989). *Qualitative research methods for the social sciences.* Boston: Allyn & Bacon.

Berger, V. W., & Exner, D. V. (1999). Detecting selection bias in randomized clinical trials. *Controlled Clinical Trials, 20,* 319–327.

Bergner, R. M. (1974). The development and evaluation of a training videotape for the resolution of marital conflict. *Dissertation Abstracts International, 34,* 3485B. (University Microfilms No. 73-32510).

Bergstralh, E., Kosanke, J., & Jocobsen, S. (1996). Software for optimal matching in observational studies. *Epidemiology, 7,* 331–332.

Berk, R. A., & DeLeeuw, J. (1999). An evaluation of California's inmate classification system using a generalized regression discontinuity design. *Journal of the American Statistical Association, 94,* 1045–1052.

Berk, R. A., Lenihan, K. J., & Rossi, P. H. (1980). Crime and poverty: Some experimental evidence from ex-offenders. *American Sociological Review, 45,* 766–786.

Berk, R. A., & Rauma, D. (1983). Capitalizing on nonrandom assignment to treatment: A regression discontinuity evaluation of a crime control program. *Journal of the American Statistical Association, 78,* 21–27.

Berk, R. A., Smyth, G. K., & Sherman, L. W. (1988). When random assignment fails: Some lessons from the Minneapolis Spouse Abuse Experiment. *Journal of Quantitative Criminology, 4,* 209–223.

Berman, J. S., & Norton, N. C. (1985). Does professional training make a therapist more effective? *Psychological Bulletin, 98,* 401–407.

Berman, P., & McLaughlin, M. W. (1977). *Federal programs supporting educational change: Vol. 8. Factors affecting implementation and continuation.* Santa Monica, CA: RAND.

Besadur, M., Graen, G. B., & Scandura, T. A. (1986). Training effects on attitudes toward divergent thinking among manufacturing engineers. *Journal of Applied Psychology, 71,* 612–617.

Beutler, L. E., & Crago, M. (Eds.). (1991). *Psychotherapy research: An international review of programmatic studies.* Washington, DC: American Psychological Association.

Bhaskar, R. (1975). *A realist theory of science.* Leeds, England: Leeds.

Bickman, L. (1985). Randomized field experiments in education: Implementation lessons. In R. F. Boruch & W. Wothke (Eds.), *Randomization and field experimentation* (pp. 39–53). San Francisco: Jossey-Bass.

Biglan, A., Hood, D., Brozovsky, P., Ochs, L., Ary, D., & Black, C. (1991). Subject attrition in prevention research. In C. G. Leukfeld & W. Bukowski (Eds.), *Drug use prevention intervention research: Methodological issues* (NIDA Research Monograph 107, DHHS Publication No. 91-1761, pp. 213–228). Rockville, MD: U.S. Government Printing Office.

Biglan, A., Metzler, C. W., & Ary, D. V. (1994). Increasing the prevalence of successful children: The case for community intervention research. *Behavior Analyst, 17,* 335–351.

Birnbaum, A. (1961). Confidence curves: An omnibus technique for estimation and testing statistical hypotheses. *Journal of the American Statistical Association, 56,* 246-249.

Bishop, R. C., & Hill, J. W. (1971). Effects of job enlargement and job change on contiguous but non-manipulated jobs as a function of worker's status. *Journal of Applied Psychology, 55,* 175–181.

Blackburn, H., Luepker, R., Kline, F. G., Bracht, N., Carlaw, R., Jacos, D., Mittelmark, M., Stauffer, L., & Taylor, H. L. (1984). The Minnesota Heart Health Program: A research and demonstration project in cardiovascular disease prevention. In J. D. Matarazzo, S. Weiss, J. A. Herd, N. E. Miller, & S. M. Weiss (Eds.), *Behavioral health* (pp. 1171–1178). New York: Wiley.

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: Handbook I. The cognitive domain.* New York: McKay.

Bloom, H. S. (1984a). Accounting for no-shows in experimental evaluation designs. *Evaluation Review, 8,* 225–246.

Bloom, H. S. (1984b). Estimating the effect of job-training programs, using longitudinal data: Ashenfelter's findings reconsidered. *Journal of Human Resources, 19,* 544–556.

Bloom, H. S. (1990). *Back to work: Testing reemployment services for displaced workers.* Kalamazoo, MI: Upjohn Institute.

Bloom, H. S., & Ladd, H. F. (1982). Property tax revaluation and tax levy growth. *Journal of Urban Economics, 11,* 73–84.

Bloor, D. (1976). *Knowledge and social imagery.* London: Routledge & Kegan Paul.

Bloor, D. (1997). Remember the strong program? *Science, Technology, and Human Values, 22,* 373–385.

Bock, R. D. (Ed.). (1989). *Multilevel analysis of educational data.* San Diego, CA: Academic Press.

Boissel, J. P., Blanchard, J., Panak, E., Peyrieux, J. C., & Sacks, H. (1989). Considerations for the meta-analysis of randomized clinical trials. *Controlled Clinical Trials, 10,* 254–281.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Bollen, K. A. (1990). Outlier screening and a distribntion-free test for vanishing tetrads. *Sociological Methods and Research, 19,* 80–92.

Boomsma, A. (1987). The robnstness of maximum likelihood estimation in structural eqnation models. In P. Cntrance & R. Ecob (Eds.), *Structural modeling by example: Applications in educational, sociological, and behavioral research* (pp. 160–188). Cambridge, England: Cambridge University Press.

Borenstein, M., & Cohen, J. (1988). *Statistical power analysis: A computer program.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Borenstein, M., Cohen, J., & Rothstein, H. (in press). *Confidence intervals, effect size, and power* [Compnter program]. Hillsdale, NJ: Erlbaum.

Borenstein, M., & Rothstein, H. (1999). *Comprehensive meta-analysis.* Englewood, NJ: Biostat.

Borkovec, T. D., & Nan, S. D. (1972). Credibility of analogne therapy rationales. *Journal of Behavior Therapy and Experimental Psychiatry, 3,* 257–260.

Boruch, R. F. (1975). Conpling randomized experiments and approximations ro experiments in social program evaluation. *Sociological Methods and Research, 4,* 31–53.

Boruch, R. F. (1982). Experimental tests in edncation: Recommendations from the Holtzman Report. *American Statistician, 36,* 1–8.

Boruch, R. F. (1997). *Randomized field experiments for planning and evaluation: A practical guide.* Thousand Oaks, CA: Sage.

Boruch, R. F., & Cecil, J. S. (1979). *Assuring the confidentiality of social research data.* Philadelphia: University of Pennsylvania Press.

Boruch, R. F., Dennis, M., & Carter-Greer, K. (1988). Lessons from the Rockefeller Foundation's experiments on the Minority Female Single Parent program. *Evaluation Review, 12,* 396–426.

Boruch, R. F., & Foley, E. (2000). The honestly experimental society: Sites and other entities as the units of allocation and analysis in randomized trials. In L. Bickman (Ed.), *Validity and social experimentation: Donald Campbell's legacy* (Vol. 1, pp. 193–238). Thousand Oaks, CA: Sage.

Boruch, R. F., & Gomez, H. (1977). Sensitivity, bias, and theory in impact evaluations. *Professional Psychology, 8,* 411–434.

Boruch, R. F., & Wothke, W. (1985). Seven kinds of randomization plans for designing field experiments. In R. F. Boruch & W. Wothke (Eds.), *Randomization and field experimentation* (pp. 95–118). San Francisco: Jossey-Bass.

Bos, H., Huston, A., Granger, R., Duncan, G., Brock, T., & McLoyd, V. (1999, April). *New hope for people with low incomes: Two-year results of a program to reduce poverty and reform welfare.* New York: Manpower Research Development.

Box, G. E. P., & Draper, N. R. (1987). *Empirical model-building and response surfaces.* New York: Wiley.

Box, G. E. P., Hunter, W. G., & Hunter, J. S. (1978). *Statistics for experimenters: An introduction to design, data analysis, and model building.* New York: Wiley.

Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis: Forecasting and control.* San Francisco: Holden-Day.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). *Time series analysis: Forecasting and control* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Bracht, G. H., & Glass, G. V. (1968). The external validity of experiments. *American Educational Research Journal, 5,* 437–474.

Bradburn, N. M. (1983). Response effects. In P. H. Rossi, J. D. Wright, & A. B. Anderson (Eds.), *Handbook of survey research* (pp. 289–328). San Diego, CA: Academic Press.

Braden, J. P., & Bryant, T. J. (1990). Regression discontinuity designs: Applications for school psychologists. *School Psychology Review, 19,* 232–239.

Bramel, D., & Friend, R. (1981). Hawthorne, the myth of the docile worker, and class bias in psychology. *American Psychologist, 36,* 867–878.

Braught, G. N., & Reichardt, C. S. (1993). A computerized approach to trickle-process, random assignment. *Evaluation Review, 17,* 79–90.

Braunholtz, D. A. (1999). A note on Zelen randomization: Attitudes of parents participating in a neonatal clinical trial. *Controlled Clinical Trials, 20,* 569–571.

Braver, M. C. W., & Braver, S. L. (1988). Statistical treatment of the Solomon Four-Group design: A meta-analytic approach. *Psychological Bulletin, 104,* 150–154.

Braver, S. L., & Smith, M. C. (1996). Maximizing both external *and* internal validity in longitudinal true experiments with voluntary treatments: The "combined modified" design. *Evaluation and Program Planning, 19,* 287–300.

Breger, M. J. (1983). Randomized social experiments and the law. In R. F. Boruch & J. S. Cecil (Eds.), *Solutions to ethical and legal problems in social research* (pp. 97–144). New York: Academic Press.

Breslau, D. (1997). Contract shop epistemology: Credibility and problem construction in applied social science. *Social Studies of Science, 27,* 363–394.

Brockwell, P. J., & Davis, R. A. (1991). *Time series: Theory and methods* (2nd ed.). New York: Springer-Verlag.

Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by nature and design.* Cambridge, Massachusetts: Harvard University Press.

Brook, J. S., Cohen, P., & Gordon, A. S. (1983). Longitudinal study of adolescent drug use. *Psychological Reports, 53,* 375–378.

Brown, R. (1986). *Social psychology: The second edition.* New York: Free Press.

Brown, H. I. (1977). *Perception, theory and commitment: The new philosophy of science.* Chicago: University of Chicago Press.

Brown, H. I. (1989). Toward a cognitive psychology of *What? Social Epistemology, 3,* 129–138.

Brunette, D. (1995). Natural disasters and commercial real estate returns. *Real Estate Finance, 11,* 67–72.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley: University of California Press.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage.

Bryk, A. S., Raudenbush, S. W., & Congdon, R. T. (1996). *HLM: Hierarchical linear modeling with the HLM/2L and HLM/3L programs.* (Available from Scientific Software International, 1525 E. 53rd Street, Suite 530, Chicago IL 60615)

Bunge, M. (1959). *Causality and modern science* (3rd ed.). New York: Dover.

Bunge, M. (1992). A critical examination of the new sociology of science (Part 2). *Philosophy of the Social Sciences, 22,* 46–76.

Burger, T. (in press). Ideal type: Understandings in the social sciences. In N. Smelser & P. Baltes (Eds.), *Encyclopedia of the behavioral and social sciences.* Amsterdam: Elsevier.

Burtless, G. (1995). The case for randomized field trials in economic and policy research. *Journal of Economic Perspectives, 9,* 63–84.

Byrne, B. (1989). *A primer of LISREL.* New York: Springer-Verlag.

Byrne, B. (1994). *Structural equation modeling with EQS and EQS/Windows: Basic concepts, applications, and programming.* Newbury Park, CA: Sage.

Cahan, S., & Davis, D. (1987). A between-grade-levels approach to the investigation of the absolute effects of schooling on achievement. *American Educational Research Journal, 24,* 1–12.

Cahan, S., Linchevski, L., Ygra, N., & Danziger, I. (1996). The cumulative effect of ability grouping on mathematical achievement: A longitudinal perspective. *Studies in Educational Evaluation, 22,* 29–40.

Cain, G. G. (1975). Regression and selection models to improve nonexperimental comparisons. In C. A. Bennett & A. A. Lumsdaine (Eds)., *Evaluation and experiment: Some critical issues in assessing social programs* (pp. 297–317). New York: Academic Press.

Caines, P. E. (1988). *Linear stochastic systems.* New York: Wiley.

Campbell, D. T. (1956). *Leadership and its effects on groups* (Ohio Studies in Personnel, Bureau of Business Research Monograph No. 83). Columbus: Ohio State University.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54,* 297–312.

Campbell, D. T. (1963). From description to experimentation: Interpreting trends as quasi-experiments. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 212–243). Madison: University of Wisconsin Press.

Campbell, D. T. (1966a). Pattern matching as an essential in distal knowing. In K. R. Hammond (Ed.), *The psychology of Egon Brunswik.* New York: Holt, Rinehart, & Winston.

Campbell, D. T. (1966b). *The principle of proximal similarity in the application of science.* Unpublished manuscript, Northwestern University.

Campbell, D. T. (1969a). Prospective: Artifact and control. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research* (pp. 351–382). New York: Academic Press.

Campbell, D. T. (1975). "Degrees of freedom" and the case study. *Comparative Political Studies, 8,* 178–193.

Campbell, D. T. (1976). Focal local indicators for social program evaluation. *Social Indicators Research, 3,* 237–256.

Campbell, D. T. (1978). Qualitative knowing in action research. In M. Brenner & P. Marsh (Eds.), *The social contexts of method* (pp. 184–209). London: Croom Helm.

Campbell, D. T. (1982). Experiments as arguments. In E. R. House (Ed.), *Evaluation studies review annual* (Volume 7, pp. 117–127). Newbury Park, CA: Sage.

Campbell, D. T. (1984). Foreword. In W. M. K. Trochim, *Research design for program evaluation: The regression discontinuity approach* (pp. 15–43). Beverly Hills, CA: Sage.

Campbell, D. T. (1986). Relabeling internal and external validity for applied social scientists. In W. M. K. Trochim (Ed.), *Advances in quasi-experimental design and analysis* (pp. 67–77). San Francisco: Jossey-Bass.

Campbell, D. T. (1988). *Methodology and epistemology for social science: Selected papers* (E. S. Overman, Ed.). Chicago: University of Chicago Press.

Campbell, D. T., & Boruch, R. F. (1975). Making the case for randomized assigument to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects. In C. A. Bennett & A. A. Lumsdaine (Eds.), *Evaluation and experiments: Some critical issues in assessing social programs* (pp. 195–296). New York: Academic Press.

Campbell, D. T., & Erlebacher, A. E. (1970). How regression artifacts can mistakenly make compensatory education programs look harmful. In J. Hellmuth (Ed.), *The Disadvantaged Child: Vol. 3, Compensatory education: A national debate* (pp. 185–210). New York: Brunner/Mazel.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts.* New York: Guilford Press.

Campbell, D. T., & Russo, M. J. (1999). *Social experimentation.* Thousand Oaks, CA: Sage.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: RandMcNally.

Campbell, F. A., & Ramey, C. T. (1995). Cognitive and school outcomes for high-risk African American students at middle adolescence: Positive effects of early intervention. *American Educational Research Journal, 32,* 743–772.

Camstra, A., & Boomsma, A. (1992). Cross-validation in regression and covariance structure analysis. *Sociological Methods and Research, 21,* 89–115.

Canner, P. (1984). How much data should be collected in a clinical trial? *Statistics in Medicine, 3,* 423–432.

Canner, P. (1991). Covariate adjustment of treatment effects in clinical trials. *Controlled Clinical Trials, 12,* 359–366.

Capaldi, D., & Patterson, G. R. (1987). An approach to the problem of recruitment and retention rates for longitudinal research. *Behavioral Assessment, 9,* 169–177.

Cappelleri, J. C. (1991). *Cutoff-based designs in comparison and combination with randomized clinical trials.* Unpublished doctoral dissertation, Cornell University, Ithaca, New York.

Cappelleri, J. C., Darlington, R. B., & Trochim, W. M. K. (1994). Power analysis of cutoff-based randomized clinical trials. *Evaluation Review, 18,* 141–152.

Cappelleri, J. C., Ioannidis, J. P. A., Schmid, C. H., deFerranti, S. D., Aubert, M., Chalmers, T. C., & Lau, J. (1996). Large trials vs meta-analysis of smaller trials: How do their results compare? *Journal of the American Medical Association, 276,* 1332–1338.

Cappelleri, J. C., & Trochim, W. M. K. (1992, May). *An illustrative statistical analysis of cutoff-based randomized clinical trials.* Paper presented at the annual meeting of the Society for Clinical Trials, Philadelphia, PA.

Cappelleri, J. C., & Trochim, W. M. K. (1994). An illustrative statistical analysis of cutoff-based randomized clinical trials. *Journal of Clinical Epidemiology, 47,* 261–270.

Cappelleri, J. C., & Trochim, W. M. K. (1995). Ethical and scientific features of cutoff-based designs of clinical trials. *Medical Decision Making, 15,* 387–394.

Cappelleri, J. C., Trochim, W. M. K., Stanley, T. D., & Reichardt, C. S. (1991). Random measurement error does not bias the treatment effect estimate in the regression-discontinuity design: I. The case of no interaction. *Evaluation Review, 15,* 395–419.

Carbonari, J. P., Wirtz, P. W., Muenz, L. R., & Stout, R. L. (1994). Alternative analytical methods for detecting matching effects in treatment outcomes. *Journal of Studies on Alcohol* (Suppl. 12), 83–90.

Card, D. (1990). The impact of the Mariel Boatlift on the Miami labor market. *Industrial and Labor Relations Review,* 43, 245–257.

Carrington, P. J., & Moyer, S. (1994). Gun availability and suicide in Canada: Testing the displacement hypothesis. *Studies on Crime and Crime Prevention, 3,* 168–178.

Carter, G. M., Winkler, J. D., & Biddle, A. K. (1987). *An evaluation of the NIH research career development award.* Santa Monica, CA: RAND.

Casella, G., & Schwartz, S.P. (2000). Comment. *Journal of the American Statistical Association, 95,* 425-428.

Catalano, R., & Serxner, S. (1987). Time series designs of potential interest to epidemiologists. *American Journal of Epidemiology, 126,* 724–731.

Cecil, J. S., & Boruch, R. F. (1988). Compelled disclosure of research data: An early warning and suggestions for psychologists. *Law and Human Behavior, 12,* 181–189.

Chaffee, S. H., Roser, C., & Flora, J. (1989). Estimating the magnitude of threats to validity of information campaign effects. In C. G. Salmon (Ed.), *Annual review of communication research* (Vol. 18). Newbury Park, CA: Sage.

Chalmers, I., Enkin, M., & Keirse, M. J. (Eds.). (1989). *Effective care in pregnancy and childbirth.* New York: Oxford University Press.

Chalmers, T. C. (1968). Prophylactic treatment of Wilson's disease. *New England Journal of Medicine, 278,* 910–911.

Chalmers, T. C., Berrier, J., Hewitt, P., Berlin, J., Reitman, D., Nagalingam, R., & Sacks, H. (1988). Meta-analysis of randomized controlled trials as a method of estimating rare complications of non-steroidal anti-inflammatory drug therapy. *Alimentary and Pharmacological Therapy, 2–5,* 9–26.

Chalmers, T. C., Celano, P., Sacks, H. S., & Smith, H. (1983). Bias in treatment assignment in controlled clinical trials. *New England Journal of Medicine, 309,* 1358–1361.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically snpported therapies. *Journal of Consulting and Clinical Psychology, 66,* 7–18.

Chan, K.-C., & Tumin, J. R. (1997). Evaluating the U.S. nuclear triad. In E. Chelimsky & W. R. Shadish (Eds.), *Evaluation for the 21st century: A handbook* (pp. 284–298). Thousand Oaks, CA: Sage.

Chaplin, W. F. (1991). The next generation of moderator research in personality psychology. *Journal of Personality, 59,* 143–178.

Chaplin, W. F. (1997). Personality, interactive relations, and applied psychology. In S. R. Briggs, R. Hogan., & W. H. Jones (Eds.), *Handbook of personality psychology* (pp. 873–890). Orlando, FL: Academic Press.

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology, 74,* 271–280.

Chelimsky, E. (1998). The role of experience in formulating theories of evaluation practice. *American Journal of Evaluation, 19,* 35–55.

Chen, H., & Rossi, P. H. (1987). The theory-driven approach to validity. *Evaluation and Program Planning, 10,* 95–103.

Chen, H.-T., & Rossi, P. H. (Eds.). (1992). *Using theory to improve program and policy evaluations.* New York: Greenwood Press.

Choi, S. C., & Pepple, P. A. (1989). Monitoring clinical trials based on predictive probability of significance. *Biometrics, 45,* 317–323.

Choi, S. C., Smith, P. J., & Becker, D. P. (1985). Early decision in clinical trials when treatment differences are small: Experience of a controlled trial in head trauma. *Controlled Clinical Trials, 6,* 280–288.

Ciarlo, J. A., Brown, T. R., Edwards, D. W., Kiresuk, T. J., & Newman, F. L. (1986). *Assessing mental health treatment outcome measurement techniques* (DHHS Publication No. ADM 86-1301). Washington, DC: U.S. Government Printing Office.

Cicirelli, V. G., and Associates. (1969). *The impact of Head Start: An evaluation of the effects of Head Start on children's cognitive and affective development: Vols. 1–2.* Athens: Ohio University and Westinghouse Learning Corporation.

Clark, P. I., & Leaverton, P. E. (1994). Scientific and ethical issues in the use of placebo controls in clinical trials. *Annual Review of Public Health, 15,* 19–38.

Clarridge, B. R., Sheehy, L. L., & Hauser, T. S. (1977). Tracing members of a panel: A 17-year follow-up. In K. F. Schuessler (Ed.), *Sociological methodology* (pp. 185–203). San Francisco: Jossey-Bass.

Cochran, W. G. (1965). The planning of observational studies in human populations. *Journal of the Royal Statistical Society* (*Series A*), *128,* 134–155.

Cochran, W. G. (1968). The effectiveness of adjustment by snbclassification in removing bias in observational studies. *Biometrics, 24,* 295–313.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.

Cochran, W. G. (1983). *Planning and analysis of observational studies.* New York: Wiley.

Cochran, W. G., & Cox, G. M. (1957). *Experimental designs* (2nd ed.). New York: Wiley.

Cochran, W. G., & Rubin, D. B. (1973). Controlling bias in observational stndies: A review. *Sankhyā, 35,* 417–446.

Cohen, E., Mowbray, C. T., Bybee, D., Yeich, S., Ribisl, K., & Freddolino, P. P. (1993). Tracking and follow-up methods for research on homelessness. *Evaluation Review, 17,* 331–352.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale NJ: Lawrence Erlbaum Associates.

Cohen, J. (1994). The earth is ronnd ($p<.05$). *American Psychologist, 49,* 997–1003.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbanm.

Cohen, P., Cohen, J., Teresi, J., Marchi, M., & Velez, C. N. (1990). Problems in the measurement of latent variables in structural equation cansal models. *Applied Psychological Measurement, 14,* 183–196.

Colditz, G. A., Miller, J. N., & Mosteller, F. (1988). The effect of stndy design on gain in evalnation of new treatments in medicine and surgery. *Drug Information Journal, 22,* 343–352.

Collins, H. M. (1981). Stages in the empirical programme of relativism. *Social Studies of Science, 11,* 3–10.

Collins, J. F., & Elkin, I. (1985). Randomization in the NIMH Treatment of Depression Collaborative Research Program. In R. F. Boruch & W. Wothke (Eds.), *Randomization and field experimentation* (pp. 27–37). San Francisco: Jossey-Bass.

Comer, J. P. (1988). Educating poor minority children. *Scientific American, 259,* 42–48.

Connell, D. B., Turner, R. R., & Mason, E. F. (1985). Summary of findings of the school health education evaluation: Health promotion effectiveness, implementation and costs. *Journal of School Health, 55,* 316–321.

Connell, J. P., Kubisch, A. C., Schorr, L. B., & Weiss, C. H. (Eds.). (1995). *New approaches to evaluating community initiatives: Concepts, methods and contexts.* Washington, DC: Aspen Institute.

Conner, R. F. (1977). Selecting a control group: An analysis of the randomization process in twelve social reform programs. *Evaluation Quarterly, 1,* 195–244.

Connor, S. (1989). *Postmodernist culture: An introduction to theories of the contemporary.* Oxford, England: Basil Blackwell.

Connors, A. F., Speroff, T., Dawson, N. V., Thomas, C., Harrell, F. E., Wagner, D., Desbiens, N., Goldman, L., Wu, A. W., Califf, R. M., Fulkerson, W. J ., Vidaillet, H., Broste, S., Bellamy, P., Lynn, J., & Knaus, W. A. (1996). The effectiveness of right heart catheterization in the initial care of critically ill patients. *Journal of the American Medical Association, 276,* 889–897.

Conrad, K. J. (Ed.). (1994). *Critically evaluating the role of experiments.* San Francisco: Jossey-Bass.

Conrad, K. J., & Conrad, K. M. (1994). Reassessing validity threats in experiments: Focus on construct validity. In K. J. Conrad (Ed.), *Critically evaluating the role of experiments* (pp. 5–25). San Francisco: Jossey-Bass.

Cook, R. D., & Weisberg, S. (1994). *An introduction to regression graphics.* New York: Wiley.

Cook, T. D. (1984). What have black children gained academically from school integration? Examination of the meta-analytic evidence. In T. D. Cook, D. Armor, R. Crain, N. Miller, W. Stephan, H. Walberg, & P. Wortman (Eds.), *School desegregation and black achievement* (pp. 6–67). Washington, DC: National Institute of Education. (ERIC Document Reproduction Service No. ED 241 671)

Cook, T. D. (1985). Postpositivist critical multiplism. In L. Shotland & M. M. Mark (Eds.), *Social science and social policy* (pp. 21–62). Newbury Park, CA: Sage.

Cook, T. D. (1990). The generalization of causal connections: Multiple theories in search of clear practice. In L. Sechrest, E. Perrin, & J. Bunker (Eds.), *Research methodology: Strengthening causal interpretations of nonexperimental data* (DHHS Publication No. PHS 90-3454, pp. 9–31). Rockville, MD: Department of Health and Human Services.

Cook, T. D. (1991). Clarifying the warrant for generalized causal inferences in quasi-experimentation. In M. W. McLaughlin & D. C. Phillips (Eds.), *Evaluation and education: At quarter-century* (pp. 115–144). Chicago: National Society for the Study of Education.

Cook, T. D. (2000). The false choice between theory-based evaluation and experimentation. In P. J. Rogers, T. A. Hasci, A. Petrosino, & T. A. Huebner (Eds.), *Program theory in evaluation: Challenges and opportunities* (pp. 27–34). San Francisco, CA: Jossey-Bass.

Cook, T. D., Appleton, H., Conner, R. F., Shaffer, A., Tamkin, G., & Weber, S. J. (1975). *"Sesame Street" revisited.* New York: Russell Sage Foundation.

Cook, T. D., Calder, B. J., & Wharton, J. D. (1978). *How the introduction of television affected a variety of social indicators* (Vols. 1–4). Arlington, VA: National Science Foundation.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand-McNally.

Cook, T. D., Cooper, H., Cordray, D. S., Hartmann, H., Hedges, L. V., Light, R. J., Louis, T. A., & Mosteller, F. (Eds.). (1992). *Meta-analysis for explanation: A casebook.* New York: Russell Sage Foundation.

Cook, T. D., Gruder, C. L., Hennigan, K. M., & Flay, B. R. (1979). History of the sleeper effect: Some logical pitfalls in accepting the null hypothesis. *Psychological Bulletin, 86,* 662–679.

Cook, T. D., Habib, F. N., Phillips, M., Settersten, R. A., Shagle, S. C., & Degirmencioglu, S. M. (1999). Comer's School Development Program in Prince George's County, Maryland: A theory-based evaluation. *American Educational Research Journal, 36,* 543–597.

Cook, T. D., Hunt, H. D., & Mnrphy R. F. (2000). Comer's School Development Program in Chicago: A theory-based evalnation. *American Educational Research Journal, 37,* 535–597.

Cook, T. D., & Shadish, W. R. (1994). Social experiments: Some developments over the past 15 years. *Annual Review of Psychology, 45,* 545–580.

Cook, T. D., Shagle, S. C., & Degirmenciogln, S. M. (1997). Capturing social process for testing mediational models of neighborhood effects. In J. Brooks-Gunn, G. J. Dnncan, & J. L. Aber (Eds.), *Neighborhood poverty: Context and consequences for children* (Vol. 2). New York: Russell Sage Fonndation.

Cooper, H. (1998). *Synthesizing research: A guide for literature reviews* (3rd ed.). Thousand Oaks, CA: Sage.

Cooper, H., & Hedges, L. V. (Eds.). (1994a). *The handbook of research synthesis.* New York: Russell Sage Foundation.

Cooper, H., & Hedges, L. V. (1994b). Research synthesis as a scientific enterprise. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 3–14). New York: Russell Sage Fonndation.

Cooper, W. H., & Richardson, A. J. (1986). Unfair comparisons. *Journal of Applied Psychology, 71,* 179–184.

Coover, J. E., & Angell, F. (1907). General practice effect of special exercise. *American Journal of Psychology, 18,* 328–340.

Copas, J., & Li, H. (1997). Inference for non-random samples (with discussion). *Journal of the Royal Statistical Society* (Series B), *59,* 55–95.

Cordray, D. S. (1986). Qnasi-experimental analysis: A mixture of methods and judgment. In W. M. K. Trochim (Ed.), *Advances in quasi-experimental design and analysis* (pp. 9–27). San Francisco: Jossey-Bass.

Corrin, W. J., & Cook, T. D. (1998). Design elements of quasi-experimentation. *Advances in Educational Productivity, 7,* 35–57.

Cosgrove, N., Borhani, N. O., Bailey, G., Borhani, P., Levin, J., Hoffmeier, M., Krieger, S., Lovato, L. C., Petrovitch, H., Vogt, T., Wilson, A. C., Breeson, V., Probstfield, J. L., and the Systolic Hypertension in the Elderly Program (SHEP) Cooperative Research Group. (1999). Mass mailing and staff experience in a total recruitment program for a clinical trial: The SHEP experience. *Controlled Clinical Trials, 19,* 133–148.

Costanza, M. C. (1995). Matching. *Preventive Medicine, 24,* 425–433.

Cowles, M. (1989). *Statistics in psychology: An historical perspective.* Hillsdale, NJ: Erlbanm.

Cox, D. R. (1958). *Planning of experiments.* New York: Wiley.

Coyle, S. L., Boruch, R. F., & Turner, C. F. (Eds.). (1991). *Evaluating AIDS prevention programs* (Expanded ed.). Washington, DC: National Academy Press.

Cramer, D. (1990). Self-esteem and close relationships: A sratistical refinement. *British Journal of Social Psychology, 29,* 189–191.

Cramer, J. A., & Spilker, B. (Eds.). (1991). *Patient compliance in medical practice and clinical trials.* New York: Raven Press.

Critelli, J. W., & Neumann, K. F. (1984). The placebo: Conceptual analysis of a construct in transition. *American Psychologist, 39,* 32–39.

Crocker, J. (1981). Judgment of covariation by social perceivers. *Psychological Bulletin, 90,* 272–292.

Cromwell, J. B., Hannan, M. J., Labys, W. C., & Terraza, M. (1994). *Multivariate tests for time series models.* Thousand Oaks, CA: Sage.

Cromwell, J. B., Labys, W. C., & Terraza, M. (1994). *Univariate tests for time series models.* Thousand Oaks, CA: Sage.

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement, 6,* 475–494.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs.* San Francisco: Jossey-Bass.

Cronbach, L. J. (1986). Social inquiry by and for earthlings. In D. W. Fiske & R. A. Shweder (Eds.), *Metatheory in social science* (pp. 83–107). Chicago: University of Chicago Press.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement, theory and public policy* (pp. 147–171). Urbana: University of Illinois Press.

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. D., Hornik, R. C., Phillips, D. C., Walker, D. F., & Weiner, S. S. (1980). *Toward reform of program evaluation.* San Francisco: Jossey-Bass.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1967). *The dependability of behavioral measurements: Multifacet studies of generalizability.* Stanford, CA: Stanford University Press.

Cronbach, L. J., Rogosa, D. R., Floden, R. E., & Price, G. G. (1977). *Analysis of covariance in nonrandomized experiments: Parameters affecting bias* (Occasional Paper). Palo Alto, CA: Stanford University, Stanford Evaluation Consortium.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions.* New York: Irvingron.

Crosbie, J. (1993). Interrupted time-series analysis with brief single-subject data. *Journal of Consulting and Clinical Psychology, 61,* 966–974.

Cullen, K. W., Koehly, L. M., Anderson, C., Baranowski, T., Prokhorov, A., Basen-Engquist, K., Wetter, D., & Hergenroeder, A. (1999). Gender differences in chronic disease risk behaviors through the transition out of high school. *American Journal of Preventive Medicine, 17,* 1–7.

Cunningham, W. R. (1991). Issues in factorial invariance. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions* (pp. 106–113). Washington, DC: American Psychological Association.

Currie, J., & Duncan, T. (1995). Does Head Start make a difference? *American Economic Review, 85,* 341–364.

Currie, J., & Duncan, T. (1999). Does Head Start help Hispanic children? *Journal of Public Economics, 74,* 235–262.

D'Agostino, R. B., & Kwan, H. (1995). Measuring effectiveness: What to expect without a randomized control group. *Medical Care, 33* (Suppl.), AS95–AS105.

D'Agostino, R. B., & Rubin, D. B. (2000). Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association, 95,* 749–759.

Dallmayr, F. R., & McCarthy, T. A. (Eds.). (1977). *Understanding and social inquiry.* Notre Dame, IN: University of Notre Dame Press.

Danziger, K. (1990). *Constructing the subject: Historical origins of psychological research.* Cambridge, England: Cambridge University Press.

Datta, L.-E. (1997). Multimethod evaluations: Using case studies together with other methods. In E. Chelimsky & W. R. Shadish (Eds.), *Evaluation for the 21st century: A handbook* (pp. 344–359). Thousand Oaks, CA: Sage.

Davidson, D. (1967). Causal relations. *Journal of Philosophy, 64,* 691–703.

Davies, P. (1984). *Superforce: The search for a grand unified theory of nature.* New York: Simon and Schuster.

Davies, P. C. W., & Brown, J. R. (Eds.). (1986). *The ghost in the atom? A discussion of the mysteries of quantum physics.* Cambridge, England: Cambridge University Press.

Davis, C. E. (1994). Generalizing from clinical trials. *Controlled Clinical Trials, 15,* 11–14.

Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association, 95,* 407–448.

Day, N. A., Dunt, D. R., & Day, S. (1995). Maximizing response to surveys in health program evaluation at minimum cost using multiple methods: Mail, telephone, and visit. *Evaluation Review, 19,* 436–450.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association, 94,* 1053–1062.

Dehue, T. (2000). From deception trials to control reagents: The introduction of the control group about a century ago. *American Psychologist, 55,* 264–268.

DeLeeuw, J., & Kreft, I. (1986). Random coefficient models for multilevel analysis. *Journal of Educational Statistics, 11,* 57–85.

Della-Piana, G. M. (1981). Film criticism. In N. L. Smith (Ed.), *New techniques for evaluation* (pp. 274–286). Newbury Park, CA: Sage.

Delucchi, K. L. (1994). Methods for the analysis of binary outcome results in the presence of missing data. *Journal of Consulting and Clinical Psychology, 62,* 569–575.

Delucchi, K. L., & Bostrom, A. (1999). Small sample longitudinal clinical trials with missing data: A comparison of analytic methods. *Psychological Methods, 4,* 158–172.

Deluse, S. R. (1999). Mandatory divorce education: A program evaluation using a "quasi-random" regression discontinuity design (Doctoral dissertation, Arizona State University, 1999). *Dissertation Abstracts International, 60*(03), 1349B.

Dennis, M. L. (1988). *Implementing randomized field experiments: An analysis of criminal and civil justice research.* Unpublished doctoral dissertation, Northwestern University.

Dennis, M. L., Lennox, R. D., & Foss, M. A. (1997). Practical power analysis for substance abuse health services research. In K. L. Bryant, M. Windell, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 367–404). Washington, DC: American Psychological Association.

Denton, F. T. (1985). Data mining as an industry. *Review of Economics and Statistics, 67,* 124–127.

Denton, T. (1994). Kinship, marriage and the family: Eight time series, 35000 B.C. to 2000 A.D. *International Journal of Comparative Sociology, 35,* 240–251.

Denzin, N. (1989). *The research act: A theoretical introduction to sociological methods.* Englewood Cliffs, NJ: Prentice-Hall.

Denzin, N. K., & Lincoln, Y. S. (2000). *Handbook of qualitative research* (2nd ed.). Newbury Park, CA: Sage.

Devine, E. C. (1992). Effects of psychoeducational care with adult surgical patients: A theory-probing meta-analysis of intervention studies. In T. Cook, H. Cooper, D. Cordray, H. Hartmann, L. Hedges, R. Light, T. Louis, & F. Mosteller (Eds.), *Meta-analysis for explanation: A casebook* (pp. 35–82). New York: Russell Sage Foundation.

Devine, E. C., & Cook, T. D. (1983). A meta-analytic analysis of effects of psychoeducational interventions on length of post-surgical hospital stay. *Nursing Research, 32,* 267–274.

Devine, E. C., & Cook, T. D. (1986). Clinical and cost-saving effects of psychoeducational interventions with surgical patients: A meta-analysis. *Research in Nursing and Health, 9,* 89–105.

Devine, E. C., O'Connor, F. W., Cook, T. D., Wenk, V. A., & Curtin, T. R. (1988). Clinical and financial effects of psychoeducational care provided by staff nurses to adult surgical patients in the post-DRG environment. *American Journal of Public Health, 78,* 1293–1297.

Devlin, B. (Ed.). (1997). *Intelligence and success. Is it all in the genes?: Scientists respond to The Bell Curve.* New York: Springer-Verlag.

Diament, C., & Colletti, G. (1978). Evaluation of behavioral group counseling for parents of learning-disabled children. *Journal of Abnormal Child Psychology, 6,* 385–400.

Diáz-Guerrero, R., & Holtzman, W. H. (1974). Learning by televised "Plaza Sesamo" in Mexico. *Journal of Educational Psychology, 66,* 632–643.

Dickerson, K. (1994). Research registers. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 71–83). New York: Russell Sage Foundation.

Dickerson, K., Higgins, K., & Meinert, C. L. (1990). Identification of meta-analyses: The need for standard terminology. *Controlled Clinical Trials, 11,* 52–66.

Diehr, P., Martin, D. C., Koepsell, T., & Cheadle, A. (1995). Breaking the matches in a paired *t*-test for community interventions when the number of pairs is small. *Statistics in Medicine, 14,* 1491–1504.

DiRaddo, J. D. (1996). The investigation and amelioration of a staff turnover problem. *Dissertation Abstracts International, 59* (04), 1133A (University Microfilms No. 316379).

Director, S. M. (1979). Underadjustment bias in the evaluation of manpower training. *Evaluation Review, 3,* 190–218.

Dixon, D. O., & Lagakos, S. W. (2000). Should data and safety monitoring boards share confidential interim data? *Controlled Clinical Trials, 21,* 1–6.

Dohrenwend, B. P., Shrout, P. E., Egri, G., & Mendelsohn, F. S. (1980). Nonspecific psychological distress and other dimensions of psychopathology. *Archives of General Psychiatry, 37,* 1229–1236.

Donner, A. (1992). Sample size requirements for stratified cluster randomization designs. *Statistics in Medicine, 11,* 743–750.

Donner, A., & Klar, N. (1994). Cluster randomization trials in epidemiology: Theory and application. *Journal of Statistical Planning and Inference, 42,* 37–56.

Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research.* London: Arnold.

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effects. *Biometrics, 49,* 1231–1236.

Drake, C., & Fisher, L. (1995). Prognostic models and the propensity score. *International Journal of Epidemiology, 24,* 183–187.

Drake, S. (1981). *Cause, experiment, and science.* Chicago: University of Chicago Press.

Draper, D. (1995). Inference and hierarchical modeling in social sciences. *Journal of Educational and Behavioral Statistics, 20,* 115–147.

Droitcour, J. A. (1997). Cross-design synthesis: Concepts and applications. In E. Chelimsky & W. R. Shadish (Eds.), *Evaluation for the 21st century: A handbook* (pp. 360–372). Thousand Oaks, CA: Sage.

Ducasse, C. J. (1951). *Nature, mind and death.* La Salle, IL: Open Court.

Duckart, J. P. (1998). An evaluation of the Baltimore Community Lead Education and Reduction Corps (CLEARCorps) Program. *Evaluation Review, 22,* 373–402.

Dukes, R. L., Ullman, J. B., & Stein, J. A. (1995). An evaluation of D.A.R.E. (Drug Abuse Resistance Education), using a Solomon Four-Group design with latent variables. *Evaluation Review, 19,* 409–435.

Duncan, G. J., Yeung, W. J., Brooks-Gunn, J., & Smith, J. R. (1998). How much does childhood poverty affect the life chances of children? *American Sociological Review, 63,* 406–423.

Dunford, F. W. (1990). Random assignment: Practical considerations from field experiments. *Evaluation and Program Planning, 13,* 125–132.

Durlak, J. A., & Lipsey, M. W. (1991). A practitioner's guide to meta-analysis. *American Journal of Community Psychology, 19,* 291–332.

Duval, S., & Tweedie, R. (2000). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95,* 89–98.

Dwyer, J. H., & Flesch-Janys, D. (1995). Agent Orange in Vietnam. *American Journal of Public Health, 85,* 476–478.

Edgington, E. S. (1987). Randomized single-subject experiments and statistical tests. *Journal of Counseling Psychology, 34,* 437–442.

Edgington, E. S. (1992). Nonparametric tests for single-case experiments. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis* (pp. 133–157). Hillsdale, NJ: Erlbaum.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5,* 155–174.

Eells, E. (1991). *Probabilistic causality.* New York: Cambridge University Press.

Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika, 58,* 403–417.

Efron, B., & Feldman, D. (1991). Compliance as an explanatory variable in clinical trials. *Journal of the American Statistical Association, 86,* 9–26.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* New York: Chapman & Hall.

Eggan, F. (1954). Social anthropology and the method of controlled comparison. *American Anthropologist, 56,* 743–763.

Einstein, A. (1949). Reply to criticisms. In P. A. Schilpp (Ed.), *Albert Einstein: Philosopher-scientist* (pp. 665–688) . Evanston, IL: Library of Living Philosophers.

Eisenhart, M., & Howe, K. (1992). Validity in educational research. In M. D. LeCompte, W. L. Millroy, & J. Preissle (Eds.), *The handbook of qualitative research in education* (pp. 643–680). San Diego: Academic Press.

Eisner, E. (1979). *The educational imagination.* New York: Macmillan.

Eisner, E. (1983). Anastasia might still be alive, but the monarchy is dead. *Educational Researcher, 12,* 5.

Elbourne, D., Garcia, J., & Snowdon, C. (1999). Reply. *Controlled Clinical Trials, 20,* 571–572.

Elkin, I., Parloff, M. B., Hadley, S. W., & Autry, J. H. (1985). NIMH Treatment of Depression Collaborative Research Program: Background and research plan. *Archives of General Psychiatry, 42,* 305–316.

Elkin, I., Shea, T., Watkins, J. T., Imber, S. D., Sotsky, S. M., Collins, J. F., Glass, D. R., Pilkonis, P. A., Leber, W. R., Docherty, J. P., Fiester, S. J., & Parloff, M. B. (1989). National Institute of Mental Health Treatment of Depression Collaborative Research Program: General effectiveness of treatments. *Archives of General Psychiatry, 46,* 971–982.

Ellenberg, J. H. (1994). Cohort studies: Selection bias in observational and experimental studies. *Statistics in Medicine, 13,* 557–567.

Ellenberg, S. S., Finkelstein, D. M., & Schoenfeld, D. A. (1992). Statistical issues arising in AIDS clinical trials. *Journal of the American Statistical Association, 87,* 562–569.

Elmore, R. F. (1996). Getting to scale with good educational practice. *Harvard Educational Review, 66,* 1–26.

Emanuel, E. J., Wendler, D., & Grady, C. (2000). What makes clinical research ethical? *Journal of the American Medical Association, 283,* 2701–2711.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *American Psychologist, 93,* 179–197.

Emerson, R. M. (1981). Observational field work. *Annual Review of Sociology, 7,* 351–378.

Emerson, S. S. (1996). Statistical packages for group sequential methods. *American Statistician, 50,* 183–192.

Epperson, D. L., Bushway, D. J., & Warman, R. E. (1983). Client self-termination after one counseling session: Effects of problem recognition, counselor gender, and counselor experience. *Journal of Counseling Psychology, 30,* 307–315.

Equal Employment Opportunity Commission, Department of Labor, Department of Justice, and the Civil Service Commission. (1978, August). Adoption by four agencies of uniform guidelines on employee selection procedures. 34 Fed. Reg. 38290–38315.

Erbland, M. L., Deupree, R. H., & Niewoehner, D. E. (1999). Systemic corticosteroids in chronic obstructive pulmonary disease exacerbations (SCCOPE): Rationale and design of an equivalence trial. *Controlled Clinical Trials, 19,* 404–417.

Erez, E. (1986). Randomized experiments in correctional context: Legal, ethical, and practical concerns. *Journal of Criminal Justice, 14,* 389–400.

Esbensen, F.-A., Deschenes, E. P., Vogel, R. E., West, J., Arboit, K., & Harris, L. (1996). Active parental consent in school-based research: An examination of ethical and methodological issues. *Evaluation Review, 20,* 737–753.

Estes, W. K. (1997). Significance testing in psychological research: Some persisting issues. *Psychological Science, 8,* 18–20.

Estroff, S. E. (1981). *Making it crazy: An ethnography of psychiatric clients in an American community.* Berkeley: University of California Press.

Etzioni, R. D., & Kadane, J. B. (1995). Bayesian statistical methods in public health and medicine. *Annual Review of Public Health, 16,* 23–41.

Everitt, D. E., Soumerai, S. B., Avorn, J., Klapholz, H., & Wessels, M. (1990). Changing surgical antimicrobial prophylaxis practices through education targeted at senior department leaders. *Infectious Control and Hospital Epidemiology, 11,* 578–583.

Expanded availability of investigational new drugs through a parallel track mechanism for people with AIDS and HIV-related diseases, 55 Fed. Reg. 20856–20860 (1990).

Eyberg, S. M., & Johnson, S. M. (1974). Multiple assessment of behavior modification with families: Effects of contingency contracting and order of treated problems. *Journal of Consulting and Clinical Psychology, 42,* 594–606.

Eysenck, H. J. (1978). An exercise in mega-silliness. *American Psychologist, 33,* 517.

Eysenck, H. J., & Eysenck, M. (1983). *Mindwatching: Why people behave the way they do.* Garden City, NY: Anchor Press.

Fagan, J. A. (1990). Natural experiments in criminal justice. In K. L. Kempf (Ed.), *Measurement issues in criminology* (pp. 108–137). New York: Springer-Verlag.

Fagerstrom, D. O. (1978). Measuring degree of physical dependence to tobacco smoking with reference to individualization of treatment. *Addictive Behaviors, 3,* 235–241.

Fairweather, G. W., & Tornatsky, L. G. (1977). *Experimental methods for social policy research.* New York: Pergamon Press.

Faith, M. S., Allison, D. B., & Gorman, B. S. (1997). Meta-analysis of single-case research. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 245–277). Hillsdale, NJ: Erlbaum.

Family Support Act, Pub. L. N. 100–485, Section 203, 102 Stat. 2380 (1988).

Farquhar, J. W., Fortmann, S. P., Flora, J. A., Taylor, C. B., Haskell, W. L., Williams, P. T., MacCoby, N., & Wood, P. D. (1990). The Stanford five-city project: Effects of community-wide education on cardiovascular disease risk factors. *Journal of the American Medical Association, 26,* 359–365.

Faust, D. (1984). *The limits of scientific reasoning.* Minneapolis: University of Minnesota Press.

Federal Judicial Center. (1981). *Experimentation in the law: Report of the Federal Judicial Center Advisory Committee on Experimentation in the Law.* Washington, DC: U.S. Government Printing Office.

Feinauer, D. M., & Havlovic, S. J. (1993). Drug testing as a strategy to reduce occupational accidents: A longitudinal analysis. *Journal of Safety Research, 24,* 1–7.

Feinberg, S. E. (1971). Randomization and social affairs: The 1970 draft lottery. *Science, 171,* 255–261.

Feinberg, S. E., Singer, B., & Tanur, J. M. (1985). Large-scale social experimentation in the United States. In A. C. Atkinson & S. E. Feinberg (Eds.), *A celebration of statistics: The ISI centenary volume* (pp. 287–326). New York: Springer-Verlag.

Feldman, H. A., & McKinlay, S. M. (1994). Cohort versus cross-sectional design in large field trials: Precision, sample size, and a unifying model. *Statistics in Medicine, 13,* 61–78.

Feldman, H. A., McKinlay, S. M., & Niknian, M. (1996). Batch sampling to improve power in a community trial: Experience from the Pawtucket Heart Health Program. *Evaluation Review, 20,* 244–274.

Feldman, R. (1968). Response to compatriot and foreigner who seek assistance. *Journal of Personality and Social Psychology, 10,* 202–214.

Festinger, L. (1953). Laboratory experiments. In L. Festinger & D. Katz (Eds.), *Research methods in the behavioral sciences* (pp. 136–172). New York: Holt, Rinehart & Winston.

Fetterman, D. M. (1982). Ibsen's baths: Reactivity and insensitivity. *Educational Evaluation and Policy Analysis, 4,* 261–279.

Fetterman, D. M. (Ed.). (1984). *Ethnography in educational evaluation.* Beverly Hills, CA: Sage.

Feyerabend, P. (1975). *Against method: Outline of an anarchisitic theory of knowledge.* Atlantic Highlands, NJ: Humanities Press.

Feyerabend, P. (1978). *Science in a free society.* London: New Left Books.

Filstead, W. (1979). Qualitative methods: A needed perspective in evaluation research. In T. Cook & C. Reichardt (Eds.), *Qualitative and quantitative methods in evaluation research* (pp. 33–48). Newbury Park, CA: Sage.

Fink, A. (1998). *Conducting research literature reviews.* Thousand Oaks, CA: Sage.

Finkelstein, M. O., Levin, B., & Robbins, H. (1996a). Clinical and prophylactic trials with assured new treatment for those at greater risk: I. A design proposal. *American Journal of Public Health, 86,* 691–695.

Finkelstein, M. O., Levin, B., & Robbins, H. (1996b). Clinical and prophylactic trials with assured new treatment for those at greater risk: II. Examples. *American Journal of Public Health, 86,* 696–705.

Finn, J. D., & Achilles, C. M. (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal, 27,* 557–577.

Fischer, R. (1994). Control construct design in evaluating campaigns. *Public Relations Review, 21,* 45–58.

Fischer-Lapp, K., & Goetghebeur, E. (1999). Practical properties of some structural mean analyses of the effect of compliance in randomized trials. *Controlled Clinical Trials, 20,* 531–546.

Fischhoff, B. (1975). Hindsight/foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance, 1,* 288–299.

Fisher, L. D. (1999). Advances in clinical trials in the twentieth century. *Annual Review of Public Health, 20,* 109–124.

Fisher, R. A. (1925). *Statistical methods for research workers.* Edinburgh: Oliver & Boyd.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain, 33,* 505–513.

Fisher, R. A. (1932). *Statistical methods for research workers* (4th ed.). London: Oliver & Boyd.

Fisher, R. A. (1935). *The design of experiments.* Edinburgh: Oliver & Boyd.

Fisher, R. A., & Yates, F. (1953). *Statistical tables for biological, agricultural, and medical research* (4th ed.). Edinburgh: Oliver & Boyd.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: Wiley.

Fleiss, J. L. (1986). *The design and analysis of clinical experiments.* New York: Wiley.

Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 245–260). New York: Russell Sage Foundation.

Flournoy, N., & Rosenberger, W. F. (Eds.). (1995). *Adaptive designs.* Hayward, CA: IMS.

Folkman, J. (1996). Fighting cancer by attacking its blood supply. *Scientific American, 275,* 150–154.

Fortin, F., & Kirouac, S. (1976). A randomized controlled trial of preoperative patient education. *International Journal of Nursing Studies, 13,* 11–24.

Foster, E. M., & Bickman, L. (1996). An evaluator's guide to detecting attrition problems. *Evaluation Review, 20,* 695–723.

Foster, E. M., & McLanahan, S. (1996). An illustration of the use of instrumental variables: Do neighborhood conditions affect a young person's chance of finishing high school? *Psychological Methods, 1,* 249–261.

Fowler, R. L. (1985). Testing for substantive significance in applied research by specifying nonzero effect null hypotheses. *Journal of Applied Psychology, 70,* 215–218.

Fraker, T., & Maynard, R. (1986, October). *The adequacy of comparison group designs for evaluations of employment-related programs.* (Available from Mathematica Policy Research, P.O. Box 2393, Princeton, NJ 08543-2393)

Fraker, T., & Maynard, R. (1987). Evaluating comparison group designs with employment-related programs. *Journal of Human Resources, 22,* 194–227.

Frangakis, C. E., & Rubin, D. B. (1999) Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika, 86,* 366–379.

Frankel, M. (1983). Sampling theory. In P. H. Rossi, J. D. Wright, & A. B. Anderson (Eds.), *Handbook of survey research* (pp. 21–67). San Diego: Academic Press.

Franklin, C., Grant, D., Corcoran, J., Miller, P. O., & Bultman, L. (1997). Effectiveness of prevention programs for adolescent pregnancy: A meta-analysis. *Journal of Marriage and the Family, 59,* 551–567.

Franklin, R. D., Allison, D. B., & Gorman, B. S. (Eds.). (1997). *Design and analysis of single-case research.* Mahwah, NJ: Erlbaum.

Freedman, D. A. (1987). As others see us: A case study in path analysis. *Journal of Educational Statistics, 12,* 101–128.

Franklin, R. D., Allison, D. B., & Gorman, B. S. (Eds.). (1997). *Design and analysis of single-case research.* Mahwah, NJ: Erlbaum.

Freedman, D. A. (1987). As others see us: A case study in path analysis. *Journal of Educational Statistics, 12,* 101–128.

Freedman, L. S., & White, S. J. (1976). On the use of Pocock and Simon's method for balancing treatment numbers over prognostic variables in the controlled clinical trial. *Biometrics, 32,* 691–694.

Freiman, J. A., Chalmers, T. C., Smith, H., & Kuebler, R. R. (1978). The importance of beta, the Type II error, and sample size in the design and interpretation of the randomized control trial. *New England Journal of Medicine, 299,* 690–694.

Frick, R. W. (1995). Accepting the null hypothesis. *Memory & Cognition, 23,* 132–138.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods, 1,* 379–390.

Friedlander, D., & Robins, P. K. (1995). Evaluating program evaluations: New evidence on commonly used nonexperimental methods. *American Economic Review, 85,* 923–937.

Friedman, J., & Weinberg, D. H. (Eds.). (1983). *Urban affairs annual review: Volume 24, The great housing experiment.* Thousand Oaks, CA: Sage.

Fuller, H. (2000). Evidence supports the expansion of the Milwaukee parental choice program. *Phi Delta Kappan, 81,* 390–391.

Fuller, W. A. (1995). *Introduction to statistical time series* (2nd ed.). New York: Wiley.

Furby, L. (1973). Interpreting regression toward the mean in development research. *Developmental Psychology, 8,* 172–179.

Furlong, M. J., Casas, J. M., Corrall, C., & Gordon, M. (1997). Changes in substance use patterns associated with the development of a community partnership project. *Evaluation and Program Planning, 20,* 299–305.

Furlong, M. J., & Wampold, B. E. (1981). Visual analysis of single-subject studies by school psychologists. *Psychology in the Schools, 18,* 80–86.

Gadenne, V. (1976). *Die Gultigkeit psychologischer Unterscuchungen.* Stuttgart, Germany: Kohlhammer.

Gail, M. H., Byar, D. P., Pechacek, T. F., & Corle, D. K. (1992). Aspects of statistical design for the Community Intervention Trial for Smoking Cessation (COMMIT). *Controlled Clinical Trials, 13,* 6–21.

Gail, M. H., Mark, S. D., Carroll, R. J., Green, S. B., & Pee, D. (1996). On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine, 15,* 1069–1092.

Gallini, J. K., & Bell, M. E. (1983). Formulation of a structural equation model for the evaluation of curriculum. *Educational Evaluation and Policy Analysis, 5,* 319–326.

Galton, F. (1872). Statistical inquiries into the efficacy of prayer. *Fortnightly Review, 12,* 124–135.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute, 15,* 246–263.

Garber, J., & Hollon, S. D. (1991). What can specificity designs say about causality in psychopathology research? *Psychological Bulletin, 110,* 129–136.

Gastwirth, J. (1992). Method for assessing the sensitivity of statistical comparisons used in Title VII cases to omitted variables. *Jurimetrics, 33,* 19–34.

Gastwirth, J., Krieger, A., & Rosenbaum, P. (1994). How a court accepted an impossible explanation. *American Statistician, 48,* 313–315.

Geertz, C. (1973). Thick description: Toward an interpretative theory of culture. In C. Geertz (Ed.), *The interpretation of culture* (pp. 3–30). New York: Basic Books.

Gephart, W. J. (1981). Watercolor painting. In N. L. Smith (Ed.), *New techniques for evaluation* (pp. 286–298). Newbury Park, CA: Sage.

Gergen, K. J. (1973). Social psychology as history. *Journal of Personality and Social Psychology, 26,* 309–320.

Geronimus, A. T., & Korenman, S. (1992). The socioeconomic consequences of teen childbearing reconsidered. *Quarterly Journal of Economics, 107,* 1187–1214.

Gholson, B., & Houts, A. C. (1989). Toward a cognitive psychology of science. *Social Epistemology, 3,* 107–127.

Gholson, B. G., Shadish, W. R., Neimeyer, R. A., & Houts, A. C. (Eds.). (1989). *Psychology of science: Contributions to metascience.* Cambridge, England: Cambridge University Press.

Gibbons, R. D., Hedeker, D. R., & Davis, J. M. (1993). Estimation of effect size from a series of experiments involving paired comparisons. *Journal of Educational Statistics, 18,* 271–279.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review, 103,* 592–596.

Gilbert, J. P., McPeek, B., & Mosteller, F. (1977a). Progress in surgery and anesthesia: Benefits and risks of innovative therapy. In J. P. Bunker, B. A. Barnes, & F. Mosteller (Eds.), *Costs, risks, and benefits of surgery* (pp. 124–169). New York: Oxford University Press.

Gilbert, J. P., McPeek, B., & Mosteller, F. (1977b). Statistics and ethics in surgery and anesthesia. *Science, 198,* 684–689.

Gillespie, R. (1988). The Hawthorne experiments and the politics of experimentation. In J. Morawski (Ed.), *The rise of experimentation in American psychology* (pp. 114–137). New Haven, CT: Yale University Press.

Gilovich, T. (1991). *How we know what isn't so: The fallibility of human reasoning in everyday life.* New York: Free Press.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research.* New York: Aldine.

Glasgow, R. E., Vogt, T. M., & Boles, S. M. (1999). Evaluating the public health impact of health promotion interventions: The RE-AIM framework. *American Journal of Public Health, 89,* 1322–1327.

Glass, G. V. (1976). Primary, secondary, and meta-analysis. *Educational Researcher, 5,* 3–8.

Glass, G. V., & Smith, M. L. (1979). Meta-analysis of research on the relationship of class-size and achievement. *Educational Evaluation and Policy Analysis, 1,* 2–16.

Gleser, L. J., & Olkin, I. (1994). Stochastically dependent effect sizes. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 339–355). New York: Russell Sage Foundation.

Glymour, C., Scheines, R., Spirtes, P., & Kelly, K. (1987). *Discovering causal structure: Artificial intelligence, philosophy of science, and statistical modeling.* San Diego, CA: Academic Press.

Goetghebeur, E., & Molenberghs, G. (1996). Causal inference in a placebo-controlled clinical trial with binary outcome and ordered compliance. *Journal of the American Statistical Association, 91,* 928–934.

Goetghebeur, E., & Shapiro, S. H. (1996). Analyzing non-compliance in clinical trials: Ethical imperative or mission impossible? *Statistics in Medicine, 15,* 2813–2826.

Goetz, J. P., & LeCompte, M. D. (1984). *Ethnography and qualitative design in educational research.* San Diego, CA: Academic Press.

Goldberg, H. B. (1997, February). Prospective payment in action: The National Home Health Agency demonstration. *CARING, 17*(2), 14–27.

Goldberger, A. S. (1972a). *Selection bias in evaluating treatment effects: Some formal illustrations* (Discussion Paper No. 123). Madison: University of Wisconsin, Institute for Research on Poverty.

Goldberger, A. S. (1972b). *Selection bias in evaluating treatment effects: The case of interaction* (Discussion paper). Madison: University of Wisconsin, Institute for Research on Poverty.

Goldman, J. (1977). A randomization procedure for "trickle-process" evaluations. *Evaluation Quarterly, 1,* 493–498.

Goldschmidt, W. (1982). [Letter to the editor]. *American Anthropologist, 84,* 641–643.

Goldstein, H. (1987). *Multilevel models in educational and social research.* London: Oxford University Press.

Goldstein, H., Yang, M., Omar, R., Turner, R., & Thompson, S. (2000). Meta-analysis using multilevel models with an application to the study of class size effects. *Applied Statistics, 49,* 399–412.

Goldstein, J. P. (1986). The effect of motorcycle helmet use on the probability of fatality and the severity of head and neck injuries. *Evaluation Review, 10,* 355–375.

Gooding, D., Pinch, T., & Schaffer, S. (1989b). Preface. In D. Gooding, T. Pinch, & S. Schaffer (Eds.), *The uses of experiment: Studies in the natural sciences* (pp. xiii–xvii). Cambridge, England: Cambridge University Press.

Goodman, J. S., & Blum, T. C. (1996). Assessing the non-random sampling effects of subject attrition in longitudinal research. *Journal of Management, 22,* 627–652.

Goodson, B. D., Layzer, J. I., St. Pierre, R. G., Bernstein, L. S. & Lopez, M. (2000). Effectiveness of a comprehensive five-year family support program on low-income children and their families: Findings from the Comprehensive Child Development Program. *Early Childhood Research Quarterly, 15,* 5–39.

Gorman, B. S., & Allison, D. B. (1997). Statistical alternatives for single-case designs. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 159–214). Hillsdale, NJ: Erlbaum.

Gorman, M. E. (1994). Toward an experimental social psychology of science: Preliminary results and reflexive observations. In W. R. Shadish & S. Fuller (Eds.), *The social psychology of science* (pp. 181–196). New York: Guilford Press.

Gosnell, H. F. (1927). *Getting out the vote.* Chicago: University of Chicago Press.

Graham, J. W., & Donaldson, S. I. (1993). Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and use of follow-up data. *Journal of Applied Psychology, 78,* 119–128.

Grandy, J. (1987). *Characteristics of examinees who leave questions unanswered on the GRE general test under rights-only scoring* (GRE Board Professional Rep. No. 83-16P; ETS Research Rep. No. 87-38). Princeton, NJ: Educational Testing Service.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica, 37,* 424–438.

Green, S. B., Corle, D. K., Gail, M. H., Mark, S. D., Pee, D., Freedman, L. S., Graubard, B. I., & Lynn, W. R. (1995). Interplay between design and analysis for behavioral intervention trials with community as the unit of randomization. *American Journal of Epidemiology, 142,* 587–593.

Greenberg, D., & Shroder, M. (1997). *The digest of social experiments* (2nd ed.). Washington, DC: Urban Institute Press.

Greenberg, J., & Folger, R. (1988). *Controversial issues in social research methods.* New York: Springer-Verlag.

Greenberg, R. P., Bornstein, R. F., Greenberg, M.D., & Fisher, S. (1992). A meta-analysis of antidepressant outcome under "blinder" conditions. *Journal of Consulting and Clinical Psychology, 60,* 664–669.

Greene, C. N., & Podsakoff, P. M. (1978). Effects of removal of a pay incentive: A field experiment. *Proceedings of the Academy of Management, 38,* 206–210.

Greene, J. P., Peterson, P. E., & Du, J. (1999). Effectiveness of school choice: The Milwaukee experiment. *Education and Urban Society, 31,* 190–213.

Greene, W. H. (1985). LIMDEP: An econometric modeling program for the IBM PC. *American Statistician, 39,* 210.

Greene, W. H. (1999). *Econometric analysis.* Upper Saddle River, NJ: Prentice-Hall.

Greenhouse, J. B., & Iyengar, S. (1994). Sensitivity analysis and diagnostics. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 383–398). New York: Russell Sage Foundation.

Greenhouse, S. W. (1982). Jerome Cornfield's contributions to epidemiology. *Biometrics, 28* (Suppl.), 33–45.

Greenland, S., & Robins, J. M. (1986). Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology, 15,* 413–419.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82,* 1–20.

Greenwald, A. G., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and *p* values: What should be reported and what should be replicated? *Psychophysiology, 33,* 175–183.

Greenwald, P., & Cullen, J. W. (1984). The scientific approach to cancer control. *CA-A Cancer Journal for Clinicians, 34,* 328–332.

Greenwood, J. D. (1989). *Explanation and experiment in social psychological science: Realism and the social constitution of action.* New York: Springer-Verlag.

Griffin, L., & Ragin, C. C. (Eds.). (1994). Formal methods of qualitative analysis [Special issue]. *Sociological Methods and Research, 23*(1).

Grilli, R., Freemantle, N., Minozzi, S., Domenighetti, G., & Finer, D. (2000). Mass media interventions: Effects on health services utilization (Cochrane Review). *The Cochrane Library,* Issue 3. Oxford, England: Update Software.

Gross, A. J. (1993). Does exposure to second-hand smoke increase lung cancer risk? *Chance: New Directions for Statistics and Computing, 6,* 11–14.

Grossarth-Maticek, R., & Eysenck, H. J. (1989). Is media information that smoking causes illness a self-fulfilling prophecy? *Psychological Reports, 65,* 177–178.

Grossman, J., & Tierney, J. P. (1993). The fallibility of comparison groups. *Evaluation Review, 17,* 556–571.

Groves, R. M. (1989). *Survey errors and survey costs.* New York: Wiley.

Gu, X. S., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics, 2,* 405–420.

Guba, E., & Lincoln, Y. (1982). *Effective evaluation.* San Francisco: Jossey-Bass.

Guba, E., & Lincoln, Y. (1989). *Fourth generation evaluation.* Newbury Park, CA: Sage.

Guba, E. G. (1981). Investigative journalism. In N. L. Smith (Ed.), *New techniques for evaluation* (pp. 167–262). Newbury Park, CA: Sage.

Guba, E. G. (Ed.). (1990). *The paradigm dialog.* Newbury Park, CA: Sage.

Gueron, J. M. (1985). The demonstration of state work/welfare initiatives. In R. F. Boruch & W. Wothke (Eds.), *Randomization and field experimentation* (pp. 5–13). San Francisco: Jossey-Bass.

Gueron, J. M. (1999, May). *The politics of random assignment: Implementing studies and impacting policy.* Paper presented at the conference on Evaluation and Social Policy in Education of the American Academy of Arts and Sciences, Cambridge, MA.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6,* 427–439.

Gunn, W. J., Iverson, D. C., & Katz, M. (1985). Design of school health education evaluation. *Journal of School Health, 55,* 301–304.

Gurman, A. S., & Kniskern, D. P. (1978). Research on marital and family therapy: Progress, perspective, and prospect. In S. L. Garfield & A. E. Bergin (Eds.), *Handbook of psychotherapy and behavior change: An empirical analysis* (2nd ed., pp. 817–901). New York: Wiley.

Gwadz, M., & Rotheram-Borus, M. J. (1992, Fall). Tracking high-risk adolescents longitudinally. *AIDS Education and Prevention* (Suppl.), 69–82.

Haavelmo, T. (1944, July). The probability approach in econometrics. *Econometrica,* 12 (Suppl.).

Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science.* Cambridge, England: Cambridge University Press.

Hacking, I. (1988). Telepathy: Origins of randomization in experimental design. *Isis,* 79, 427–451.

Hackman, J. R., Pearce, J. L., & Wolfe, J. C. (1978). Effects of changes in job characteristics on work attitudes and behaviors: A naturally occurring quasi-experiment. *Organizational Behavior and Human Performance, 21,* 289–304.

Haddock, C. K., Rindskopf, D., & Shadish, W. R. (1998). Using odds ratios as effect sizes for meta-analysis of dichotomous data: A primer on methods and issues. *Psychological Methods, 3,* 339–353.

Haddock, C. K., Shadish, W. R., Klesges, R. C., & Stein, R. J. (1994). Treatments for childhood and adolescent obesity: A meta-analysis. *Annals of Behavioral Medicine, 16,* 235–244.

Hahn, G. J. (1984). Experimental design in the complex world. *Technometrics, 26,* 19–31.

Halvorsen, K. T. (1994). The reporting format. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 425–437). New York: Russell Sage Foundation.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Thousand Oaks, CA:: Sage.

Hamilton, J. D. (1994). *Time series analysis.* Princeton, NJ: Princeton University Press.

Hand, H. H., & Slocum, J. W., Jr. (1972). A longitudinal study of the effects of a human relations training program on managerial effectiveness. *Journal of Applied Psychology, 56,* 412–417.

Hankin, J. R., Sloan, J. J., Firestone, I. J., Ager, J. W., Sokol, R. J., & Martier, S. S. (1993). A time series analysis of the impact of the alcohol warning label on antenatal drinking. *Alcoholism: Clinical and Experimental Research, 17,* 284–289.

Hannan, E. G., & Deistler, M. (1988). *The statistical theory of linear systems.* New York: Wiley.

Hannan, P. J., & Murray, D. M. (1996). Gauss or Bernoulli? A Monte Carlo comparison of the performance of the linear mixed-model and the logistic mixed-model analyses in simulated community trials with a dichotomous outcome variable at the individual level. *Evaluation Review, 20,* 338–352.

Hansen, M. H., & Hurwitz, W. N. (1996, March). The problem of non-response in sample surveys. *Amstat News,* 25–26.

Hansen, M. H., Hurwitz, W. N., & Madow, W. G. (1993). *Sample survey methods and theory* (Vols. 1–2). Somerset, NJ: Wiley.

Hansen, W. B., Tobler, N. S., & Graham, J. W. (1990). Attrition in substance abuse prevention research: A meta-analysis of 85 longitudinally followed cohorts. *Evaluation Review, 14,* 677–685.

Hanson, N. R. (1958). *Patterns of discovery: An inquiry into the conceptual foundations of science.* Cambridge, England: Cambridge University Press.

Hanushek, E. A. (1999). The evidence on class size. In S. E. Mayer & P. E. Peterson (Eds.) *Earning and learning: How schools matter* (pp. 131–168). Washington, DC: Brookings.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Hillsdale, NJ: Erlbaum.

Harré, R. (1981). *Great scientific experiments.* Oxford, England: Phaidon Press.

Harris, M. J., & Rosenthal, R. (1985). Mediation of inrerpersonal expectancy effects: 31 meta-analyses. *Psychological Bulletin, 97,* 363–386.

Harris, R. J. (1997). Significance tests have their place. *Psychological Science, 8,* 8–11.

Harrop, J. W., & Velicer, W. F. (1990a). Computer programs for interrupted time series analysis: I. A qualitative evaluation. *Multivariate Behavioral Research, 25,* 219–231.

Harrop, J. W., & Velicer, W. F. (1990b). Computer programs for interrupted time series analysis: II. A quantitative evaluation. *Multivariate Behavioral Research, 25,* 233–248.

Hart, H. L. A., & Honore, T. (1985). *Causation in the law.* Oxford, England: Clarendon Press.

Hartman, R. S. (1991). A Monte Carlo analysis of alternative estimators in models involving selectivity. *Journal of Business and Economic Statistics, 9,* 41–49.

Hartmann, D. P., & Hall, R. V. (1976). The changing criterion design. *Journal of Applied Behavior Analysis, 9,* 527–532.

Hartmann, G. W. (1936). A field experiment on the comparative effectiveness of "emotional" and "rational" political leaflets in determining election results. *Journal of Abnormal and Social Psychology, 31,* 99–114.

Harvey, A. (1990). *The econometric analysis of time series* (2nd ed.). Cambridge, MA: MIT Press.

Harwell, M. (1997). An empirical study of Hedges's homogeneity test. *Psychological Methods, 2,* 219–231.

Hathaway, S. R., & McKinley, J. C. (1989). *MMPI-2: Manual for Administration and Scoring.* Minneapolis: University of Minnesota Press.

Hauk, W. W., & Anderson, S. (1986). A proposal for interpreting and reporting negative studies. *Statistics in Medicine, 5,* 203–209.

Hausman, J. A., & Wise, D. A. (Eds.). (1985). *Social experimentation.* Chicago: University of Chicago Press.

Havassey, B. (1988). *Efficacy of cocaine treatments: A collaborative study* (NIDA Grant Number DA05582). San Francisco: University of California.

Haveman, R. H. (1987). *Poverty policy and poverty research: The Great Society and the social sciences.* Madison: University of Wisconsin Press.

Hayduk, L. A. (1987). *Structural equation modeling with LISREL.* Baltimore: Johns Hopkins University Press.

Haynes, R. B., Taylor, D. W., & Sackett, D. L. (Eds.). (1979). *Compliance in health care.* Baltimore: Johns Hopkins University Press.

Heap, J. L. (1995). Constructionism in the rhetoric and practice of Fourth Generation Evaluation. *Evaluation and Program Planning, 18,* 51–61.

Hearst, N., Newman, T., & Hulley, S. (1986). Delayed effects of the military draft on mortality: A randomized natural experiment. *New England Journal of Medicine, 314,* 620–634.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica, 47,* 153–161.

Heckman, J. J. (1992). Randomization and social policy evaluation. In C. F. Manski & I. Garfinkel (Eds.), *Evaluating welfare and training programs* (pp. 201–230). Cambridge, MA: Harvard University Press.

Heckman, J. J. (1996). Comment. *Journal of the American Statistical Association, 91,* 459–462.

Heckman, J. J., & Hotz, V. J. (1989a). Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training. *Journal of the American Statistical Association, 84,* 862–874.

Heckman, J. J., & Hotz, V. J. (1989b). Rejoinder. *Journal of the American Statistical Association, 84,* 878–880.

Heckman, J. J., Hotz, V. J., & Dabos, M. (1987). Do we need experimental data to evaluate the impact of manpower training on earnings? *Evaluation Review, 11,* 395–427.

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies, 64,* 605–654.

Heckman, J. J., LaLonde, R. J., & Smith, J. A. (1999). The economics and econometrics of active labor market programs. In A. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (Vol. 3, pp. 1–160). Amsterdam: Elsevier Science.

Heckman, J. J., & Robb, R. (1985). Alternative methods for evaluating the impact of interventions. In J. J. Heckman & B. Singer (Eds.), *Longitudinal analysis of labor market data* (pp. 156–245). Cambridge, England: Cambridge University Press.

Heckman, J. J., & Robb, R. (1986a). Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In H. Wainer (Ed.), *Drawing inferences from self-selected samples* (pp. 63–107). New York: Springer-Verlag.

Heckman, J. J., & Robb, R. (1986b). Postscript: A rejoinder to Tukey. In H. Wainer (Ed.), *Drawing inferences from self-selected samples* (pp. 111–113). New York: Springer-Verlag.

Heckman, J. J., & Roselius, R. L. (1994, August). *Evaluating the impact of training on the earnings and labor force status of young women: Better data help a lot.* (Available from the Department of Economics, University of Chicago)

Heckman, J. J., & Roselius, R. L. (1995, August). *Non-experimental evaluation of job training programs for young men.* (Available from the Department of Economics, University of Chicago)

Heckman, J. J., & Todd, P. E. (1996, December). *Assessing the performance of alternative estimators of program impacts: A study of adult men and women in JTPA.* (Available from the Department of Economics, University of Chicago)

Hedeker, D., & Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel data. *Biometrics, 50,* 933–944.

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics, 9,* 61–85.

Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science, 7,* 246–255.

Hedges, L. V. (1994). Fixed effects models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 285–299). New York: Russell Sage Foundation.

Hedges, L. V. (1997a). The promise of replication in labour economics. *Labour Economics, 4,* 111–114.

Hedges, L. V. (1997b). The role of construct validity in causal generalization: The concept of total causal inference error. In V. R. McKim & S. P. Turner (Eds.), *Causality in crisis? Statistical methods and the search for causal knowledge in the social sciences* (pp. 325–341). Notre Dame, IN: University of Notre Dame Press.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Hedges, L. V., & Olkin, I. (in press). *Statistical methods for meta-analysis in the medical and social sciences.* Orlando, FL: Academic Press.

Hedges, L. V., & Vevea, J. L. (1996). Estimating effect size under publication bias: Small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics, 21,* 299–332.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 4,* 486–504.

Heider, F. (1944). Social perception and phenomenal causality. *Psychological Review, 51,* 358–374.

Heinsman, D. T., & Shadish, W. R. (1996). Assignment methods in experimentation: When do nonrandomized experiments approximate the answers from randomized experiments? *Psychological Methods, 1,* 154–169.

Heitjan, D. F. (1999). Causal inference in a clinical trial: A comparative example. *Controlled Clinical Trials, 20,* 309–318.

Hennigan, K. M., Del Rosario, M. L., Heath, L., Cook, T. D., Wharton, J. D., & Calder, B. J. (1982). Impact of the introduction of television on crime in the United States: Empirical findings and theoretical implications. *Journal of Personality and Social Psychology, 55,* 239–247.

Henry, G. T., & McMillan, J. H. (1993). Performance data: Three comparison methods. *Evaluation Review, 17,* 643–652.

Herbst, A., Ulfelder, H., & Poskanzer, D. (1971). Adenocarcinoma of the vagina: Association of maternal stilbestrol therapy with tumor appearance in young women. *New England Journal of Medicine, 284,* 878–881.

Herrnstein, R. J., & Murray, C. (1994). *The bell curve.* New York: The Free Press.

Hill, C. E., O'Grady, K. E., & Elkin, I. (1992). Applying the Collaborative Study Psychotherapy Rating Scale to rate therapist adherence in cognitive-behavior therapy, interpersonal therapy, and clinical management. *Journal of Consulting and Clinical Psychology, 60,* 73–79.

Hill, J. L., Rubin, D. B., & Thomas, N. (2000). The design of the New York School Choice Scholarship program evaluation. In L. Bickman (Ed.), *Validity and social experimentation: Donald Campbell's legacy* (Vol. 1, pp. 155–180). Thousand Oaks, CA: Sage.

Hillis, A., Rajab, M. H., Baisden, C. E., Villamaria, F. J., Ashley, P., & Cummings, C. (1998). Three years of experience with prospective randomized effectiveness studies. *Controlled Clinical Trials, 19,* 419–426.

Hillis, J. W., & Wortman, C. B. (1976). Some determinants of public acceptance of randomized control group experimental designs. *Sociometry, 39,* 91–96.

Hintze, J. L. (1996). *PASS User's Guide: PASS 6.0 Power Analysis and Sample Size for Windows.* (Available from Number Cruncher Statistical Systems, 329 North 1000 East, Kaysville, Utah 84037)

Hogg, R. V., & Tanis, E. A. (1988). *Probability and statistical inference* (3rd ed.). New York: Macmillan.

Hohmann, A. A., & Parron, D. L. (1996). How the new NIH guidelines on inclusion of women and minorities apply: Efficacy trials, effectiveness trials, and validity. *Journal of Consulting and Clinical Psychology, 64,* 851–855.

Holder, H. D., & Wagenaar, A. C. (1994). Mandated server training and reduced alcohol-involved traffic crashes: A time series analysis of the Oregon experience. *Accident Analysis and Prevention, 26,* 89–97.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81,* 945–970.

Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equation models. In C. C. Clogg (Ed.), *Sociological methodology* (pp. 449–493). Washington, DC: American Sociological Association.

Holland, P. W. (1989). Comment: It's very clear. *Journal of the American Statistical Association, 84,* 875–877.

Holland, P. W. (1994). Probabilistic causation without probability. In P. Humphreys (Ed.), *Pattrick Suppes: Scientific philosopher* (Vol. 1, pp. 257–292). Dordrecht, Netherlands: Kluwer.

Holland, P. W., & Rubin, D. B. (1983). On Lord's paradox. In H. Wainer & S. Messick (Eds.), *Principles of modern psychological measurement* (pp. 3–25). Hillsdale, NJ: Erlbaum.

Holland, P. W., & Rubin, D. B. (1988). Causal inference in retrospective studies. *Evaluation Review, 12,* 203–231.

Hollister, R. G., & Hill, J. (1995). Problems in the evaluation of community-wide initiatives. In J. P. Connell, A. C. Kubisch, L. B. Schorr, & C. H. Weiss (Eds.), *New approaches to evaluating community initiatives: Concepts, methods, and contexts* (pp. 127–172). Washington, DC: Aspen Institute.

Holton, G. (1986). *The advancement of science, and its burdens.* Cambridge, England: Cambridge University Press.

Hopson, R. K. (Ed.). (2000). *How and why language matters in evaluation.* San Francisco: Jossey-Bass.

Horn, J. L. (1991). Comments on "Issues in factorial invariance." In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions* (pp. 114–125). Washington, DC: American Psychological Association.

Horowich, P. (1990). *Truth.* Worcester, England: Basil Blackwell.

Houts, A., & Gholson, B. (1989). Brownian notions: One historicist philosopher's resistance to psychology of science via three truisms and ecological validity. *Social Epistemology, 3,* 139–146.

Howard, G. S., Maxwell, S. E., & Fleming, K. J. (2000). The proof of the pudding: An illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. *Psychological Methods, 5,* 315–332.

Howard, G. S., Millham, J., Slaten, S., & O'Donnell, L. (1981). The effect of subject response style factors on retrospective measures. *Applied Psychological Measurement, 5,* 89–100.

Howard, G. S., Ralph, K. M., Gulanick, N. A., Maxwell, S. E., Nance, D. W., & Gerber, S. K. (1979). Internal invalidity in pretest–posttest self-reports and a re-evaluation of retrospective pretests. *Applied Psychological Measurement, 3,* 1–23.

Howard, K. I., Cox, W. M., & Saunders, S. M. (1988). Attrition in substance abuse comparative treatment research: The illusion of randomization. In L. S. Onken & J. D. Blaine (Eds.), *Psychotherapy and counseling in the treatment of drug abuse* (pp. 66–79). Rockville, MD: National Institute on Drug Abuse.

Howard, K. I., Kopta, S. M., Krause, M. S., & Orlinsky, D. E. (1986). The dose-effect relationship in psychotherapy. *American Psychologist, 41,* 159–164.

Howard, K. I., Krause, M. S., & Orlinsky, D. E. (1986). The attrition dilemma: Toward a new strategy for psychotherapy research. *Journal of Consulting and Clinical Psychology, 54,* 106–110.

Hox, J. J. (1995). AMOS, EQS, and LISREL for Windows: A comparative review. *Structural Equation Modeling, 2,* 79–91.

Hrobjartsson, A., Gotzche, P. C., & Gluud, C. (1998). The controlled clinical trial turns 100: Fibieger's trial of serum treatment of diptheria. *British Medical Journal, 317,* 1243–1245.

Hsiao, C. (1986). *Analysis of panel data.* New York: Cambridge University Press.

Hsiao, C., Lahiri, K., Lee, L.-F., & Pesaran, M. H. (Eds.). (1999). *Analysis of panels and limited dependent variable models: In honour of G. S. Maddala.* Cambridge, England: Cambridge University Press.

Huitema, B. E. (1980). *The analysis of covariance and alternatives.* New York: Wiley.

Hultsch, D. F., & Hickey, T. (1978). External validity in the study of human development: Theoretical and methodological issues. *Human Development, 21,* 76–91.

Humphreys, P. (Ed.). (1986a). Causality in the social sciences [Special issue]. *Synthese, 68*(1).

Humphreys, P. (1989). *The chances of explanation: Causal explanation in the social, medical, and physical sciences.* Princeton, NJ: Princeton University Press.

Hunter, J. E. (1997). Needed: A ban on significance tests. *Psychological Science, 8,* 3–7.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park CA: Sage.

Hunter, J. E., & Schmidt, F. L. (1994). Correcting for sources of artificial variation across studies. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 323–336). New York: Russell Sage Foundation.

Imbens, G. W., & Rubin, D. B. (1997a). Bayesian inference for causal effects in randomized experiments with non-compliance. *Annals of Statistics, 25,* 305–327.

Imbens, G. W., & Rubin, D. B. (1997b). Estimating outcome distributions for compliers in instrumental variables models. *Review of Economic Studies, 64,* 555–574.

Imber, S. D., Pilkonis, P. A., Sotsky, S. M., Elkin, I., Watkins, J. T., Collins, J. F., Shea, M. T., Leber, W. R., & Glass, D. R. (1990). Mode-specific effects among three treatments for depression. *Journal of Consulting and Clinical Psychology, 58,* 352–359.

Innes, J. M. (1979). Attitudes towards randomized control group experimental designs in the field of community welfare. *Psychological Reports, 44,* 1207–1213.

International Conference on Harmonization. (1999, May 7). *Draft consensus guideline: Choice of control group in clinical trials* [On-line]. Available: *http://www.ifpma.org/ich1.html,* or from ICH Secretariat, c/o IFPMA, 30 rue de St-Jean, P.O. Box 9, 1211 Geneva 18, Switzerland.

Ioannidis, J. P. A., Dixon, D. O., McIntosh, M., Albert, J. M., Bozzette, S. A., & Schnittman, S. M. (1999). Relationship between event rates and treatment effects in clinical site differences within mnlticenter trials: An example from primary Pneumocystic carinii prophylaxis. *Controlled Clinical Trials, 20,* 253–266.

Isserman, A., & Rephann, T. (1995). The economic effects of the Appalachian Regional Commission: An empirical assessment of 26 years of regional development planning. *Journal of the American Planning Association, 61,* 345–364.

Iyengar, S., & Greenhouse, J. B. (1988). Selection models and the file-drawer problem [with discussion]. *Statistical Science, 3,* 109–135.

Jacobson, N. S., & Baucom, D. H. (1977). Design and assessment of nonspecific control groups in behavior modification research. *Behavior Therapy, 8,* 709–719.

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy, 15,* 336–352.

Jacobson, N. S., Schmaling, K. B., & Holtzworth-Munroe, A. (1987). Component analysis of behavioral marital therapy: Two-year follow-up and prediction of relapse. *Journal of Marital and Family Therapy, 13,* 187–195.

Jadad, A. R., Moore, A., Carroll, D., Jenkinson, C., Reynolds, D. J. M., Gavaghan, D. J., & McQuay, H. J. (1996). Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Controlled Clinical Trials, 17,* 1–12.

James, L. R., & Brett, J. M. (1984). Mediators, moderators, and tests for mediation. *Journal of Applied Psychology, 69,* 307–321.

Jason, L. A., McCoy, K., Blanco, D., & Zolik, E. S. (1981). Decreasing dog litter: Behavioral consultation to help a community group. In H. E. Freeman & M. A. Solomon (Eds.), *Evaluation studies review annual* (Vol. 6, pp. 660–674). Thousand Oaks, CA: Sage.

Jennrich, R. I., & Schlucter, M. D. (1986). Unbalanced repeated measures models with structured covariance matrices. *Biometrics, 42,* 805–820.

Jensen, K. B. (1989). Discourses of interviewing: Validating qualitative research findings through textual analysis. In S. Kvale (Ed.), *Issues of validity in qualitative research* (pp. 93–108). Lund, Sweden: Studentliteratur.

Johnson, B. T. (1989). *DSTAT: Software for the meta-analytic review of research literatures.* Hillsdale NJ: Erlbaum.

Johnson, B. T. (1993). *DSTAT 1.10: Software for the meta-analytic review of research literatures* [Upgrade documentation]. Hillsdale NJ: Erlbaum.

Johnson, M., Yazdi, K., & Gelb, B. D. (1993). Attorney advertising and changes in the demand for wills. *Journal of Advertising, 22,* 35–45.

Jones, B. J., & Meiners, M. R. (1986, August). *Nursing home discharges: The results of an incentive reimbursement experiment* (Long-Term Care Studies Program Research Report; DHHS Publication No. PHS 86-3399). Rockville, MD: U.S. Department of Health and Human Services, Public Health Service, National Center for Health Services Research and Health Care Technology Assessment.

Jones, E. E. (1985). Major developments in social psychology during the past five decades. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (Vol. 1, pp. 47–107). New York: Random House.

Jones, E. E., & Sigall, H. (1971). The bogus pipeline: A new paradigm for measuring affect and attitude. *Psychological Bulletin, 76,* 349–364.

Jones, J. H. (1981). *Bad blood: The Tuskegee syphilis experiment.* New York: Free Press.

Jones, K. (1991). The application of time series methods to moderate span longitudinal data. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions* (pp. 75–87). Washington, DC: American Psychological Association.

Jones, W. T. (1969a). *A history of Western philosophy: Vol. 1. The classical mind* (2nd ed.). New York: Harcourt, Brace, & World.

Jones, W. T. (1969b). *A history of Western philosophy: Vol. 3. Hobbes to Hume* (2nd ed.). New York: Harcourt, Brace, & World.

Joreskog, K. G., & Sorbom, D. (1988). *LISREL 7: A Guide to the Program and Applications.* (Available from SPSS, Inc., 444 N. Michigan Ave., Chicago, IL)

Joreskog, K. G., & Sorbom, D. (1990). Model search with TETRAD II and LISREL. *Sociological Methods and Research, 19,* 93–106.

Joreskog, K. G., & Sorbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language.* (Available from Scientific Software International, Inc., 1525 East 53rd Street, Suite 906, Chicago IL)

Judd, C. M., & Kenny, D. A. (1981a). *Estimating the effects of social interventions.* Cambridge, England: Cambridge University Press.

Judd, C. M., & Kenny, D. A. (1981b). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review, 5,* 602–619.

Judd, C. M., McClelland, G. H., & Culhane, S. E. (1995). Data analysis: Continuing issues in the everyday analysis of psychological data. *Annual Review of Psychology, 46,* 433–465.

Judge, G., Hill, C., Griffiths, W., & Lee, T. (1985). *The theory and practice of econometrics.* New York: Wiley.

Kadane, J. B. (Ed.). (1996). *Bayesian methods and ethics in clinical trial design.* New York: Wiley.

Kalaian, H. A., & Raudenbush, S. W. (1996). A multivariate mixed linear model for meta-analysis. *Psychological Methods, 1,* 227–235.

Kalish, L. A., & Begg, C. B. (1985). Treatment allocation methods in clinical trials: A review. *Statistics in Medicine, 4,* 129–144.

Karlin, S. (1987). Path analysis in genetic epidemiology and alternatives. *Journal of Educational Statistics, 12,* 165–177.

Katz, L. F., Kling, J., & Liebman, J. (1997, November). *Moving to opportunity in Boston: Early impacts of a housing mobility program.* Unpublished manuscript, Kennedy School of Government, Harvard University.

Kazdin, A. E. (1992). *Research design in clinical psychology* (2nd ed.). Boston: Allyn & Bacon.

Kazdin, A. E. (1996). Dropping out of child psychotherapy: Issues for research and implications for practice. *Clinical Child Psychology and Psychiatry, 1,* 133–156.

Kazdin, A. E., & Bass, D. (1989). Power to detect differences between alternative treatments in comparative psychotherapy outcome research. *Journal of Consulting and Clinical Psychology, 57,* 138–147.

Kazdin, A. E., Mazurick, J. L., & Bass, D. (1993). Risk for attrition in treatment of antisocial children and families. *Journal of Clinical Child Psychology, 22,* 2–16.

Kazdin, A. E., & Wilcoxon, L. A. (1976). Systematic desensitization and non-specific treatment effects: A methodological evaluation. *Psychological Bulletin, 83,* 729–758.

Keller, R. T., & Holland, W. E. (1981). Job change: A naturally occurring field experiment. *Human Relations, 134,* 1053–1067.

Kelling, G. L., Pate, T., Dieckman, D., & Brown, C. E. (1976). The Kansas City Preventive Patrol Experiment: A summary report. In G. V. Glass (Ed.), *Evaluation studies review annual* (Vol. 1, pp. 605–657). Beverly Hills, CA: Sage.

Kelly, J. A., Murphy, D. A., Sikkema, K. J., McAuliffe, T. L., Roffman, R. A., Solomon, L. J., Winett, R. A., Kalichman, S. C., & The Community HIV Prevention Research Collaborative. (1997). Randomised, controlled, community-level HIV-prevention intervention for sexual-risk behaviour among homosexual men in US cities. *Lancet, 350,* 1500–1505.

Kendall, M., & Ord, J. K. (1990). *Time series* (3rd ed.). London: Arnold.

Kendall, P. C. (1998). Empirically supported psychological therapies. *Journal of Consulting and Clinical Psychology, 66,* 3–6.

Kenny, D. A. (1979). *Correlation and causality.* New York: Wiley.

Kenny, D. A., & Harackiewicz, J. M. (1979). Cross-lagged panel correlation: Practice and promise. *Journal of Applied Psychology, 64,* 372–379.

Kenny, D. A., & Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin, 96,* 201–210.

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Kershaw, D., & Fair, J. (1976). *The New Jersey income-maintenance experiment: Vol. 1. Operations, surveys, and administration.* New York: Academic Press.

Kershaw, D., & Fair, J. (1977). *The New Jersey income-maintenance experiment: Vol. 3. Expenditures, health, and social behavior.* New York: Academic Press.

Kiecolt, K. J., & Nathan, L. E. (1990). *Secondary analysis of survey data.* Thousand Oaks, CA: Sage.

Kim, J., & Trivedi, P. K. (1994). Econometric time series analysis software: A review. *American Statistician, 48,* 336–346.

Kirk, J., & Miller, M. L. (1986). *Reliability and validity in qualitative research.* Thousand Oaks, CA: Sage.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences* (2nd ed.). Belmont, CA: Brooks/Cole.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* 746–759.

Kirkhart, K. E. (1995). Seeking multicultural validity: A postcard from the road. *Evaluation Practice, 16,* 1–12.

Kirsch, I. (1996). Hypnotic enhancement of cognitive-behavioral weight loss treatments: Another meta-reanalysis. *Journal of Consulting and Clinical Psychology, 64,* 517–519.

Kish, L. (1987). *Statistical design for research.* New York: Wiley.

Kisker, E. E., & Love, J. M. (1999, December). *Leading the way: Characteristics and early experiences of selected Early Head Start programs.* Washington, DC: U.S. Department of Health and Human Services, Administration on Children, Youth, and Families.

Kitzman, H., Olds, D. L., Henderson, C. R., Hanks, C., Cole, R., Tatelbaum, R., McConnochie, K. M., Sidora, K., Luckey, D. W., Shaver, D., Engelhardt, K., James, D., & Barnard, K. (1997). Effect of prenatal and infancy home visitation by nurses on pregnancy outcomes, childhood injuries, and repeated childbearing. A randomized controlled trial. *Journal of the American Medical Association, 278,* 644–652.

Klein, L. R. (1992). Self-concept, enhancement, computer education and remediation: A study of the relationship between a multifaceted intervention program and academic achievement. *Dissertation Abstracts International, 53* (05), 1471A. (University Microfilms No. 9227700)

Kleinbaum, D. G., Kupper, L. L., & Morgenstern, H. (1982). *Epidemiologic research: Principles and quantitative methods.* New York: Van Nostrand Reinhold.

Klesges, R. C., Brown, K., Pascale, R. W., Murphy, M., Williams, E., & Cigrang, J. A. (1988). Factors associated with participation, attrition, and outcome in a smoking cessation program at the workplace. *Health Psychology, 7,* 575–589.

Klesges, R. C., Haddock, C. K., Lando, H., & Talcott, G. W. (1999). Efficacy of a forced smoking cessation and an adjunctive behavioral treatment on long-term smoking rates. *Journal of Consulting and Clinical Psychology, 67,* 952–958.

Klesges, R. C., Vasey, M. M., & Glasgow, R. E. (1986). A worksite smoking modification competition: Potential for public health impact. *American Journal of Public Health, 76,* 198–200.

Kline, R. B., Canter, W. A., & Robin, A. (1987). Parameters of teenage alcohol abuse: A path analytic conceptual model. *Journal of Consulting and Clinical Psychology, 55,* 521–528.

Knatterud, G. L., Rockhold, F. W., George, S. L., Barton, F. B., Davis, C. E., Fairweather, W. R., Honohan, T., Mowery, R., & O'Neill, R. (1998). Guidelines for quality assurance in multicenter trials: A position paper. *Controlled Clinical Trials, 19,* 477–493.

Knight, G. P., Fabes, R. A., & Higgins, D. A. (1996). Concerns about drawing causal inferences from meta-analyses: An example in the study of gender differences in aggression. *Psychological Bulletin, 119,* 410–421.

Knorr-Cetina, K.D. (1981). *The manufacture of knowledge: An essay on the constructivist and contextual nature of science.* Oxford, England: Pergamon.

Koehler, M. J., & Levin, J. R. (1998). Regulated randomization: A potentially sharper analytical tool for the multiple baseline design. *Psychological Methods, 3,* 206–217.

Koepke, D., & Flay, B. R. (1989). Levels of analysis. In M. T. Braverman (Ed.), *Evaluating health promotion programs* (pp. 75–87). San Francisco: Jossey-Bass.

Koepsell, T. D., Martin, D. C., Diehr, P. H., Psaty, B. M., Wagner, E. G., Perrin, E. B., & Cheadle, A. (1991). Data analysis and sample size issues in evaluations of community-based health promotion and disease prevention programs: A mixed-model analysis of variance approach. *Journal of Clinical Epidemiology, 44,* 701–713.

Kollins, S. H., Newland, M. C., & Critchfield, T. S. (1999). Quantitative integration of single-subject studies: Methods and misinterpretations. *Behavior Analyst, 22,* 149–157.

Kopta, S. M., Howard, K. I., Lowry, J. L., & Beutler, L. E. (1994). Patterns of symptomatic recovery in psychotherapy. *Journal of Consulting and Clinical Psychology, 62,* 1009–1016.

Korfmacher, J., O'Brien, R., Hiatt, S., & Olds, D. (1999). Differences in program implementation between nurses and paraprofessionals providing home visits during pregnancy and infancy: A randomized trial. *American Journal of Public Health, 89,* 1847–1851.

Koricheva, J., Larsson, S., & Haukioja, E. (1998). Insect performance on experimentally stressed woody plants: A meta-analysis. *Annual Review of Entomology, 43,* 195–216.

Kraemer, H. C., Gardner, C., Brooks, J. L., & Yesavage, J. A. (1998). Advantages of excluding underpowered studies in meta-analysis: Inclusionist versus exclusionist viewpoints. *Psychological Methods, 3,* 23–31.

Kraemer, H. C., & Thiemann, S. (1989). A strategy to use soft data effectively in randomized controlled clinical trials. *Journal of Consulting and Clinical Psychology, 57,* 148–154.

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association, 94,* 1372–1381.

Kratochwill, T. R., & Levin, J. R. (1992). *Single-case research design and analysis: New directions for psychology and education.* Hillsdale, NJ: Erlbaum.

Kromrey, J. D., & Foster-Johnson, L. (1996). Determining the efficacy of intervention: The use of effect sizes for data analysis in single-subject research. *Journal of Experimental Education, 65,* 73–93.

Kruglanski, A. W., & Kroy, M. (1976). Outcome validity in experimental research: A reconceptualization. *Journal of Representative Research in Social Psychology, 7,* 168–178.

Krull, J. L., & MacKinnon, D. P. (1999). Multilevel mediation modeling in group-based intervention studies. *Evaluation Review, 23,* 418–444.

Kruse, A. Y., Kjaergard, L. L., Krogsgaard, K., Gluud, C., Mortensen, E. L., Gottschau, A., Bjerg, A., and the INFO Trial Group. (2000). A randomized trial assessing the impact of written information on outpatients' knowledge about and attitude toward randomized clinical trials. *Controlled Clinical Trials, 21,* 223–240.

Kruskal, W., & Mosteller, F. (1979a). Representative sampling: I. Non-scientific literature. *International Statistical Review, 47,* 13–24.

Kruskal, W., & Mosteller, F. (1979b). Representative sampling: II. Scientific literature, excluding statistics. *International Statistical Review, 47,* 111–127.

Kruskal, W., & Mosteller, F. (1979c). Representative sampling: III. The current statistical literature. *International Statistical Review, 47,* 245–265.

Kuhn, T. S. (1962). *The structure of scientific revolutions.* Chicago: University of Chicago Press.

Kunz, R., & Oxman, D. (1998). The unpredictability paradox: Review of empirical comparisons of randomised and non-randomised clinical trials. *British Medical Journal, 317,* 1185–1190.

Kvale, S. (Ed.). (1989). *Issues of validity in qualitative research.* Lund, Sweden: Studentlitteratur.

Lachin, J. M. (1988). Statistical properties of randomization in clinical trials. *Controlled Clinical Trials, 9,* 289–311.

Lachin, J. M. (2000). Statistical considerations in the intent-to-treat principle. *Controlled Clinical Trials, 21,* 167–189.

Lachin, J. M., Matts, J. P., & Wei, L. J. (1988). Randomization in clinical trials: Conclusions and recommendations. *Statistics in Medicine, 9,* 365–374.

Lakatos, I. (1978). *The methodology of scientific research programmes.* Cambridge, England: Cambridge University Press.

Lakoff, G. (1985). *Women, fire, and dangerous things.* Chicago: University of Chicago Press.

LaLonde, R. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review, 76,* 604–620.

LaLonde, R., & Maynard, R. (1987). How precise are evaluations of employment and training experiments: Evidence from a field experiment. *Evaluation Review, 11,* 428–451.

Lam, J. A., Hartwell, S. W., & Jekel, J. F. (1994). "I prayed real hard, so I know I'll get in": Living with randomization. In K. J. Conrad (Ed.), *Critically evaluating the role of experiments* (pp. 55–66). San Francisco: Jossey-Bass.

Lana, R. C. (1969). Pretest sensitization. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research* (pp. 119–141). New York: Academic Press.

Larson, R. C. (1976). What happened to patrol operations in Kansas City? *Evaluation, 3,* 117–123.

Latané, B., & Darley, J. M. (1970). *The unresponsive bystander: Why doesn't he help?* New York: Appleton-Century-Crofts.

Latané, B., & Nida, S. (1981). Ten years of research on group size and helping. *Psychological Bulletin, 89,* 308–324.

Latour, B. (1987). *Science in action.* Cambridge, MA: Harvard University Press.

Latour, B., & Woolgar, S. (1979). *Laboratory life: The social construction of scientific facts.* Beverly Hills, CA: Sage.

Lavori, P. W. (1992). Clinical trials in psychiatry: Should protocol deviation censor patient data? *Neuropsychopharmacology, 6,* 39–48.

Lavori, P. W., Louis, T. A., Bailar, J. C., & Polansky, H. (1986). Designs for experiments: Parallel comparisons of treatment. In J. C. Bailar & F. Mosteller (Eds.), *Medical uses of statistics* (pp. 61–82). Waltham, MA: New England Journal of Medicine.

Lazar, I., & Darlington, R. (1982). Lasting effects of early education. *Monographs of the Society for Research in Child Development, 47* (2–3, Serial No. 195).

Lazarsfeld, P. F. (1947). *The mutual effects of statistical variables.* Unpublished manuscript, Columbia University, Bureau of Applied Social Research.

Lazarsfeld, P. F. (1948). The use of panels in social research. *Proceedings of the American Philosophical Society, 92,* 405–410.

Leaf, R. C., DiGiuseppe, R., Mass, R., & Alington, D. E. (1993). Statistical methods for analyses of incomplete service records: Concurrent use of longitudinal and cross-sectional data. *Journal of Consulting and Clinical Psychology, 61,* 495–505.

Lee, S., & Hershberger, S. (1990). A simple rule for generating equivalent models in covariance structure modeling. *Multivariate Behavioral Research, 25,* 313–334.

Lee, Y., Ellenberg, J., Hirtz, D., & Nelson, K. (1991). Analysis of clinical trials by treatment actually received: Is it really an option? *Statistics in Medicine, 10,* 1595–1605.

Lehman, D. R., Lempert, R. O., & Nisbett, R. E. (1988). The effects of graduate training on reasoning: Formal discipline and thinking about everyday-life events. *American Psychologist, 43,* 431–442.

LeLorier, J., Gregoire, G., Benhaddad, A., Lapierre, J., & Derderian, F. (1997). Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *New England Journal of Medicine, 337,* 536–542.

Leviton, L. C., & Cook, T. D. (1983). Evaluation findings in education and social work textbooks. *Evaluation Review, 7,* 497–518.

Leviton, L. C., Finnegan, J. R., Zapka, J. G., Meischke, H., Estabrook, B., Gilliland, J., Linares, A., Weitzman, E. R., Raczynski, J., & Stone, E. (1999). Formative research methods to understand patient and provider responses to heart attack symptoms. *Evaluation and Program Planning, 22,* 385–397.

Levy, A. S., Mathews, O., Stephenson, M., Tenney, J. E., & Schucker, R. E. (1985). The impact of a nutrition information program on food purchases. *Journal of Public Policy and Marketing, 4,* 1–13.

Lewin, K. (1935). *A dynamic theory of personality: Selected papers.* New York: McGraw-Hill.

Lewis, D. (1973). Causation. *Journal of Philosophy, 70,* 556–567.

Lewontin, R. (1997, January 9). Billions and billions of demons. *New York Review of Books, 64*(1), 28–32.

Li, Z., & Begg, C. B. (1994). Random effects models for combining results from controlled and uncontrolled studies in a meta-analysis. *Journal of the American Statistical Association, 89,* 1523–1527.

Lichstein, K. L. (1988). *Clinical relaxation strategies.* New York: Wiley.

Lichstein, K. L., Riedel, B. W., & Grieve, R. (1994). Fair tests of clinical trials: A treatment implementation model. *Advances in Behavior Research and Therapy, 16,* 1–29.

Lichtenstein, E. L., Glasgow, R. E., & Abrams, D. B. (1986). Social support in smoking cessation: In search of effective interventions. *Behavior Therapy, 17,* 607–619.

Lichtenstein, L. M. (1993). Allergy and the immune system. *Scientific American, 269,* 117–124.

Lieberman, S. (1956). The effects of changes in roles on the attitudes of role occupants. *Human Relations, 9,* 385–402.

Lieberson, S. (1985). *Making it count: The improvement of social research and theory.* Berkeley: University of California Press.

Liebman, B. (1996). Vitamin E and fat: Anatomy of a flip-flop. *Nutrition Action Newsletter, 23,* 10–11.

Light, R. J., Singer, J. D., & Willett, J. B. (1990). *By design: Planning research in higher education.* Cambridge, MA: Harvard University Press.

Light, R. J., Singer, J. D., & Willett, J. B. (1994). The visual presentation and interpretation of meta-analysis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 439–453). New York: Russell Sage Foundation.

Lincoln, Y. S. (1990). Campbell's retrospective and a constructivist's perspective. *Harvard Educational Review, 60,* 501–504.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry.* Newbury Park, CA: Sage.

Lind, E. A. (1985). Randomized experiments in the Federal courts. In R. F. Boruch & W. Wothke (Eds.), *Randomization and field experimentation* (pp. 73–80). San Francisco: Jossey-Bass.

Lind, J. (1753). *A treatise of the scurvy. Of three parts containing an inquiry into the nature, causes and cure of that disease.* Edinburgh: Sands, Murray, & Cochran.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research.* Thousand Oaks, CA: Sage.

Lipsey, M. W. (1992). Juvenile delinquency treatment: A meta-analytic inquiry into the variability of effects. In T. D. Cook, H. M. Cooper, D. S. Cordray, H. Hartmann, L. V. Hedges, R. J. Light, T. A. Louis, & F. Mosteller (Eds.), *Meta-analysis for explanation: A casebook* (pp. 83–127). New York: Russell Sage Foundation.

Lipsey, M. W., Cordray, D. S., & Berger, D. E. (1981). Evaluation of a juvenile diversion program. *Evaluation Review, 5,* 283–306.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48,* 1181–1209.

Lipsey, M. W., & Wilson, D. B. (2000). *Practical meta-analysis.* Newbury Park, CA: Sage.

Little, R. J. (1985). A note about models for selectivity bias. *Econometrica, 53,* 1469–1474.

Little, R. J. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association, 90,* 1112–1121.

Little, R. J., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley.

Little, R. J., & Rubin, D. B. (1999). Comment. *Journal of the American Statistical Association, 94,* 1130–1132.

Little, R. J., & Schenker, N. (1995). Missing data. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 39–75). New York: Plenum Press.

Little, R. J., & Yau, L. (1996). Intent-to-treat analysis for longitudinal studies with dropouts. *Biometrics, 52,* 1324–1333.

Little, R. J., & Yau, L. (1998). Statistical techniques for analyzing data from prevention trials: Treatment of no-shows using Rubin's causal model. *Psychological Methods, 3,* 147–159.

Locke, E. A. (1986). *Generalizing from laboratory to field settings.* Lexington, MA: Lexington Books.

Locke, H. J., & Wallace, K. M. (1959). Short-term marital adjustment and prediction tests: Their reliability and validity. *Journal of Marriage and Family Living, 21,* 251–255.

Locke, J. (1975). *An essay concerning human understanding.* Oxford, England: Clarendon Press. (Original work published in 1690)

Lockhart, D. C. (Ed.). (1984). *Making effective use of mailed questionnaires.* San Francisco: Jossey-Bass.

Loehlin, J. C. (1992, January). *Latent variable models: An introduction to factor, path, and structural analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Lohr, B. W. (1972). *An historical view of the research on the factors related to the utilization of health services.* Unpublished mauuscript. Rockville, MD: Bureau for Health Services Research and Evaluation, Social and Economic Analysis Division.

Looney, M. A., Feltz, C. J., & Van Vleet, C. N. (1994). The reporting and analysis of research findings for within-subject designs: Methodological issues for meta-analysis. *Research Quarterly for Exercise & Sport, 65,* 363–366.

Lord, F. M. (1963). Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 21–38). Madison: University of Wisconsin Press.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin, 68,* 304–305.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Louis, T. A., Fineberg, H. V., & Mosteller, F. (1985). Findings for public health from meta-analyses. *Annual Review of Public Health, 6,* 1–20.

Louis, T. A., & Zelterman, D. (1994). Bayesian approaches to research synthesis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 411–422). New York: Russell Sage Foundation.

Ludwig, J., Duncan, G. J., & Hirschfield, P. (1998, September). *Urban poverty and juvenile crime: Evidence from a randomized housing-mobility experiment.* (Available from author, Ludwig, Georgetown Public Policy Institute, 3600 N Street NW, Suite 200, Washington, DC 20007)

Lufr, H. S. (1990). The applicability of the regression discontinuity design in health services research. In L. Sechrest, E. Perrin, & J. Bunker (Eds.), *Research methodology: Strengthening causal interpretations of nonexperimental data* (pp. 141–143). Rockville, MD: Public Health Service, Agency for Health Care Policy and Research.

Lund, E. (1989). The validity of different control groups in a case-control study: Oral contraceptive use and breast cancer in young women. *Journal of Clinical Epidemiology, 42,* 987–993.

MacCallum, R. C. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin, 100,* 107–120.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1,* 130–149.

MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin, 114,* 185–199.

MacCready, R. A. (1974). Admissions of phenylketonuric patients to residential institutions before and after screening programs of the newborn infant. *Journal of Pediatrics, 85,* 383–385.

MacIntyre, A. (1981). *After virtue.* Notre Dame, IN: University of Notre Dame Press.

MacKenzie, A., Funderburk, F. R., Allen, R. P., & Stefan, R. L. (1987). The characteristics of alcoholics frequently lost to follow-up. *Journal of Studies on Alcohol, 48,* 119–123.

Mackie, J. L. (1974). *The cement of the universe: A study of causation.* Oxford, England: Oxford University Press.

MacKinnon, D. P., Johnson, C. A., Pentz, M. A., Dwyer, J. H., Hansen, W. B., Flay, B. R., & Wang, E.Y.-I. (1991). Mediating mechanisms in a school-based drug prevention program: First-year effects of the Midwestern Prevention Project. *Health Psychology, 10,* 164–172.

MacKinnon, D. P., Warsi, G., & Dwyer, J. H. (1995). A simulation study of mediated effect measures. *Multivariate Behavioral Research, 30,* 41–62.

Madden, E. H., & Humber, J. (1971). Nomological necessity and C. J. Ducasse. *Ratio, 13,* 119–138.

Madaus, G. F., & Greaney, V. (1985). The Irish Experience in competency testing: Implications for American education. *American Journal of Education, 93,* 268–294.

Magidson, J. (1977). Toward a causal model approach for adjusting for preexisting differences in the nonequivalent control group situation. *Evaluation Quarterly, 1,* 399–420.

Magidson, J. (1978). Reply to Bentler and Woodward: The .05 significance level is not all-powerful. *Evaluation Quarterly, 2,* 511–520.

Magidson, J. (2000). On models used to adjust for preexisting differences. In L. Bickman (Ed.), *Research design: Donald Campbell's legacy* (Vol. 2, pp. 181–194). Thousand Oaks, CA: Sage.

Magnusson, D. (2000). The individual as the organizing principle in psychological inquiry: A holistic approach. In L. R. Bergman, R. B. Cairns, L. Nilsson, & L. Nystedt (Eds.), *Developmental science and the holistic approach* (pp. 33–47). Mahwah, NJ: Erlbaum.

Makuch, R., & Simon, R. (1978). Sample size requirements for evaluating a conservative therapy. *Cancer Treatment Reports, 62,* 1037–1040.

Mallams, J. H., Godley, M. D., Hall, G. M., & Meyers, R. J. (1982). A social systems approach to resocializing alcoholics in the community. *Journal of Studies on Alcohol, 43,* 1115–1123.

Maltz, M. D., Gordon, A. C., McDowall, D., & McCleary, R. (1980). An artifact in pretest-posttest designs: How it can mistakenly make delinquency programs look effective. *Evaluation Review, 4,* 225–240.

Mann, C. (1994). Can meta-analysis make policy? *Science, 266,* 960–962.

Mann, T. (1994). Informed consent for psychological research: Do subjects comprehend consent forms and understand their legal rights? *Psychological Science, 5,* 140–143.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *American Economic Review Papers and Proceedings, 80,* 319–323.

Manski, C. F., & Garfinkel, I. (1992). Introduction. In C. F. Manski & I. Garfinkel (Eds.), *Evaluating welfare and training programs* (pp. 1–22). Cambridge, MA: Harvard University Press.

Manski, C. F., & Nagin, D. S. (1998). Bounding disagreements about treatment effects: A case study of sentencing and recidivism. In A. Raftery (Ed.), *Sociological methodology* (pp. 99–137). Cambridge, MA: Blackwell.

Manski, C. F., Sandefur, G. D., McLanahan, S., & Powers, D. (1992). Alternative estimates of the effect of family structure during adolescence on high school graduation. *Journal of the American Statistical Association, 87,* 25–37.

Marascuilo, L. A., & Busk, P. L. (1988). Combining statistics for multiple-baseline AB and replicated ABAB designs across subjects. *Behavioral Assessment, 10,* 1–28.

Marcantonio, R. J. (1998). *ESTIMATE: Statistical software to estimate the impact of missing data* [Computer software]. Lake in the Hills, IL: Statistical Research Associates.

Marcus, S. M. (1997a). Assessing non-consent bias with parallel randomized and nonrandomized clinical trials. *Journal of Clinical Epidemiology, 50,* 823–828.

Marcus, S. M. (1997b). Using omitted variable bias to assess uncertainty in the estimation of an AIDS education treatment effect. *Journal of Educational and Behavioral Statistics, 22,* 193–201.

Marcus, S. M. (2001). A sensitivity analysis for subverting randomization in controlled trials. *Statistics in Medicine, 20, 545–555.*

Margraf, J., Ehlers, A., Roth, W. T., Clark, D. B., Sheikh, J., Agras, W. S., & Taylor, C. B. (1991). How "blind" are double-blind studies? *Journal of Consulting and Clinical Psychology, 59,* 184–187.

Marin, G., Marin, B. V., Perez-Stable, E. J., Sabogal, F., & Ostero-Sabogal, R. (1990). Changes in information as a function of a culturally appropriate smoking cessation community intervention for Hispanics. *American Journal of Community Psychology, 18,* 847–864.

Mark, M. M. (1986). Validity typologies and the logic and practice of quasi-experimentaton. In W. M. K. Trochim (Ed.), *Advances in quasi-experimental design and analysis* (pp. 47–66). San Francisco: Jossey-Bass.

Mark, M. M. (2000). Realism, validity, and the experimenting society. In L. Bickman (Ed.), *Validity and social experimentation: Donald Campbell's legacy* (Vol. 1, pp. 141–166). Thousand Oaks, CA: Sage.

Mark, M. M., & Mellor, S. (1991). Effect of self-relevance of an event on hindsight bias: The foreseeability of a layoff. *Journal of Applied Psychology, 76, 569–577.*

Marquis, D. (1983). Leaving therapy to chance. *Hastings Center Report, 13,* 40–47.

Marriott, F. H. C. (1990). *A dictionary of statistical terms* (5th ed.). Essex, England: Longman Scientific and Technical.

Marschak, J. (1953). Economic measurements for policy and prediction. In W. C. Hood & T. C. Koopmans (Ed.), *Studies in econometric method* (Cowles Commission Monograph No. 13). New York: Wiley.

Marsh, H. W. (1998). Simulation study of non-equivalent group-matching and regression-discontinuity designs: Evaluation of gifted and talented programs. *Journal of Experimental Education,* 66, 163–192.

Marshall, E. (1989). Quick release of AIDS drugs. *Science, 245,* 345, 347.

Martin, D. C., Diehr, P., Perrin, E. B., & Koepsell, T. D. (1993). The effect of matching on the power of randomized community intervention studies. *Statistics in Medicine, 12,* 329–338.

Martin, S. E., Annan, S., & Forst, B. (1993). The special deterrent effects of a jail sanction on first-time drunk drivers: A quasi-experimental study. *Accident Analysis and Prevention, 25,* 561–568.

Marx, J. L. (1989). Drug availability is an issue for cancer patients, too. *Science, 245,* 346–347.

Mase, B. F. (1971). *Changes in self-actualization as a result of two types of residential group experience.* Unpublished doctoral dissertation, Northwestern University, Evanston, IL.

Mastroianni, A. C., Faden, R., & Federman, D. (Eds.). (1994). *Women and health research* (Vols. 1–2). Washington, DC: National Academy Press.

Matt, G. E. (1989). Decision rules for selecting effect sizes in meta-analysis: A review and reanalysis of psychotherapy outcome studies. *Psychological Bulletin, 105,* 106–115.

Matt, G. E., & Cook, T. D. (1994). Threats to the validity of research syntheses. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 503–520). New York: Russell Sage Foundation.

Matt, G. E., Cook, T. D., & Shadish, W. R. (2000). *Generalizing about causal inferences.* Manuscript in preparation.

Mauro, R. (1990). Understanding L.O.V.E. (Left Out Variables Error): A method for examining the effects of omitted variables. *Psychological Bulletin, 108,* 314–329.

Maxwell, J. A. (1990). Response to "Campbell's retrospective and a constructivist's perspective." *Harvard Educational Review, 60,* 504–508.

Maxwell, J. A. (1992). Understanding and validity in qualitative research. *Harvard Educational Review, 62,* 279–300.

Maxwell, J. A., Bashook, P. G., & Sandlow, L. J. (1986). Combining ethnographic and experimental methods in educational evaluation: A case study. In D. M. Fetterman & M. A. Pittman (Eds.), *Educational evaluation: Ethnography in theory, practice, and politics* (pp. 121–143). Newbury Park, CA: Sage.

Maxwell, J. A., & Lincoln, Y. S. (1990). Methodology and epistemology: A dialogue. *Harvard Educational Review, 60,* 497–512.

Maxwell, S. E. (1993). Covariate imbalance and conditional size: Dependence on model-based adjustments. *Statistics in Medicine, 12,* 101–109.

Maxwell, S. E. (1994). Optimal allocation of assessment time in randomized pretest-posttest designs. *Psychological Bulletin, 115,* 142–152.

Maxwell, S. E. (1998), Longitudinal designs in randomized group comparisons: When will intermediate observations increase statistical power? *Psychological Methods, 3,* 275–290.

Maxwell, S. E., Cole, D. A., Arvey, R. D., & Salas, E. (1991). A comparison of methods for increasing power in randomized between-subjects designs. *Psychological Bulletin, 110,* 328–337.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison approach.* Pacific Grove, CA: Brooks/Cole.

Maxwell, S. E., Delaney, H. D., & Dill, C. A. (1984). Another look at ANCOVA versus blocking. *Psychological Bulletin, 95,* 136–147.

Mayer, L. S. (1986). Statistical inferences for cross-lagged panel models without the assumption of normal errors. *Social Science Research, 15,* 28–42.

McAweeney, M. J., & Klockars, A. J. (1998). Maximizing power in skewed distributions: Analysis and assignment. *Psychological Methods, 3,* 117–122.

McCall, W. A. (1923). *How to experiment in education.* New York: MacMillan.

McCardel, J. B. (1972). Interpersonal effects of structured and unstructured human relations groups. *Dissertation Abstracts International, 33,* 4518–4519. (University Microfilms No. 73-5828)

McClannahan, L. E., McGee, G. G., MacDuff, G. S., & Krantz, P. J. (1990). Assessing and improving child care: A personal appearance index for children with autism. *Journal of Applied Behavior Analysis, 23,* 469–482.

McCleary, R. D. (2000). The evolution of the time series experiment. In L. Bickman (Ed.), *Research design: Donald Campbell's legacy* (Vol. 2, pp. 215–234). Thousand Oaks, CA: Sage.

McCleary, R. D., & Welsh, W. N. (1992). Philosophical and statistical foundations of time-series experiments. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis* (pp. 41–91). Hillsdale, NJ: Erlbaum.

McClelland, G. H. (1997). Optimal design in psychological research. *Psychological Methods, 2,* 3–19.

McClelland, G. H. (2000). Increasing statistical power without increasing sample size. *American Psychologist, 55,* 963–964.

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin, 114,* 376–390.

McCord, J. (1978). A thirty-year followup of treatment effects. *American Psychologist, 33,* 284–289.

McCord, W., & McCord, J. (1959). *Origins of crime.* New York: Columbia University Press.

McCoy, H. V., & Nurco, D. M. (1991). Locating subjects by traditional techniques. In D. M. Nurco (Ed.), *Follow-up fieldwork: AIDS outreach and IV drug abuse* (DHHS Publication No. ADM 91-1736, pp. 31–73). Rockville, MD: National Institute on Drug Abuse.

McCullough, B. D., & Wilson, B. (1999). On the accuracy of statistical procedure in Microsoft Excel 97. *Computational Statistics and Data Analysis, 31,* 27–37.

McDonald, R. P. (1994). The bilevel reticular action model for path analysis with latent variables. *Sociological Methods and Research, 22,* 399–413.

McDowall, D., McCleary, R., Meidinger, E. E., & Hay, R. A. (1980). *Interrupted time series analysis.* Newbury Park CA: Sage.

McFadden, E. (1998). *Management of data in clinical trials.* New York: Wiley.

McGuire, W. J. (1984). A contextualist theory of knowledge: Its implications for innovation and return in psychological research. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (pp. 1–47). New York: Academic Press.

McGuire, W. J. (1997). Creative hypothesis generating in psychology: Some useful heuristics. *Annual Review of Psychology, 48,* 1–30.

McKay, H., Sinisterra, L., McKay, A., Gomez, H., & Lloreda, P. (1978). Improving cognitive ability in chronically deprived children. *Science, 200,* 270–278.

McKillip, J. (1992). Research without control groups: A control construct design. In F. B. Bryant, J. Edwards, R. S. Tindale, E. J. Posavac, L. Heath, & E. Henderson (Eds.), *Methodological issues in applied psychology* (pp. 159–175). New York: Plenum.

McKillip, J., & Baldwin, K. (1990). Evaluation of an STD education media campaign: A control construcr design. *Evaluation Review, 14,* 331–346.

McLeod, R. S., Taylor, D. W., Cohen, A., & Cullen, J. B. (1986, March 29). Single patient randomized climical trial: Its use in determining optimal treatment for patient with inflammation of a Kock continent ileostomy reservoir. *Lancet, 1,* 726–728.

McNees, P., Gilliam, S. W., Schnelle, J. F., & Risley, T. (1979). Controlling employee theft through time and product identification. *Journal of Organizational Behavior Management, 2,* 113–119.

McSweeny, A. J. (1978). The effects of response cost on the behavior of a million persons: Charging for directory assistance in Cincinnati. *Journal of Applied Behavioral Analysis, 11,* 47–51.

Mead, R. (1988). *The design of experiments: Statistical principles for practical application.* Cambridge, England: Cambridge University Press.

Medin, D. L. (1989). Concepts and conceptual structures. *American Psychologist, 44,* 1469–1481.

Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34,* 103–115.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.

Meier, P. (1972). The biggest public health experiment ever: The 1954 field trial of the Salk poliomyelitis vaccine. In J. M. Tanur, F. Mosteller, W. H. Kruskal, R. F. Liuk, R. S. Pieters, & G. R. Rising (Eds.), *Statistics: A guide to the unknown* (pp. 120–129). San Francisco: Holden-Day.

Meinert, C. L., Gilpin, A. K., Unalp, A., & Dawson, C. (2000). Gender representation in trials. *Controlled Clinical Trials, 21,* 462–475.

Menard, S. (1991). *Longitudinal research.* Thousand Oaks, CA: Sage.

Mennicke, S. A., Lent, R. W., & Burgoyne, K. L. (1988). Premature termination from university counseling centers: A review. *Journal of Counseling and Development, 66,* 458–464.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.), (pp. 13–103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validatiou of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741–749.

Meyer, B. D. (1995). Natural and quasi-experiments in economics. *Journal of Business and Economic Statistics, 13,* 151–161.

Meyer, D. L. (1991). Misinterpretation of interaction effects: A reply to Rosnow and Rosenthal. *Psychological Bulletin, 110,* 571–573.

Meyer, T. J., & Mark, M. M. (1995). Effects of psychosocial interveutions with adult cancer patients: A meta-aualysis of randomized experiments. *Health Psychology, 14,* 101–108.

Miettineu, O. S. (1985). The "case-control" study: Valid selection of subjects. *Journal of Chronic Diseases, 38,* 543–548.

Miké, V. (1989). Philosophers assess randomized clinical trials: The need for dialogue. *Controlled Clinical Trials, 10,* 244–253.

Miké, V. (1990). Suspended judgment: Ethics, evidence, and uncertainty. *Controlled Clinical Trials, 11,* 153–156.

Miles, M. B., & Huberman, A. M. (1984). *Qualitative data analysis: A sourcebook of new methods.* Newbury Park, CA: Sage.

Miller, N., Pedersen, W. C., & Pollock, V. E. (2000). Discriminative validity. In L. Bickman (Ed.), *Validity and social experimentation: Donald Campbell's legacy* (Vol. 1, pp. 65–99). Thousand Oaks, CA: Sage.

Miller, T. Q., Turner, C. W., Tindale, R. S., Posavac, E. J., & Dugoni, B. L. (1991). Reasons for the trend toward null findings in research on Type A behavior. *Psychological Bulletin, 110,* 469–485.

Millsap, M. A., Goodson, B., Chase, A., & Gamse, B. (1997, December). *Evaluation of "Spreading the Comer School Development Program and Philosophy."* Report submitted to the Rockefeller Foundation, 420 Fifth Avenue, New York, NY 10018 by Abt Associates Inc., 55 Wheeler Street, Cambridge MA 02138.

Minton, J. H. (1975). The impact of "Sesame Street" on reading readiness of kindergarten children. *Sociology of Education, 48,* 141–151.

Mishler, E. G. (1990). Validation in inquiry-guided research: The role of exemplars in narrative studies. *Harvard Educational Review, 60,* 415–442.

Mitroff, I. I., & Fitzgerald, I. (1977). On the psychology of the Apollo moon scientists: A chapter in the psychology of science. *Human Relations, 30,* 657–674.

Moberg, D. P., Piper, D. L., Wu, J., & Serlin, R. C. (1993). When total randomization is impossible: Nested random assignment. *Evaluation Review, 17,* 271–291.

Moerbeek, M., van Breukelen, G. J. P., & Berger, M. P. F. (2000). Design issues for experiments in multilevel populations. *Journal of Educational and Behavioral Statistics, 25,* 271–284.

Moffitt, R. A. (1989). Comment. *Journal of the American Statistical Association, 84,* 877–878.

Moffitt, R. A. (1991). Program evaluation with nonexperimental data. *Evaluation Review, 15,* 291–314.

Moffitt, R. A. (1996). Comment. *Journal of the American Statistical Association, 91,* 462–465.

Moher, D., Jadad, A. R., Nichol, G., Penman, M., Tugwell, P., & Walsh, S. (1995). Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists. *Controlled Clinical Trials, 16,* 62–73.

Moher, D., & Olkin, I. (1995). Meta-analysis of randomized controlled trials: A concern for standards. *Journal of the American Medical Association, 274,* 1962–1963.

Mohr, L. B. (1988). *Impact analysis for program evaluation.* Chicago: Dorsey Press.

Mohr, L. B. (1995). *Impact analysis for program evaluation* (2nd ed.). Thousand Oaks, CA: Sage.

Moos, R. H. (1997). *Evaluating treatment environments: The quality of psychiatric and substance abuse programs* (2nd ed.). New Brunswick, NJ: Transaction.

Morales, A. (2000, November). *Investigating rules and principles for combining qualitative and quantitative data.* Paper presented at the annual conference of the American Evaluation Association, Honolulu, Hawaii.

Morawski, J. G. (1988). *The rise of experimentation in American psychology.* New Haven, CT: Yale University Press.

Mosteller, F. (1990). Improving research methodology: An overview. In L. Sechrest, E. Perrin, & J. Bunker (Eds.), *Research methodology: Strengthening causal interpretations of nonexperimental data* (pp. 221–230). Rockville, MD: U.S. Public Health Service, Agency for Health Care Policy and Research.

Mosteller, F., Gilbert, J. P., & McPeek, B. (1980). Reporting standards and research strategies for controlled trials: Agenda for the editor. *Controlled Clinical Trials, 1,* 37–58.

Mosteller, F., Light, R. J., & Sachs, J. A. (1996). Sustained inquiry in education: Lessons from skill grouping and class size. *Harvard Educational Review, 66,* 797–842.

Mulford, H. A., Ledolrer, J., & Fitzgerald, J. L. (1992). Alcohol availability and consumption: Iowa sales data revisited. *Journal of Studies on Alcohol, 53,* 487–494.

Mulkay, M. (1979). *Science and the sociology of knowledge.* London: Allen & Unwin.

Mullin, B. (1993). *Advanced BASIC meta-analysis.* Hillsdale, NJ: Erlbaum.

Mumford, E., Schlesinger, H. J., Glass, G. V., Patrick, C., & Cuerdon, T. (1984). A new look at evidence about reduced cost of medical utilization following mental health treatment. *American Journal of Psychiatry, 141,* 1145–1158.

Murdoch, J. C., Singh, H., & Thayer, M. (1993). The impact of natural hazards on housing values: The Loma Prieta earthquake. *Journal of the American Real Estate and Urban Economics Association, 21,* 167–184.

Murnane, R. J., Newstead, S., & Olsen, R. J. (1985). Comparing public and private schools: The puzzling role of selectivity bias. *Journal of Business and Economic Statistics, 3,* 23–35.

Murray, C. (1984). *Losing ground: American social policy, 1950–1980.* New York: Basic Books.

Murray, D. M. (1998). *Design and analysis of group-randomized trials.* New York: Oxford University Press.

Murray, D. M., & Hannan, P. J. (1990). Planning for the appropriate analysis in school-based drug-use prevention studies. *Journal of Consulting and Clinical Psychology, 58,* 458–468.

Murray, D. M., Hannan, P. J., & Baker, W. L. (1996). A Monte Carlo study of alternative responses to intraclass correlation in community trials: Is it ever possible to avoid Cornfield's penalties? *Evaluation Review, 20,* 313–337.

Murray, D. M., McKinlay, S. M., Martin, D., Donner, A. P., Dwyer, J. H., Raudenbush, S. W., & Graubard, B. I. (1994). Design and analysis issues in community trials. *Evaluation Review, 18,* 493–514.

Murray, D. M., Moskowitz, J. M., & Dent, C. W. (1996). Design and analysis issues in community-based drug abuse prevention. *American Behavioral Scientist, 39,* 853–867.

Muthen, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research, 22,* 376–398.

Muthen, B. O., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52,* 431–462.

Narayanan, V. K., & Nath, R. (1982). A field test of some attitudinal and behavioral consequences of flexitime. *Journal of Applied Psychology, 67,* 214–218.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research* (OPRR Report; FR Doc. No. 79-12065). Washington, DC: U.S. Government Printing Office.

National Institutes of Health. (1994). *NIH guidelines on the inclusion of women and minorities as subjects in clinical research,* 59 Del. Reg. 14, 508 (Document No. 94-5435).

Naylor, R. H. (1989). Galileo's experimental discourse. In D. Gooding, T. Pinch, & S. Schaffer (Eds.), *The uses of experiment: Studies in the natural sciences* (pp. 117–134). Cambridge, England: Cambridge University Press.

Neal-Schuman Publishers. (Eds.). (1980). *National Directory of Mental Health.* New York: Wiley.

Nesselroade, J. R., Stigler, S. M., & Baltes, P. B. (1980). Regression to the mean and the study of change. *Psychological Bulletin, 88,* 622–637.

Neter, J., Wasserman, W., & Kutner, M. H. (1983). *Applied linear regression models.* New York: Gardner Press.

Neustrom, M. W., & Norton, W. M. (1993). The impact of drunk driving legislation in Louisiana. *Journal of Safety Research, 24,* 107–121.

Newbold, P., Agiakloglou, C., & Miller, J. (1994). Adventures with ARIMA software. *International Journal of Forecasting, 10,* 573–581.

Newhouse, J. P. (1993). *Free for all? Lessons from the RAND Health Insurance Experiment.* Cambridge, MA: Harvard University Press.

Newhouse, J. P., & McClellan, M. (1998). Econometrics in outcomes research: The use of instrumental variables. *Annual Review of Public Health, 19,* 17–34.

Nicholson, R. A., & Berman, J. S. (1983). Is follow-up necessary in evaluating psychotherapy? *Psychological Bulletin, 93,* 261–278.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241–301.

Nietzel, M. T., Russell, R. L., Hemmings, K. A., & Gretter, M. L. (1987). Clinical significance of psychotherapy for unipolar depression: A meta-analytic approach to social comparison. *Journal of Consulting and Clinical Psychology, 55,* 156–161.

Notz, W. W., Staw, B. M., & Cook, T. D. (1971). Attitude toward troop withdrawal from Indochina as a function of draft number: Dissonance or self-interest? *Journal of Personality and Social Psychology, 20,* 118–126.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Nurco, D. N., Robins, L. N., & O'Donnel, J. A. (1977). Locating respondents. In L. D. Johnston, D. N. Nurco, & L. N. Robins (Eds.), *Conducting follow-up research on drug treatment programs* (NIDA Treatment Program Monograph Series, No. 2, pp. 71–84). Rockville, MD: National Institute on Drug Abuse.

Nuremberg Code. (1949). *Trials of war criminals before the Nuremberg Military Tribunals under Control Council Law No. 10* (Vol. 2). Washington, DC: U.S. Government Printing Office.

Oakes, D., Moss, A. J., Fleiss, J. L., Bigger, J. T., Jr., Therneau, T., Eberly, S. W., McDermott, M. P., Manatunga, A., Carleen, E., Benhorn, J., and The Multicenter Diltiazem Post-Infarction Trial Research Group. (1993). Use of compliance measures in an analysis of the effect of Diltiazem on mortality and reinfarction after myocardial infarction. *Journal of the American Statistical Association, 88,* 44–49.

O'Carroll, P. W., Loftin, C., Waller, J. B., McDowall, D., Bukoff, A., Scott, R. O., Mercy, J. A., & Wiersema, B. (1991). Preventing homicide: An evaluation of the efficacy of a Detroit gun ordinance. *American Journal of Public Health, 81,* 576–581.

O'Connor, F. R., Devine, E. C., Cook, T. D. & Curtin, T. R. (1990). Enhancing surgical nurses' patient education. *Patient Education and Counseling, 16,* 7–20.

Okene, J. (Ed.). (1990). *Adoption and maintenance of behaviors for optimal health.* New York: Academic Press.

Oldroyd, D. (1986). *The arch of knowledge: An introductory study of the history of the philosophy and methodology of science.* New York: Methuen.

Olds, D. L., Eckenrode, D., Henderson, C. R., Kitzman, H., Powers, J., Cole, R., Sidora, K., Morris, P., Pettitt, L. M., & Luckey, D. (1997). Long-term effects of home visitation on maternal life course and child abuse and neglect. *Journal of the American Medical Association, 278,* 637–643.

Olds, D. L, Henderson, C. R., Cole, R., Eckenrode, J., Kitzman, H., Luckey, D., Pettitt, L. M., Sidora, K., Morris, P., & Powers, J. (1998). Long-term effects of nurse home visitation on children's criminal and antisocial behavior: 15-year follow-up of a randomized controlled trial. *Journal of the American Medical Association, 280,* 1238–1244.

Olds, D., Henderson, C. R., Kitzman, H., & Cole, R. (1995). Effects of prenatal and infancy nurse home visitation on surveillance of child maltreatment. *Pediatrics, 95,* 365–372.

O'Leary, K. D., Becker, W. C., Evans, M. B., & Saudargas, R. A. (1969). A token reinforcement program in a pnblic school: A replication and systematic analysis. *Journal of Applied Behavior Analysis, 3,* 3–13.

O'Leary, K. D., & Borkovec, T. D. (1978). Conceptual, methodological, and ethical problems of placebo groups in psychotherapy research. *American Psychologist, 33,* 821–830.

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance [Monograph]. *Journal of Applied Psychology, 78,* 679–703.

Orne, M. T. (1959). The nature of hypnosis: Artifact and essence. *Journal of Abnormal and Social Psychology, 58,* 277–299.

Orne, M. T. (1962). On the social psychology of the psychological experiment. *American Psychologist, 17,* 776–783.

Orne, M. T. (1969). Demand characteristics and the concept of quasi-controls. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research* (pp. 143–179). New York: Academic Press.

Orr, L. L. (1999). *Social experiments: Evaluating public programs with experimental methods.* Thousand Oaks, CA: Sage.

Orr, L. L., Johnston, T., Montgomery, M., & Hojnacki, M. (1989). *Design of the Washington Self-Employment and Enterprise Development (SEED) Demonstration.* Bethesda, MD: Abt/Battell Memorial Institute.

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics, 8,* 157–159.

Orwin, R. G. (1984). Evaluating the life cycle of a product warning: Saccharin and diet soft drimks. *Evaluation Review, 8,* 801–822.

Orwin, R. G. (1994). Evaluating coding decisions. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 140–162). New York: Russell Sage Foundation.

Orwin, R. G., & Cordray, D. S. (1985). Effects of deficient reporting on meta-analysis: A conceptual framework and reanalysis. *Psychological Bulletin, 85,* 185–193.

Orwin, R. G., Cordray, D. S., & Huebner, R. B. (1994). Judicious application of randomized designs. In K. J. Conrad (Ed.), *Critically evaluating the role of experiments* (pp. 73–86). San Francisco: Jossey-Bass.

Ostrom, C. W. (1990). *Time series analysis: Regression techniques* (2nd ed.). Thousand Oaks, CA: Sage.

Ottenbacher, K. J. (1986). Reliability and accuracy of visually analyzing graphed data from single-subject designs. *American Journal of Occupational Therapy, 40,* 464–469.

Overall, J. E., & Woodward, J. A. (1977). Nonrandom assignment and the analysis of covariance. *Psychological Bulletin, 84,* 588–594.

Overton, R. C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. *Psychological Methods, 3,* 354–379.

Ozminkowski, R. J., Wortman, P. M., & Roloff, D. W. (1989). Inborn/outborn status and neonatal survival: A meta-analysis of non-randomized studies. *Statistics in Medicine, 7,* 1207–1221.

Page, E. B. (1958). Teacher comments and student performance: A seventy-four classroom experiment in school motivation. *Journal of Educational Psychology, 49,* 173–181.

Palmer, C. R., & Rosenberger, W. F. (1999). Ethics and practice: Alternative designs for Phase III randomized clinical trials. *Controlled Clinical Trials, 20,* 172–186.

Patterson, G. R. (1986). Performance models for antisocial boys. *American Psychologist, 41,* 432–444.

Patton, M. Q. (1980). *Qualitative evaluation methods.* Beverly Hills, CA: Sage Publications.

Paulos, J. A. (1988). *Innumeracy: Mathematical illiteracy and its consequences.* New York: Hill & Wang.

Pearl, J. (2000). *Causality: Models, reasoning, and inference.* Cambridge, England: Cambridge University Press.

Pearlman, S., Zweben, A., & Li, S. (1989). The comparability of solicited versus clinic subjects in alcohol treatment research. *British Journal of Addictions, 84,* 523–532.

Pearson, K. (1904). Report on certain entiric fever inoculation statistics. *British Medical Journal,* 2, 1243–1246.

Peirce, C. S., & Jastrow, J. (1884). On small differences of sensation. *Memoirs of the National Academy of Sciences, 3,* 75–83.

Pelz, D. C., & Andrews, F. M. (1964). Detecting causal priorities in panel study data. *American Sociological Review, 29,* 836–848.

Permutt, T. (1990). Testing for imbalance of covariates in controlled experiments. *Statistics in Medicine, 12,* 1455–1462.

Perng, S. S. (1985). Accounts receivable treatments study. In R. F. Boruch & W. Wothke (Eds.), *Randomization and field experimentation* (pp. 55–62). San Francisco: Jossey-Bass.

Petersen, W. M. (1999, April 20). Economic status quo [Letter to the editor]. *New York Times,* p. A18.

Petty, R. E., Fabrigar, L. R., Wegener, D. T., & Priester, J. R. (1996). Understanding data when interactions are present or hypothesized. *Psychological Science, 7,* 247–252.

Pfungst, O. (1911). *Clever Hans (The horse of Mr. Von Osten)*. New York: Henry Holt.

Phillips, D. C. (1987). Validity in qualitative research: Why the worry about warrant will not wane. *Education and Urban Society, 20*, 9–24.

Phillips, D. C. (1990). Postpositivistic science: Myths and realities. In E. G. Guba (Ed.), *The paradigm dialog* (pp. 31–45). Newbury Park, CA: Sage.

Pigott, T. D. (1994). Methods for handling missing data in research syntheses. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 163–175). New York: Russell Sage Foundation.

Pinch, T. (1986). *Confronting nature*. Dordrecht, Holland: Reidel.

Pirie, P. L., Thomson, M. A., Mann, S. L., Peterson, A. V., Murray, D. M., Flay, B. R., & Best, J. A. (1989). Tracking and attrition in longitudinal school-based smoking prevention research. *Preventive Medicine, 18*, 249–256.

Platt, J. R. (1964). Strong inference. *Science, 146*, 347–353.

Plomin, R., & Daniels, D. (1987). Why are children from the same family so different from one another? *Behavioral and Brain Sciences, 10*, 1–60.

Pocock, S. J. (1983). *Clinical trials: A practical approach*. Chichester, England: Wiley.

Pocock, S. J., & Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics, 31*, 103–115.

Polanyi, M. (1958). *Personal knowledge: Toward a post-critical philosophy*. London: Routledge & Kegan Paul.

Polsby, N. W. (1984). *Political innovation in America: The politics of policy initiation*. New Haven, CT: Yale University Press.

Popper, K. R. (1959). *The logic of scientific discovery*. New York: Basic Books.

Population Council. (1986). *An experimental study of the efficiency and effectiveness of an IUD insertion and back-up component* (English summary; Report No. PC PES86). Lima, Peru: Author.

Potvin, L., & Campbell, D. T. (1996). *Exposure opportunity, the experimental paradigm and the case-control study*. Unpublished manuscript.

Pound, C. R., Partin, A. W., Eisenberger, M. A., Chan, D. W., Pearson, J. D., & Walsh, P. C. (1999). Natural history of progression after PSA elevation following radical prostatectomy. *Journal of the American Medical Association, 281*, 1591–1597.

Powers, K. I., & Anglin, M. D. (1993). Cumulative versus stabilizing effects of methadone maintenance: A quasi-experimental study using longitudinal self-report data. *Evaluation Review, 17*, 243–270.

Powers, E., & Witmer, H. (1951). *An experiment in the prevention of delinquency*. New York: Columbia University Press.

Premack, S. L., & Hunter, J. E. (1988). Individual unionization decisions. *Psychological Bulletin, 103*, 223–234.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin, 112*, 160–164.

Presby, S. (1978). Overly broad categories obscure important differences between therapies. *American Psychologist, 33*, 514–515.

Pressman, J. L., & Wildavsky, A. (1984). *Implementation* (3rd ed.). Berkeley: University of California Press.

Project MATCH Research Group (1993). Project MATCH: Rationale and methods for a multisite clinical trial matching patients to alcoholism treatment. *Alcoholism: Clinical and Experimental Research, 17,* 1130–1145.

Protection of Human Subjects, Title 45 C.F.R. Part 46, Subparts A–D (Revised 1991).

Psaty, B. M., Koepsell, T. D., Lin, D., Weiss, N. S., Siscovick, D. S., Rosendaal, F. R., Pahor, M., & Furberg, C. D. (1999). Assessment and control for confounding by indication in observational studies. *Journal of the American Geriatric Society, 47,* 749–754.

Puma, M. J., Burstein, N. R., Merrell, K., & Silverstein, G. (1990). *Evaluation of the Food Stamp Employment and Training Program. Final report: Volume I.* Bethesda, MD: Abt.

Quine, W. V. (1951). Two dogmas of empiricism. *Philosophical Review, 60,* 20–43.

Quine, W. V. (1969). *Ontological relativity and other essays.* New York: Columbia University Press.

Quirk, P. J. (1986). Public policy [Review of *Political Innovation in America* and *Agendas, Alternatives, and Public Policies*]. *Journal of Policy Analysis and Management, 6,* 607–613.

Ralston, D. A., Anthony, W. P., & Gustafson, D. J. (1985). Employees may love flextime, but what does it do to the organization's productivity? *Journal of Applied Psychology, 70,* 272–279.

Ranstam, J., Buyse, M., George, S. L., Evans, S., Geller, N. L., Scherrer, B., Lasaffre, E., Murray, G., Edler, L., Hutton, J. L., Colton, T., & Lachenbruch, P. (2000). Fraud in medical research: An international survey of biostatisticians. *Controlled Clinical Trials, 21,* 415–427.

Raudenbush, S. W. (1994). Random effects models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301–321). New York: Russell Sage Foundation.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized design. *Psychological Methods, 2,* 173–185.

Raudenbush, S. W., Becker, B. J., & Kalaian, H. (1988). Modeling multivariate effect sizes. *Psychological Bulletin, 103,* 111–120.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5,* 199–213.

Raudenbush, S. W., & Willms, J. D. (Eds.). (1991). *Schools, classrooms, and pupils: International studies of schooling from a multilevel perspective.* San Diego, CA: Academic Press.

Raudenbush, S. W., & Willms, J. D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics, 20,* 307–335.

Rauma, D., & Berk, R. A. (1987). Remuneration and recidivism: The long-term impact of unemployment compensation on ex-offenders. *Journal of Quantitative Criminology, 3,* 3–27.

Ray, J. W., & Shadish, W. R. (1996). How interchangeable are different estimators of effect size? *Journal of Consulting and Clinical Psychology, 64,* 1316–1325. (See also "Correction to Ray and Shadish (1996)," *Journal of Consulting and Clinical Psychology, 66,* 532, 1998)

Reding, G. R., & Raphelson, M. (1995). Around-the-clock mobile psychiatric crisis intervention: Another effective alternative to psychiatric hospitalization. *Community Mental Health Journal, 31,* 179–187.

Reed, J. G., & Baxter, P. M. (1994). Using reference databases. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 57–70). New York: Russell Sage Foundation.

Rees, A. (1974). The graduated work incentive experiment: An overview of the labor-supply results. *Journal of Human Resources, 9,* 158–180.

Reichardt, C. S. (1985). Reinterpreting Seaver's (1973) study of teacher expectancies as a regression artifact. *Journal of Educational Psychology, 77,* 231–236.

Reichardt, C. S. (1991). Comments on "The application of time series methods to moderate span longitudinal data." In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions* (pp. 88–91). Washington, DC: American Psychological Association.

Reichardt, C. S. (2000). A typology of strategies for ruling out threats to validity. In L. Bickman (Ed.), *Research design: Donald Campbell's legacy* (Vol. 2, pp. 89–115). Thousand Oaks, CA: Sage.

Reichardt, C. S., & Gollob, H. F. (1986). Satisfying the constraints of causal modeling. In W. M. K. Trochim (Ed.), *Advances in quasi-experimental design and analysis* (pp. 91–107). San Francisco: Jossey-Bass.

Reichardt, C. S., & Gollob, H. F. (1987, October). *Setting limits on the bias due to omitted variables.* Paper presented at the meeting of the Society of Multivariate Experimental Psychology, Denver, CO.

Reichardt, C. S., & Gollob, H. F. (1997). When confidence intervals should be used instead of statistical tests, and vice versa. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 259–284). Hillsdale, NJ: Erlbaum.

Reichardt, C. S., & Rallis, S. F. (Eds.). (1994). *The qualitative-quantitative debate: New perspectives.* San Francisco: Jossey-Bass.

Reichardt, C. S., Trochim, W. M. K., & Cappelleri, J. C. (1995). Reports of the death of the regression-discontinuity design are greatly exaggerated. *Evaluation Review, 19,* 39–63.

Reicken, H. W., Boruch, R. F., Campbell, D. T., Caplan, N., Glennan, T. K., Pratt, J. W., Rees, A., & Williams, W. (1974). *Social experimentation: A method for planning and evaluating social intervention.* New York: Academic Press.

Reinsel, G. (1993). *Elements of multivariate time series analysis.* New York: Springer-Verlag.

Rescher, N. (1969). *Introduction to value theory.* Englewood Cliffs, NJ: Prentice-Hall.

Revelle, W., Humphreys, M. S., Simon, L., & Gilliland, K. (1980). The interactive effect of personality, time of day, and caffeine: A test of the arousal model. *Journal of Experimental Psychology: General, 109,* 1–31.

Reynolds, A. J., & Temple, J. A. (1995). Quasi-experimental estimates of the effects of a preschool intervention: Psychometric and econometric comparisons. *Evaluation Review, 19,* 347–373.

Reynolds, K. D., & West, S. G. (1987). A multiplist strategy for strengthening nonequivalent control group designs. *Evaluation Review, 11,* 691–714.

Rezmovic, E. L., Cook, T. J., & Dobson, L. D. (1981). Beyond random assignment: Factors affecting evaluation integrity. *Evaluation Review, 5,* 51–67.

Ribisl, K. M., Walton, M. A., Mowbray, C. T., Luke, D. A., Davidson, W. S., & Bootsmiller, B. J. (1996). Minimizing participant attrition in panel studies through the use of effective retention and tracking strategies: Review and recommendations. *Evaluation and Program Planning, 19,* 1–25.

Richet, C. (1884). La suggestion mentale et le calcul des probabilites. *Revue Philosophique de la France et de l'Etranger, 18,* 609–674.

Rindskopf, D. (1986). New developments in selection modeling for quasi-experimentation. In W. M. K. Trochim (Ed.), *Advances in quasi-experimental design and analysis* (pp. 79–89). San Francisco: Jossey-Bass.

Rindskopf, D. (2000). Plausible rival hypotheses in measurement, design, and scientific theory. In L. Bickman (Ed.), *Research design: Donald Campbell's legacy* (Vol. 1, pp. 1–12). Thousand Oaks, CA: Sage.

Rindskopf, D. M. (1981). Structural equation models in analysis of nonexperimental data. In R. F. Boruch, P. M. Wortman, & D. S. Cordray (Eds.), *Reanalyzing program evaluations* (pp. 163–193). San Francisco: Jossey-Bass.

Rivlin, A. M., & Timpane, P. M. (1975). (Eds.). *Planned variation in education.* Washington, DC: Brookings.

Robbins, H., & Zhang, C.-H. (1988). Estimating a treatment effect under biased sampling. *Proceedings of the National Academy of Science, USA, 85,* 3670–3672.

Robbins, H., & Zhang, C.-H. (1989). Estimating the superiority of a drug to a placebo when all and only those patients at risk are treated with the drug. *Proceedings of the National Academy of Science, USA, 86,* 3003–3005.

Robbins, H., & Zhang, C.-H. (1990). *Estimating a treatment effect under biased allocation.* (Working paper.) New Brunswick, NJ: Rutgers University, Institute of Biostatistics and Department of Statistics.

Roberts, J. V., & Gebotys, R. J. (1992). Reforming rape laws: Effects of legislative change in Canada. *Law and Human Behavior, 16,* 555–573.

Robertson, T. S., & Rossiter, J. R. (1976). Short-run advertising effects on children: A field study. *Journal of Marketing Research, 8,* 68–70.

Robins, J. M. (1998). Correction for non-compliance in equivalence trials. *Statistics in Medicine, 17,* 269–302.

Robins, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In M. E. Halloran & D. Berry (Eds.), *Statistical models in epidemiology, the environment, and clinical trials* (pp. 95–133). New York: Springer-Verlag.

Robins, J. M., & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology, 3,* 143–155.

Robins, J. M., & Greenland, S. (1996). Comment. *Journal of the American Statistical Association, 91,* 456–458.

Robins, J. M., Greenland, S., & Hu, F.-C. (1999). Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome (with discussion). *Journal of the American Statistical Association, 94,* 687–712.

Robinson, A., Bradley, R. D., & Stanley, T. D. (1990). Opportunity to achieve: Identifying and serving mathematically talented black students. *Contemporary Educational Psychology, 15,* 1–12.

Robinson, A., & Stanley, T. D. (1989). Teaching to talent: Evaluating an enriched and accelerated mathematics program. *Journal of the Education of the Gifted, 12,* 253–267.

Robinson, L. A., Berman, J. S., & Neimeyer, R. A. (1990). Psychotherapy for the treatment of depression: A comprehensive review of controlled outcome research. *Psychological Bulletin, 108,* 30–49.

Rockette, H. E. (1993). What evidence is needed to link lung cancer to second-hand smoke? *Chance: New Directions for Statistics and Computing, 6,* 15–18.

Roese, N. J., & Jamieson, D. W. (1993). Twenty years of bogus pipeline research: A critical review and meta-analysis. *Psychological Bulletin, 114,* 363–375.

Roethlisberger, F. S., & Dickson, W. J. (1939). *Management and the worker.* Cambridge, MA: Harvard University Press.

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113,* 553–565.

Rogers, P. J., Hacsi, T. A., Petrosino, A., & Huebner, T. A. (Eds.). (2000). *Program theory in evaluation: Challenges and opportunities.* San Francisco: Jossey-Bass.

Rogosa, D. (1980). A critique of cross-lagged correlation. *Psychological Bulletin, 88,* 245–258.

Rogosa, D. (1988). Myths about longitudinal research. In K. W. Schaie, R. T. Campbell, W. Meredith, & S. C. Rawlings (Eds.), *Methodological issues in aging research* (pp. 171–210). New York: Springer.

Rosch, E. H. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale, NJ: Erlbaum.

Rosenbaum, P. R. (1984). From association to causation in observational studies: The role of tests of strongly ignorable treatment assumptions. *Journal of the American Statistical Association, 79,* 41–48.

Rosenbaum, P. R. (1986). Dropping out of high school in the United States: An observational study. *Journal of Educational Statistics, 11,* 207–224.

Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation tests in matched observational studies. *Biometrika, 74,* 13–26.

Rosenbaum, P. R. (1988). Sensitivity analysis for matching with multiple controls. *Biometrika, 75,* 577–581.

Rosenbaum, P. R. (1989). Sensitivity analysis for matched observational studies with many ordered treatments. *Scandinavian Journal of Statistics, 16,* 227–236.

Rosenbaum, P. R. (1991a). Discussing hidden bias in observational studies. *Annals of Internal Medicine, 115,* 901–905.

Rosenbaum, P. R. (1991b). Sensitivity analysis for matched case-control studies. *Biometrics, 47,* 87–100.

Rosenbaum, P. R. (1993). Hodges-Lehmann point estimates of treatment effect in observational studies. *Journal of the American Statistical Association, 88,* 1250–1253.

Rosenbaum, P. R. (1995a). *Observational studies.* New York: Springer-Verlag.

Rosenbaum, P. R. (1995b). Quantiles in nonrandom samples and observational studies. *Journal of the American Statistical Association, 90,* 1424–1431.

Rosenbaum, P. R. (1996a). Comment. *Journal of the American Statistical Association, 91,* 465–468.

Rosenbaum, P. R. (1996b). Observational studies and nonrandomized experiments. In S. Ghosh & C. R. Rao (Eds.), *Handbook of statistics* (Vol. 13, pp. 181–197). Amsterdam: Elsevier Science.

Rosenbaum, P. R. (1998). Multivariate matching methods. In S. Kotz, N. L. Johnson, L. Norman, & C. B. Read (Eds.), *Encyclopedia of statistical sciences* (Update Volume 2, pp. 435–438). New York: Wiley.

Rosenbaum, P. R. (1999a). Choice as an alternative to control in observational studies. *Statistical Science, 14,* 259–304.

Rosenbaum, P. R. (1999b). Using quantile averages in matched observational studies. *Applied Statistics, 48,* 63–78.

Rosenbaum, P. R. (in press). Observational studies. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of social and behavioral sciences.* Oxford, England: Elsevier Science.

Rosenbaum, P. R., & Krieger, A. (1990). Sensitivity analysis for matched case-control studies. *Journal of the American Statistical Association, 85,* 493–498.

Rosenbaum, P. R., & Rubin, D. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79,* 516–524.

Rosenberg, M., Adams, D. C., & Gurevitch, J. (1997). *MetaWin: Statistical software for meta-analysis with resampling tests.* (Available from Sinauer Associates, Inc., P.O. Box 407, Sunderland, MA 01375-0407)

Rosenberg, M. J. (1969). The conditions and consequences of evaluation apprehension. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research* (pp. 279–349). New York: Academic Press.

Rosenberger, W. F. (1999). Randomized play-the-winner clinical trials: Review and recommendations. *Controlled Clinical Trials, 20,* 328–342.

Rosenthal, M. C. (1994). The fugitive literature. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 85–94). New York: Russell Sage Foundation.

Rosenthal, R. (1956). *An attempt at the experimental induction of the defense mechanism of projection.* Unpublished doctoral dissertation, University of California, Los Angeles.

Rosenthal, R. (1966). *Experimenter effects in behavioral research.* New York: Appleton-Century-Crofts.

Rosenthal, R. (1973a). The mediation of Pygmalion effects: A four-factor theory. *Papua New Guinea Journal of Education, 9,* 1–12.

Rosenthal, R. (1973b). *On the social psychology of the self-fulfilling prophecy: Further evidence for Pygmalion effects and their mediating mechanisms.* New York: MSS Modular Publication, Module 53.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86,* 638–641.

Rosenthal, R. (1986). Meta-analytic procedures and the nature of replication: The ganzfeld debate. *Journal of Parapsychology, 50,* 315–336.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis.* New York: McGraw-Hill.

Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences, 3,* 377–386.

Rosenthal, R., & Rubin, D. B. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin, 99,* 400–406.

Rosenthal, R., & Rubin, D. B. (1994). The counternull value of an effect size: A new statistic. *Psychological Science, 5,* 329–334.

Rosenzweig, S. (1933). The experimental situation as a psychological problem. *Psychological Review, 40,* 337–354.

Rosnow, R. L., & Rosenthal, R. (1989). Definition and interpretation of interaction effects. *Psychological Bulletin, 105,* 143–146.

Rosnow, R. L., & Rosenthal, R. (1996). Contrasts and interaction effects redux: Five easy pieces. *Psychological Science, 7,* 253–257.

Rosnow, R. L., & Rosenthal, R. (1997). *People studying people: Artifacts and ethics in behavioral research.* New York: Freeman.

Ross, A. S., & Lacey, B. (1983). A regression discontinuity analysis of a remedial education programme. *Canadian Journal of Higher Education, 13,* 1–15.

Ross, H. L. (1973). Law, science and accidents: The British Road Safety Act of 1967. *Journal of Legal Studies, 2,* 1–75.

Ross, H. L., Campbell, D. T., & Glass, G. V. (1970). Determining the social effects of a legal reform: The British "breathalyser" crackdown of 1967. *American Behavioral Scientist, 13,* 493–509.

Rossi, P. H. (1987). The iron law of evaluation and other metallic rules. In J. Miller & M. Lewis (Eds.), *Research in social problems and public policy* (Vol. 4, pp. 3–20). Greenwich, CT: JAI Press.

Rossi, P. H. (1995). Doing good and getting it right. In W. R. Shadish, D. L.Newman, M. A. Scheirer, & C. Wye (Eds.), *Guiding principles for evaluators* (pp. 55–59). San Francisco: Jossey-Bass.

Rossi, P. H., Berk, R. A., & Lenihan, K. J. (1980). *Money, work and crime: Some experimental findings.* New York: Academic Press.

Rossi, P. H., & Freeman, H. E. (1989). *Evaluation: A systematic approach* (4th ed). Thousand Oaks, CA: Sage.

Rossi, P. H., Freeman, H. E., & Lipsey, M. W. (1999). *Evaluation: A systematic approach* (6th ed.). Thousand Oaks, CA: Sage.

Rossi, P. H., & Lyall, K. C. (1976). *Reforming public welfare.* New York: Russell Sage.

Rossi, P. H., & Lyall, K. C. (1978). An overview evaluation of the NIT Experiment. In T. D. Cook, M. L. DelRosario, K. M. Hennigan, M. M. Mark, & W. M. K. Trochim (Eds.), *Evaluation studies review annual* (Volume 3, pp. 412–428). Newbury Park, CA: Sage.

Rossi, P. H., & Wright, J. D. (1984). Evaluation research: An assessment. *Annual Review of Sociology, 10,* 331–352.

Rossi, P. H., Wright, J. D., & Anderson, A. B. (Eds.). (1983). *Handbook of survey research.* San Diego, CA: Academic Press.

Rothman, K. (1986). *Modern epidemiology.* Boston: Little Brown.

Rouanet, H. (1996). Bayesian methods for assessing importance of effects. *Psychological Bulletin, 119,* 149–158.

Rouse, C. E. (1998). Private school vouchers and student achievement: An evaluation of the Milwaukee Parental Choice Program. *Quarterly Journal of Economics, 113,* 553–602.

Rowe, P. M. (1999). What is all the hullabaloo about endostatin? *Lancet, 353,* 732.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin, 57,* 416–428.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66,* 688–701.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics, 2,* 1–26.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics, 6,* 34–58.

Rubin, D. B. (1986). Comment: Which ifs have causal answers. *Journal of the American Statistical Association, 81,* 961–962.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Rubin, D. B. (1990). A new perspective. In K. W. Wachter & M. L. Straf (Eds.), *The future of meta-analysis* (pp. 155–165). New York: Russell Sage Foundation.

Rubin, D. B. (1991). Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics, 47,* 1213–1234.

Rubin, D. B. (1992a). Clinical trials in psychiatry: Should protocol deviation censor patient data? A comment. *Neuropsychopharmacology, 6,* 59–60.

Rubin, D. B. (1992b). Meta-analysis: Literature synthesis or effect-size surface estimation? *Journal of Educational Statistics, 17,* 363–374.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine, 127,* 757–763.

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics, 52,* 249–264.

Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association, 95,* 573–585.

Rubins, H. B. (1994). From clinical trials to clinical practice: Generalizing from participant to patient. *Controlled Clinical Trials, 15,* 7–10.

Rudd, P., Ahmed, S., Zachary, V., Barton, C., & Bonduelle, D. (1990). Improved compliance measures: Applications in an ambulatory hypertensive drug trial. *Clinical Pharmacological Therapy, 48,* 676–685.

Ryle, G. (1971). *Collected papers* (Vol. 2). London: Hutchinson.

Sackett, D. L. (1979). Bias in analytic research. *Journal of Chronic Diseases, 32,* 51–63.

Sackett, D. L. (2000). Why randomized controlled trials fail but needn't: 1. Failure to gain "coal-face" commitment and to use the uncertainty principle. *Canadian Medical Association Journal, 162,* 1311–1314.

Sackett, D. L., & Haynes, R. B. (1976). *Compliance with therapeutic regimens.* Baltimore: Johns Hopkins University Press.

Sackett, D. L., & Hoey, J. (2000). Why randomized controlled trials fail but needn't: A new series is launched. *Canadian Medical Association Journal, 162,* 1301–1302.

Sacks, H. S., Chalmers, T. C., & Smith, H. (1982). Randomized versus historical controls for clinical trials. *American Journal of Medicine, 72,* 233–240.

Sacks, H. S., Chalmers, T. C., & Smith, H. (1983). Sensitivity and specificity of clinical trials: Randomized v historical controls. *Archives of Internal Medicine, 143,* 753–755.

St. Pierre, R. G., Cook, T. D., & Straw, R. B. (1981). An evaluation of the Nutrition Education and Training Program: Findings from Nebraska. *Evaluation and Program Planning, 4,* 335–344.

St. Pierre, R. G., Ricciuti, A., & Creps, C. (1998). *Summary of state and local Even Start evaluations.* Cambridge, MA: Abt.

Sales, B. D., & Folkman, S. (Eds.). (2000). *Ethics in research with human participants.* Washington, DC: American Psychological Association.

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world.* Princeton, NJ: Princeton University Press.

Salmon, W. C. (1989). *Four decades of scientific explanation.* Minneapolis: University of Minnesota Press.

Salner, M. (1989). Validity in human science research. In S. Kvale (Ed.), *Issues of validity in qualitative research* (pp. 47–71). Lund, Sweden: Studentlitteratur.

Sampson, R. J., Raudenbush, S. W., & Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science, 277,* 918–924.

Sanchez-Meca, J., & Marin-Martinez, F. (1998). Weighting by inverse variance or by sample size in meta-analysis: A simulation study. *Educational and Psychological Measurement, 58,* 211–220.

Sarason, S. B. (1978). The nature of problem solving in social action. *American Psychologist, 33,* 370–380.

Saretsky, G. (1972). The OEO PC experiment and the John Henry effect. *Phi Delta Kappan, 53,* 579–581.

Sargent, D. J., Sloan, J. A., & Cha, S. S. (1999). Sample size and design considerations for Phase II clinical trials with correlated observations. *Controlled Clinical Trials, 19,* 242–252.

Saunders, L. D., Irwig, L. M., Gear, J. S., & Ramushu, D. L. (1991). A randomized controlled trial of compliance improving strategies in Soweto hypertensives. *Medical Care, 29,* 669–678.

Sayrs, L. W. (1989). *Pooled time series analysis.* Thousand Oaks, CA: Sage.

Scarr, S. (1997). Rules of evidence: A larger context for statistical debate. *Psychological Science, 8,* 16–17.

Schaffer, S. (1989). Glass works: Newton's prisms and the uses of experiment. In D. Gooding, T. Pinch, & S. Schaffer (Eds.), *The uses of experiment: Studies in the natural sciences* (pp. 105–114). Cambridge, England: Cambridge University Press.

Schaffner, K. F. (Ed.). (1986). Ethical issues in the use of clinical controls. *Journal of Medical Philosophy, 11,* 297–404.

Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association, 94,* 1096–1120.

Schlesselman, J. J. (1982). *Case-control studies: Design, conduct, analysis.* New York: Oxford University Press.

Schmidt, D., & Leppik, I. E. (Eds.). (1988). *Compliance in epilepsy.* Amsterdam: Elsevier.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1,* 115–129.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62,* 529–540.

Schmitt, F. F. (1995). *Truth: A primer.* Boulder, CO: Westview Press.

Schoenberg, R. (1989). Covariance structure models. *Annual Review of Sociology, 15,* 425–440.

Schonemann, P. H. (1991). In praise of randomness. *Behavioral and Brain Sciences, 14,* 162–163.

Schulz, K. F. (1995). Subverting randomization in controlled trials. *Journal of the American Medical Association, 274,* 1456–1458.

Schulz, K. F., Chalmers, I., Hayes, R. J., & Altman, D. G. (1995). Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Journal of the American Medical Association, 273,* 408–412.

Schumacher, J. E., Milby, J. B., Raczynski, J. M., Engle, M., Caldwell, E. S., & Carr, J. A. (1994). Demoralization and threats to validity in Birmingham's Homeless Project. In K. J. Conrad (Ed.), *Critically evaluating the role of experiments* (pp. 41–44). San Francisco: Jossey-Bass.

Schweinhart, L. J., Barnes, H. V., & Weikart, D. P. (1993). *Significant benefits: The High/Scope Perry Preschool study through age 27.* (Available from High/Scope Foundation, 600 North River Street, Ypsilanti, MI 48198)

Scientists quibble on calling discovery "planets." (2000, October 6). *The Memphis Commercial Appeal,* p. A5.

Scott, A. G., & Sechrest, L. (1989). Strength of theory and theory of strength. *Evaluation and Program Planning, 12,* 329–336.

Scriven, M. (1976). Maximizing the power of causal investigation: The Modus Operandi method. In G. V. Glass (Ed.), *Evaluation studies review annual* (Vol. 1, pp. 101–118). Newbury Park, CA: Sage.

Scriven, M. (1980). *The logic of evaluation.* Inverness, CA: Edgepress.

Scriven, M. (1991). *Evaluation thesaurus* (4th ed.). Thousand Oaks, CA: Sage.

Seamon, F., & Feiock, R. C. (1995). Political participation and city/county consolidation: Jacksonville-Duval County. *International Journal of Public Administration, 18,* 1741–1752.

Seaver, W. B. (1973). Effects of naturally induced teacher expectancies. *Journal of Personality and Social Psychology, 28,* 333–342.

Seaver, W. B., & Quarton, R. J. (1976). Regression-discontinuity analysis of dean's list effects. *Journal of Educational Psychology, 66,* 459–465.

Sechrest, L., West, S. G., Phillips, M. A., Redner, R., & Yeaton, W. (1979). Some neglected problems in evaluation research: Strength and integrity of research. In L. Sechrest, S. G. West, M. A. Phillips, R. Redner, & W. Yeaton (Eds.), *Evaluation studies review annual* (Vol. 4, pp. 15–35). Beverly Hills, CA: Sage.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309–316.

Seligman, M. E. P. (1969). Control group and conditioning: A comment on operationalism. *Psychological Review, 76,* 484–491.

SenGupta, S. (1995). A similarity-based single study approach to construct and external validity. *Dissertation Abstracts International, 55*(11), 3458A. (University Microfilms No. 9509453)

Serlin, R. A., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 199–228). Hillsdale, NJ: Erlbaum.

Shadish, W. R. (1984). Policy research: Lessons from the implementation of deinstitutionalization. *American Psychologist, 39,* 725–738.

Shadish, W. R. (1989). Critical multiplism: A research strategy and its attendant tactics. In L. Sechrest, H. Freeman, & A. Mulley (Eds.), *Health services research methodology: A focus on AIDS* (DHHS Publication No. PHS 89-3439, pp. 5–28). Rockville, MD: U.S. Department of Health and Human Services, Public Health Service, National Center for Health Services Research and Health Care Technology Assessment.

Shadish, W. R. (1992a). Do family and marital psychotherapies change what people do? A meta-analysis of behavioral outcomes. In T. D. Cook, H. M. Cooper, D. S. Cordray, H. Hartmann, L. V. Hedges, R. J. Light, T. A. Louis, & F. Mosteller (Eds.), *Meta-analysis for explanation: A casebook* (pp. 129–208). New York: Russell Sage Foundation.

Shadish, W. R. (1992b, August). *Mediators and moderators in psychotherapy meta-analysis.* Paper presented at the aunual convention of the American Psychological Association, Washington, DC.

Shadish, W. R. (1994). Critical multiplism: A research strategy and its attendant tactics. In L. B. Sechrest & A. J. Figueredo (Eds.), *New directions for program evaluation* (pp. 13–57). San Francisco: Jossey-Bass.

Shadish, W. R. (1995a). Philosophy of science and the quantitative-qualitative debates: Thirteen common errors. *Evaluation and Program Planning, 18,* 63–75.

Shadish, W. R. (1995b). The logic of generalization: Five principles common to experiments and ethnographies. *American Journal of Community Psychology, 23,* 419–428.

Shadish, W. R. (1996). Meta-analysis and the exploration of causal mediating processes: A primer of examples, methods, and issues. *Psychological Methods, 1,* 47–65.

Shadish, W. R. (2000). The empirical program of quasi-experimentation. In L. Bickman (Ed.), *Validity and social experimentation: Donald Campbell's legacy* (pp. 13–35). Thousand Oaks, CA: Sage.

Shadish, W. R., & Cook, T. D. (1999). Design rules: More steps towards a complete theory of quasi-experimentation. *Statistical Science, 14,* 294–300.

Shadish, W. R., Cook, T. D., & Houts, A. C. (1986). Quasi-experimentation in a critical multiplist mode. In W. M. K. Trochim (Ed.), *Advances in quasi-experimental design and analysis* (pp. 29–46). San Francisco: Jossey-Bass.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice.* Newbury Park, CA: Sage.

Shadish, W. R., Doherty, M., & Montgomery, L. M. (1989). How many studies are in the file drawer? An estimate from the family/marital psychotherapy literature. *Clinical Psychology Review, 9,* 589–603.

Shadish, W. R., & Fuller, S. (Eds.). (1994). *The social psychology of science*. New York: Guilford Press.

Shadish, W. R., Fuller, S., Gorman, M. E., Amabile, T. M., Kruglanski, A. W., Rosenthal, R., & Rosenwein, R. E. (1994). Social psychology of science: A conceptual and research program. In W. R. Shadish & S. Fuller (Eds.), *Social psychology of science* (pp. 3–123). New York: Guilford Press.

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 261–281). New York: Russell Sage Foundation.

Shadish, W. R., & Heinsman, D. T. (1997). Experiments versus quasi-experiments: Do you get the same answer? In W. J. Bukoski (Ed.), *Meta-analysis of drug abuse prevention programs* (NIDA Research Monograph, DHHS Publication No. ADM 97-170, pp. 147–164). Washington, DC: Superintendent of Documents.

Shadish, W. R., Hu, X., Glaser, R. R., Kownacki, R. J., & Wong, T. (1998). A method for exploring the effects of attrition in randomized experiments with dichotomous outcomes. *Psychological Methods, 3*, 3–22.

Shadish, W. R., Matt, G. E., Navarro, A. M., & Phillips, G. (2000). The effects of psychological therapies under clinically representative conditions: A meta-analysis, *Psychological Bulletin, 126*, 512–529.

Shadish, W. R., Matt, G., Navarro, A., Siegle, G., Crits-Christoph, P., Hazelrigg, M., Jorm, A., Lyons, L. S., Nietzel, M. T., Prout, H. T., Robinson, L., Smith, M. L., Svartberg, M., & Weiss, B. (1997). Evidence that therapy works in clinically representative conditions. *Journal of Consulting and Clinical Psychology, 65*, 355–365.

Shadish, W. R., Montgomery, L. M., Wilson, P., Wilson, M. R., Bright, I., & Okwumabua, T. (1993). The effects of family and marital psychotherapies: A meta-analysis. *Journal of Consulting and Clinical Psychology, 61*, 992–1002.

Shadish, W. R., Newman, D. L., Scheirer, M. A., & Wye, C. (1995). Developing the guiding principles. In W. R. Shadish, D. L. Newman, M. A. Scheirer, & C. Wye (Eds.), *Guiding principles for evaluators* (pp. 3–18). San Francisco: Jossey-Bass.

Shadish, W. R., & Ragsdale, K. (1996). Random versus nonrandom assignment in psychotherapy experiments: Do you get the same answer? *Journal of Consulting and Clinical Psychology, 64*, 1290–1305.

Shadish, W. R., & Reis, J. (1984). A review of studies of the effectiveness of programs to improve pregnancy outcome. *Evaluation Review, 8*, 747–776.

Shadish, W. R., Robinson, L., & Lu, C. (1999). *ES: A computer program and manual for effect size calculation.* St. Paul, MN: Assessment Systems Corporation.

Shadish, W. R., Silber, B., Orwin, R. G., & Bootzin, R. R. (1985). The subjective well-being of mental patients in nursing homes. *Evaluation and Program Planning, 8*, 239–250.

Shadish, W. R., Straw, R. B., McSweeny, A. J., Koller, D. L., & Bootzin, R. R.(1981). Nursing home care for mental patients: Descriptive data and some propositions. *American Journal of Community Psychology, 9*, 617–633.

Shadish, W. R., & Sweeney, R. (1991). Mediators and moderators in meta-analysis: There's a reason we don't let dodo birds tell us which psychotherapies should have prizes. *Journal of Consulting and Clinical Psychology, 59*, 883–893.

Shapin, S. (1994). *A social history of truth: Civility and science in seventeenth-century England.* Chicago: University of Chicago Press.

Shapiro, A. K., & Shapiro, E. (1997). *The powerful placebo.* Baltimore: Johns Hopkins University Press.

Shapiro, J. Z. (1984). The social costs of methodological rigor: A note on the problem of massive attrition. *Evaluation Review, 8,* 705–712.

Sharpe, T. R., & Wetherbee, H. (1980). *Final report: Evaluation of the Improved Pregnancy Outcome Program.* Tupelo, MS: Mississippi State Board of Health, Three Rivers District Health Department.

Shaw, R. A., Rosati, M. J., Salzman, P., Coles, C. R., & McGeary, C. (1997). Effects on adolescent ATOD behaviors and attitudes of a 5-year community partnership. *Evaluation and Program Planning, 20,* 307–313.

Sherif, J., Harvey, O. J., White, B. J., Hood, W. R., & Sherif, C. W. (1961). *Intergroup conflict and cooperation: The Robbers Cave Experiment.* Norman: University of Oklahoma Book Exchange.

Sherman, L. W., & Berk, R. A. (1985). The randomization of arrest. In R. F. Boruch & W. Wothke (Eds.), *Randomization and field experimentation* (pp. 15–25). San Francisco: Jossey-Bass.

Shih, W. J., & Quan, H. (1997). Testing for treatment differences with dropouts present in clinical trials: A composite approach. *Statistics in Medicine, 16,* 1225–1239.

Shoham-Salomon, V., & Rosenthal, R. (1987). Paradoxical interventions: A meta-analysis. *Journal of Consulting and Clinical Psychology, 55,* 22–28.

Shönemann, P. H. (1991). In praise of randomness. *Behavioral and Brain Sciences, 14,* 162–163.

Shonkoff, J. P., & Phillips, D. A. (Eds.). (2000). *From neurons to neighborhoods: The science of early childhood development.* Washington, DC: National Academy Press.

Shrout, P. E. (1997). Should significance tests be banned? Introduction to a special section exploring the pros and cons. *Psychological Science, 8,* 1–2.

Shumaker, S. A., & Rejeski, W. J. (Eds.). (2000). Adherence to behavioral and pharmacological interventions in clinical research in older adults [Special issue]. *Controlled Clinical Trials, 21*(5S).

Shumway, R. H. (1988). *Applied statistical time series analysis.* Englewood Cliffs, NJ: Prentice-Hall.

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology.* Boston: Authors Cooperative.

Sieber, J. E. (1992). *Planning ethically responsible research: A guide for students and internal review boards.* Newbury Park, CA: Sage.

Siegel, A. E., & Siegel, S. (1957). Reference groups, membership groups, and attitude change. *Journal of Abnormal and Social Psychology, 55,* 360–364.

Siemiatycki, J. (1989). Friendly control bias. *Journal of Clinical Epidemiology, 42,* 687–688.

Siemiatycki, J., Colle, S., Campbell, S., Dewar, R., & Belmonte, M. M. (1989). Case-control study of insulin dependent (type I) diabetes mellitus. *Diabetes Care, 12,* 209–216.

Silka, L. (1989). *Intuitive judgments of change.* New York: Springer-Verlag.

Silliman, N. P. (1997). Hierarchical selection models with applications in meta-analysis. *Journal of the American Statistical Association, 92,* 926–936.

Silverman, W. A. (1977). The lesson of retrolental fibroplasia. *Scientific American, 236,* 100–107.

Simes, R. J. (1987). Confronting publication bias: A cohort design for meta-analysis. *Statistics in Medicine, 6,* 11–29.

Simon, H. A. (1976). *Administrative behavior.* New York: Free Press.

Simpson, J. M., Klar, N., & Donner, A. (1995). Accounting for cluster randomization: A review of primary prevention trials, 1990 through 1993. *American Journal of Public Health, 85,* 1378–1383.

Skinner, B. F. (1961). *Cumulative record.* New York: Appleton-Century-Crofts.

Slavin, R. E. (1986). Best-evidence synthesis: An alternative to meta-analysis and traditional reviews. *Educational Researchers, 15,* 5–11.

Smith, B., & Sechrest, L. (1991). Treatment of aptitude × treatment interactions. *Journal of Consulting and Clinical Psychology, 59,* 233–244.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts.* Cambridge, MA: Harvard University Press.

Smith, E. R. (1992). Beliefs, attributions, and evaluations: Nonhierarchical models of mediation in social cognition. *Journal of Personality and Social Psychology, 43,* 248–259.

Smith, G. (1997). Do statistics test scores regress toward the mean? *Chance, 10,* 42–45.

Smith, H. (1997). Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology, 27,* 325–353.

Smith, M. L. (1980). Publication bias and meta-analysis. *Evaluation in Education, 4,* 22–24.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32,* 752–760.

Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy.* Baltimore: Johns Hopkins University Press.

Smith, N. L. (Ed.). (1992). *Varieties of investigative journalism.* San Francisco: Jossey-Bass.

Smoot, S. L. (1989). Meta-analysis of single subject research in education: A comparison of four metrics (Doctoral dissertation, Georgia State University). *Dissertation Abstracts International, 50*(07), 1928A.

Snow, R. E. (1991). Aptitude-treatment interactions as a framework for research on individual differences in psychotherapy. *Journal of Consulting and Clinical Psychology, 59,* 205–216.

Snowdon, C., Elbourne, D., & Garcia, J. (1999). Zelen randomization: Attitudes of parents participating in a neonatal clinical trial. *Controlled Clinical Trials, 20,* 149–171.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London: Sage.

Snyder, D. K., & Wills, R. M. (1989). Behavioral versus insight-oriented marital therapy: Effects on individual and interspousal functioning. *Journal of Consulting and Clinical Psychology, 57,* 39–46.

Snyder, D. K., Wills, R. M., & Grady-Fletcher, A. (1991). Long-term effectiveness of behavioral versus insight-oriented marital therapy: A 4-year follow-up study. *Journal of Consulting and Clinical Psychology, 59,* 138–141.

Sobel, M. E. (1993). Causal inference in the social and behavioral sciences. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook for statistical modeling in the social and behavioral sciences* (pp. 1–38). New York: Plenum.

Sobel, M. E. (2000). Causal inference in the social sciences. *Journal of the American Statistical Association, 95,* 647–650.

Solomon, R. L. (1949). An extension of control group design. *Psychological Bulletin, 46,* 137–150.

Sommer, A., & Zeger, S. L. (1991). On estimating efficacy from clinical trials. *Statistics in Medicine, 10,* 45–52.

Sommer, B. (1987). The file drawer effect and publication rates in menstrual cycle research. *Psychology of Women Quarterly, 11,* 233–242.

Sorensen, G., Emmons, K., Hunt, M., & Johnston, D. (1998). Implications of the results of community trials. *Annual Review of Public Health, 19,* 379–416.

Speer, D. C. (1994). Can treatment research inform decision makers? Nonexperimental method issues and examples among older outpatients. *Journal of Consulting and Clinical Psychology, 62,* 560–568.

Speer, D. C., & Swindle, R. (1982). The "monitoring model" and the mortality × treatment interaction threat to validity in mental health outcome evaluation. *American Journal of Community Psychology, 10,* 541–552.

Spiegelhalter, D. J., Freedman, L. S., & Blackburn, P. R. (1986). Monitoring clinical trials: Conditional or predictive power? *Controlled Clinical Trials, 7,* 8–17.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction and search* (2nd ed.). Cambridge, MA: MIT Press.

Spirtes, P., Scheines, R., & Glymour, C. (1990a). Reply to comments. *Sociological Methods and Research, 19,* 107–121.

Spirtes, P., Scheines, R., & Glymour, C. (1990b). Simulation studies of the reliability of computer-aided model specification using the TETRAD II, EQS, and LISREL programs. *Sociological Methods and Research, 19,* 3–66.

Spitz, H. H. (1993). Were children randomly assigned in the Perry Preschool Project? *American Psychologist, 48,* 915.

Stadthaus, A. M. (1972). A comparison of the subsequent academic achievement of marginal selectees and rejectees for the Cincinnati public schools special college preparatory program: An application of Campbell's regression discontinuity design (Doctoral dissertation, University of Cincinnati, 1972). *Dissertation Abstracts International, 33*(06), 2825A.

Staines, G. L., McKendrick, K., Perlis, T., Sacks, S., & DeLeon, G. (1999). Sequential assignment and treatment-as-usual: Alternatives to standard experimental designs in field studies of treatment efficacy. *Evaluation Review, 23,* 47–76.

Stake, R. E., & Trumbull, D. J. (1982). Naturalistic generalizations. *Review Journal of Philosophy and Social Science, 7,* 1–12.

Stanley, B., & Sieber, J. E. (Eds.). (1992). *Social research on children and adolescents: Ethical issues.* Newbury Park, CA: Sage.

Stanley, T. D. (1991). "Regression-discontinuity design" by any other name might be less problematic. *Evaluation Review, 15,* 605–624.

Stanley, T. D., & Robinson, A. (1990). Sifting statistical significance from the artifact of regression-discontinuity design. *Evaluation Review, 14,* 166–181.

Stanley, W. D. (1987). Economic migrants or refugees from violence? A time series analysis of Salvadoran migration to the United States. *Latin American Law Review, 12,* 132–154.

Stanton, M. D., & Shadish, W. R. (1997). Outcome, attrition and family-couples treatment for drug abuse: A meta-analysis and review of the controlled, comparative studies. *Psychological Bulletin, 122,* 170–191.

Starfield, B. (1977). Efficacy and effectiveness of primary medical care for children. In *Harvard Child Health Project, Children's medical care needs and treatment: Report of the Harvard Child Health Project.* Cambridge, MA: Ballinger.

Statistical Solutions. (1998). *SOLAS for Missing Data Analysis 1.0* [Computer software]. (Available from Statistical Solutions, 8 South Bank, Crosse's Green, Cork, Ireland)

Steering Committee of the Physicians' Health Study Research Group. (1988). Preliminary report: Findings from the aspirin component of the ongoing physicians' health study. *New England Journal of Medicine, 318,* 271–295.

Stein, M. A., & Test, L. I. (1980). Alternative to mental hospital treatment: I. Conceptual model, treatment program, clinical evaluation. *Archives of General Psychiatry, 37,* 392–397.

Steiner, M. S., & Gingrich, J. R. (2000). Gene therapy for prostate cancer: Where are we now? *Journal of Urology, 164,* 1121–1136.

Stelzl, I. (1986). Changing a causal hypothesis without changing the fit: Some rules for generating equivalent path models. *Multivariate Behavioral Research, 21,* 309–331.

Stevens, S. J. (1994). Common implementation issues in three large-scale social experiments. In K. J. Conrad (Ed.), *Critically evaluating the role of experiments* (pp. 45–53). San Francisco: Jossey-Bass.

Stewart, A. L., Sherbourne, C. D., Wells, K. B., Burnam, M. A., Rogers, W. H., Hays, R. D., & Ware, J. E. (1993). Do depressed patients in different treatment settings have different levels of well-being and functioning? *Journal of Consulting and Clinical Psychology, 61,* 849–857.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900.* Cambridge, MA, and London, England: Harvard University/Belnap Press.

Stock, W. A. (1994). Systematic coding for research synthesis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 126–138). New York: Russell Sage Foundation.

Stolzenberg, R., & Relles, D. (1990). Theory testing in a world of constrained research design: The significance of Heckman's censored sampling bias correction for nonexperimental research. *Sociological Methods and Research, 18,* 395–415.

Stone, R. A., Obrosky, D. S., Singer, D. E., Kapoor, W. N., Fine, M. J., & The Pneumonia Patient Outcomes Research Team (PORT) Investigators. (1995). Propensity score adjustment for pretreatment differences between hospitalized and ambulatory patients with community-acquired pneumonia. *Medical Care, 33* (Suppl.), AS56–AS66.

Stout, R. L., Brown, P. J., Longabaugh, R., & Noel, N. (1996). Determinants of research follow-up participation in an alcohol treatment outcome trial. *Journal of Consulting and Clinical Psychology, 64,* 614–618.

Stromsdorfer, E. W., & Farkas, G. (Eds.). (1980). *Evaluation studies review annual* (Vol. 5). Beverly Hills, CA: Sage.

Stroup, D. F., Berlin, J. A., Morton, S. C., Olkin, I., Williamson, G. D., Rennie, D., Moher, D., Becker, B. J., Sipe, T. A., Thacker, S. B., for the Meta-Analysis of Observational Studies in Epidemiology (MOOSE) Group. (2000). Meta-analysis of observational studies in epidemiology: A proposal for reporting. *Journal of the American Medical Association, 283,* 2008–2012.

Sullivan, C. M., Rumptz, M. H., Campbell, R., Eby, K. K., & Davidson, W. S. (1996). Retaining participants in longitudinal community research: A comprehensive protocol. *Journal of Applied Behavioral Science, 32,* 262–276.

Swanson, H. L., & Sachselee, C. (2000). A meta-analysis of single-subject-design intervention research for students with LD. *Journal of Learning Disabilities, 33,* 114–136.

Tallmadge, G. K., & Horst, D. P. (1976). *A procedural guide for validating achievement gains in educational projects.* (Evaluation in Education Monograph No. 2). Washington, DC: U.S. Department of Health, Education, and Welfare.

Tallmadge, G. K., & Wood, C. T. (1978). *User's guide: ESEA Title I evaluation and reporting system.* Mountain View, CA: RMC Research.

Tamura, R. N., Faries, D. E., Andersen, J. S., & Heiligenstein, J. H. (1994). A case study of an adaptive clinical trial in the treatment of out-patients with depressive disorder. *Journal of the American Statistical Association, 89,* 768–776.

Tan, M., & Xiong, X. (1996). Continuous and group sequential conditional probability ratio tests for Phase II clinical trials. *Statistics in Medicine, 15,* 2037–2051.

Tanaka, J. S., & Huba, G. J. (1984). Confirmatory hierarchical factor analysis of psychological distress measures. *Journal of Personality and Social Psychology, 46,* 621–635.

Tanaka, J. S., Panter, A. T., Winborne, W. C., & Huba, G. J. (1990). Theory testing in personality and social psychology with structural equation models: A primer in 20 questions. In C. Hendrick & M. S. Clark (Eds.), *Research methods in personality and social psychology* (pp. 217–242). Newbury Park, CA: Sage.

Taylor, S. J., & Bogdan, R. (1984). *Introduction to qualitative research methods: The search for meanings.* New York: Wiley.

Taylor, G. J., Rubin, R., Tucker, M., Greene, H. L., Rudikoff, M. D., & Weisfeldt, M. L. (1978). External cardiac compression: A randomized comparison of mechanical and manual techniques. *Journal of the American Medical Association, 240,* 644–646.

Teague, M. L., Bernardo, D. J., & Mapp, H. P. (1995). Farm-level economic analysis incorporating stochastic environmental risk assessment. *American Journal of Agricultural Economics, 77,* 8–19.

Tebes, J. K., Snow, D. L., & Arthur, M. W. (1992). Panel attrition and external validity in the short-term follow-up study of adolescent substance use. *Evaluation Review, 16,* 151–170.

[illegible] M. D., Burrows, K. A., & LaChance-McCullough, M. L. (1995).

Test, M. A., & Burke, S. S. (1985). Random assignment of chronically mentally ill persons to hospital or community treatment. In R. F. Boruch & W. Wothke (Eds.), *Randomization and field experimentation* (pp. 81–93). San Francisco: Jossey-Bass.

Test, M. A., & Stein, L. I. (1980). Alternative to mental hospital treatment: III. Social cost. *Archives of Geneal Psychiatry, 37,* 409–412.

Tester, K. (1993). *The life and times of post-modernity.* London: Routledge.

Thistlewaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology, 51,* 309–317.

Thomas, A., Spiegelhalter, D. J., & Gilks, W. R. (1992). BUGS: A program to perform Bayesian inference using Gibbs sampling. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4.* Oxford, England: Clarendon Press.

Thomas, G. B. (1997). A program evaluation of the remedial and developmental studies program at Tennessee State University (Doctoral dissertation, Vanderbilt University, 1997). *Dissertation Abstracts International, 58*(08), 3042A.

Thomas, L., & Krebs, C. J. (1997). A review of statistical power analysis software. *Bulletin of the Ecological Society of America, 78,* 128–139.

Thompson, B. (1993). Statistical significance testing in contemporary practice: Some proposed alternatives with comments from journal editors [Special issue]. *Journal of Experimental Education, 61*(4).

Thompson, D. C., Rivara, F. P., & Thompson, R. (2000). Helmets for preventing head and facial injuries in bicyclists (Cochrane Review). *The Cochrane Library,* Issue 3. Oxford, England: Update Software.

Thompson, S. K. (1992). *Sampling.* New York: Wiley.

Tilden, V. P., & Shepherd, P. (1987). Increasing the rate of identification of battered women in an emergency department: Use of a nursing protocol. *Research in Nursing and Health, 10,* 209–215.

Time series database of U.S. and international statistics ready for manipulation. (1992). *Database Searcher, 8,* 27–29.

Tinbergen, J. (1956). *Economic policy principles and design.* Amsterdam: North-Holland.

Tong, H. (1990). *Non-linear time series: A dynamical system approach.* New York: Oxford University Press.

Toulmin, S. E. (1961). *Foresight and understanding: An inquiry into the aims of science.* Bloomington: Indiana University Press.

Tracey, T. J., Sherry, P., & Keitel, M. (1986). Distress and help-seeking as a function of person-environment fit and self-efficacy: A causal model. *American Journal of Community Psychology, 14,* 657–676.

Trend, M. G. (1979). On the reconciliation of qualitative and quantitative analyses: A case study. In T. D. Cook & C. S. Reichardt (Eds.), *Qualitative and quantitative methods in evaluation research* (pp. 68–86). Newbury Park, CA: Sage.

Triplett, N. (1898). The dynamogenic factors in pacemaking and competition. *American Journal of Psychology, 9,* 507–533.

Trochim, W. M. K. (1980). *The regression-discontinuity design in Title I evaluation: Implementation, analysis, and variations.* Unpublished doctoral dissertation, Northwestern University, Evanston, IL.

Trochim, W. M. K. (1984). *Research design for program evaluation: The regression-discontinuity approach.* Newbury Park, CA: Sage.

Trochim, W. M. K. (1985). Pattern matching, validity, and conceptualization in program evaluation. *Evaluation Review, 9,* 575–604.

Trochim, W. M. K. (1990). The regression discontinuity design. In L. Sechrest, E. Perrin, & J. Bunker (Eds.), *Research methodology: Strengthening causal interpretations of nonexperimental data* (pp. 119–140). Rockville, MD: Public Health Service, Agency for Health Care Policy and Research.

Trochim, W. M. K., & Campbell, D. T. (1996). *The regression point displacement design for evaluating community-based pilot programs and demonstration projects.* Unpublished manuscript. (Available from the Department of Policy Analysis and Management, Cornell University, Room N136C, MVR Hall, Ithaca, NY 14853)

Trochim, W. M. K., & Cappelleri, J. C. (1992). Cntoff assignment strategies for enhancing randomized clinical trials. *Controlled Clinical Trials, 13,* 190–212.

Trochim, W. M. K., Cappelleri, J. C., & Reichardt, C. S. (1991). Random measurement error does not bias the treatment effect estimate in the regression-discontinuity design: II. When an interaction effect is present. *Evaluation Review, 15,* 571–604.

Tufte, E. R. (1983). *The visual display of quantitative information.* Cheshire, CT: Graphics Press.

Tufte, E. R. (1990). *Envisioning information.* Cheshire, CT: Graphics Press.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6,* 100–116.

Turpin, R. S., & Sinacore, J. M. (Eds.). (1991). *Multisite evaluations.* San Francisco: Jossey-Bass.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131.

Uebel, T. E. (1992). *Overcoming logical positivism from within: The emergence of Neurath's naturalism in the Vienna Circle's protocol sentence debate.* Amsterdam and Atlanta, GA: Editions Rodopi B.V.

Valins, S., & Baum, A. (1973). Residential group size, social interaction, and crowding. *Environment and Behavior, 5,* 421–439.

Varnell, S., Murray, D. M., & Baker, W. L. (in press). An evaluation of analysis options for the one group per condition design: Can any of the alternatives overcome the problems inherent in this design? *Evaluation Review.*

Veatch, R., & Sollitto, S. (1973). Human experimentation: The ethical questions persist. *Hastings Center Report, 3,* 1–3.

Velicer, W. F. (1994). Time series models of individual substance abusers. In L. M. Collins & L. A. Seitz (Eds.), *Advances in data analysis for prevention intervention research* (NIDA Research Monograph No. 142, pp. 264–299). Rockville MD: National Institute on Drug Abuse.

Velicer, W. F., & Harrop, J. (1983). The reliability and accuracy of time series model identification. *Evaluation Review, 7,* 551–560.

Veney, J. E. (1993). Evaluation applications of regression analysis with time-series data. *Evaluation Practice, 14,* 259–274.

Verbeek, M., & Nijman, T. (1992). Testing for selectivity bias in panel data models. *International Economic Review, 33,* 681–703.

Vinokur, A. D., Price, R. H., & Caplan, R. D. (1991). From field experiments to program implementation: Assessing the potential outcomes of an experimental intervention program for unemployed persons. *American Journal of Community Psychology, 19,* 543–562.

Vinovskis, M. A. (1998). *Changing federal strategies for supporting educational research, development and statistics.* Unpublished manuscript, National Educational Research Policy and Priorities Board, U.S. Department of Education.

Virdin, L. M. (1993). *A test of the robustness of estimators that model selection in the nonequivalent control group design.* Unpublished doctoral dissertation, Arizona State University, Tempe.

Vessey, M. P. (1979). Comment. *Journal of Chronic Diseases, 32,* 64–66.

Visser, R. A., & deLeeuw, J. (1984). Maximum likelihood analysis for a generalized regression-discontinuity design. *Journal of Educational Statistics, 9,* 45–60.

Viswesvaran, C., & Schmidt, F. L. (1992). A meta-analytic comparison of the effectiveness of smoking cessation methods. *Journal of Applied Psychology, 77,* 554–561.

Vosniadou, S., & Ortony, A. (1989). Similarity and analogical reasoning: A synthesis. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 1–17). New York: Cambridge University Press.

Wagenaar, A. C., & Holder, H. D. (1991). A change from public to private sale of wine: Results from natural experiments in Iowa and West Virginia. *Journal of Studies on Alcohol, 52,* 162–173.

Wagoner, J. L. (1992). The contribution of group therapy to the successful completion of probation for adult substance abusers. *Dissertation Abstracts International, 53*(03), 724A. (University Microfilms No. AAC92-20873)

Wainer, H. (Ed.). (1986). *Drawing inferences from self-selected samples.* New York: Springer-Verlag.

Wallace, L. W. (1987). *The Community Penalties Act of 1983: An evaluation of the law, its implementation, and its impact in North Carolina.* Unpublished doctoral dissertation, University of Nebraska.

Wallace, P., Cutler, S., & Haines, A. (1988). Randomised controlled trial of general practitioner intervention in patients with excessive alcohol consumption. *British Medical Journal, 297,* 663–668.

Walther, B. J., & Ross, A. S. (1982). The effect on behavior of being in a control group. *Basic and Applied Social Psychology, 3,* 259–266.

Wampler, K. S., & Serovich, J. M. (1996). Meta-analysis in family therapy research. In D. H. Sprenkle & S. M. Moon (Eds.), *Research methods in family therapy* (pp. 286–303). New York: Guilford Press.

Wampold, B. E. (1992). The intensive examination of social interactions. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis* (pp. 93–131). Hillsdale, NJ: Erlbaum.

Wampold, B. E., & Furlong, M. J. (1981). The heuristics of visual inspection. *Behavioral Assessment, 3,* 79–92.

Wampold, B. E., Mondin, G. W., Moody, M., Stich, F., Benson, K., & Ahn, H. (1997). A meta-analysis of outcome studies comparing bona fide psychotherapies: Empirically, "All must have prizes." *Psychological Bulletin, 122,* 203–215.

Wampold, B. E., & Worsham, N. L. (1986). Randomization tests for multiple-baseline designs. *Behavioral Assessment, 8,* 135–143.

Wang, M. C., & Bushman, B. J. (1998). Using the normal quantile plot to explore meta-analytic data sets. *Psychological Methods, 3,* 46–54.

Wang, M. C., & Bushman, B. J. (1999). *Integrating results through meta-analytic review using SAS software.* Cary, NC: SAS Institute.

Washington v. Davis, 426 U.S. 229 (1976).

Watts, H., & Rees, A.W. (1976). *The New Jersey income-maintenance experiment: Vol. 2. Labor-supply responses.* New York: Academic Press.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). *Unobtrusive measures.* Skokie, IL: Rand McNally.

Webb, E. J., Campbell, D. T., Schwartz, R. D., Sechrest, L., & Grove, J.B. (1981). *Nonreactive measures in the social sciences* (2nd ed.). Boston, MA: Houghton Mifflin.

Webb, J. F., Khazen, R. S., Hanley, W. B., Partington, M. S., Percy, W. J. L., & Rathburn, J. C. (1973). PKU screening: Is it worth it? *Canadian Medical Association Journal, 108,* 328–329.

Weber, S. J., Cook, T. D., & Campbell, D. T. (1971). *The effects of school integration on the academic self-concept of public-school children.* Paper presented at the annual meeting of the Midwestern Psychological Association, Detroit, MI.

Wei, L. J. (1978). An application of an urn model to the design of sequential controlled trials. *Journal of the American Statistical Association, 73,* 559–563.

Wei, W. W. S. (1990). *Time series analysis: Univariate and multivariate methods.* Redwood City, CA: Addison-Wesley.

Weiss, B., Williams, J. H., Margen, S., Abrams, B., Caan, B., Citron, L. J., Cox, C., McKibben, J., Ogar, D., & Schultz, S. (1980). Behavioral responses to artificial food colors. *Science, 207,* 1487–1489.

Weiss, C. H. (1980). Knowledge creep and decision accretion. *Knowledge: Creation, Diffusion, Utilization, 1,* 381–404.

Weiss, C. H. (1988). Evaluation for decisions: Is anybody there? Does anybody care? *Evaluation Practice, 9,* 5–20.

Weiss, C. H. (1998). *Evaluation* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall.

Weiss, C. H., & Bucuvalas, M. J. (1980). *Social science research and decision-making.* New York: Columbia University Press.

Weisz, J. R., Weiss, B., & Donenberg, G. R. (1992). The lab versus the clinic: Effects of child and adolescent psychotherapy. *American Psychologist, 47,* 1578–1585.

Weisz, J. R., Weiss, B., & Langmeyer, D. B. (1987). Giving up on child psychotherapy: Who drops out? *Journal of Consulting and Clinical Psychology, 55,* 916–918.

Welch, W. P., Frank, R. G., & Costello, A. J. (1983). Missing data in psychiatric research: A solution. *Psychological Bulletin, 94,* 177–180.

West, S. G., Aiken, L. S., & Todd, M. (1993). Probing the effects of individual componenrs in multiple component prevention programs. *American Journal of Community Psychology, 21,* 571–605.

West, S. G., Biesanz, J., & Pitts, S. C. (2000). Causal inference and generalization in field settings: Experimental and quasi-experimental designs. In H. T. Reis & C. M. Judd

(Eds.), *Handbook of research methods in social psychology* (pp. 40–84). New York: Cambridge University Press.

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. Hoyle (Ed.), *Structural equation modeling: Issues and applications* (pp. 56–75). Thousand Oaks, CA: Sage.

West, S. G., & Hepworth, J. T. (1991). Statistical issues in the study of temporal data: Daily experiences. *Journal of Personality, 59,* 609–662.

West, S. G., Hepworth, J. T., McCall, M. A., & Reich, J. W. (1989). An evaluation of Arizona's July 1982 drunk driving law: Effects on the City of Phoenix. *Journal of Applied Social Psychology, 19,* 1212–1237.

West, S. G., & Sagarin, B. (2000). Subject selection and loss in randomized experiments. In L. Bickman (Ed.), *Research design: Donald Campell's legacy* (Vol. 2, pp. 117–154). Thousand Oaks, CA: Sage.

Westlake, W. J. (1988). Bioavailability and bioequivalence of pharmaceutical formulations. In K. E. Peace (Ed.), *Biopharmaceutical statistics for drug development* (pp. 329–352). New York: Dekker.

Westmeyer, H. (in press). Explanation in the social sciences. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences.* Oxford, England: Elsevier.

White, D. M., Rusch, F. R., Kazdin, A. E., & Hartmann, D. P. (1989). Applications of meta-analysis in individual-subject research. *Behavioral Assessment, 11,* 281–296.

White, H. (1981). Consequences and detection of misspecified nonlinear regression models. *Journal of the American Statistical Association, 76,* 419–433.

White, H. D. (1994). Scientific communication and literature retrieval. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 41–55). New York: Russell Sage Foundation.

White, L., Tursky, B., & Schwartz, G. E. (Eds.). (1985). *Placebo: Theory, research, and mechanisms.* New York: Guilford Press.

White, P. A. (1990). Ideas about causation in philosophy and psychology. *Psychological Bulletin, 108,* 3–18.

Whittier, J.G. (1989). *The works of John Greenleaf Whittier* (Vol. 1). New York: Houghton Mifflin.

Widom, C. S., Weiler, B. L., & Cottler, L. B. (1999). Childhood victimization and drug abuse: A comparison of prospective and retrospective findings. *Journal of Consulting and Clinical Psychology, 67,* 867–880.

Wilcox, R. R. (1995). ANOVA: A paradigm for low power and misleading measures of effect size? *Review of Educational Research, 65,* 51–77.

Wilcox, R. R. (1996). *Statistics for the social sciences.* New York: Academic Press.

Wilder, C. S. (1972, July). *Physician visits, volume, and interval since last visit, U.S., 1969* (Series 10, No. 75; DHEW Pub. No. HSM 72-1064). Rockville, MD: National Center for Health Statistics.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Willett, J. B. (1988). Questions and answers in the measurement of change. *Review of Research in Education, 15*, 345–422.

Willett, J. B., & Singer, J. D. (1991). From whether to when: New methods for studying student dropout and teacher attrition. *Review of Educational Research, 61*, 407–450.

Williams, S. V. (1990). Regression-discontinuity design in health evaluation. In L. Sechrest, B. Perrin, & J. Bunker (Eds.), *Health services research methodology: Strengthening causal interpretations of nonexperimental data* (DHHS Publication No. PHS 90–3454, pp. 145–149). Rockville, MD: U.S. Department of Health and Human Services.

Willms, J. D. (1992). *Monitoring school performance: A guide for educators.* Washington, DC: The Falmer Press.

Willner, P. (1991). Methods for assessing the validity of animal models of human psychopathology. In A. A. Boulton, G. B. Baker, & M. T. Martin-Iverson (eds.), *Neuromethods* (Vol. 18, pp. 1–23). Clifton, NJ: Humana Press.

Willson, V. L., & Putnam, R. R. (1982). A meta-analysis of pretest sensitization effects in experimental design. *American Educational Research Journal, 19*, 249–258.

Wilson, E. B. (1952). *An introduction to scientific research.* New York: McGraw-Hill.

Wilson, M. C., Hayward, R. S. A., Tunis, S. R., Bass, E. B. & Guyatt, G. (1995). Users' guides to the medical literature: Part 8. How to use clinical practice guidelines. B. What are the recommendations and will they help you in caring for your patients? *Journal of the American Medical Association, 274*, 1630–1632.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). New York: McGraw-Hill.

Winship, C., & Mare, R. D. (1992). Models for sample selection bias. *Annual Review of Sociology, 18*, 327–350.

Winship, C., & Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology, 25*, 659–707.

Winston, A. S. (1990). Robert Sessions Woodworth and the "Columbia Bible": How the psychological experiment was redefined. *American Journal of Psychology, 103*, 391–401.

Winston, A. S., & Blais, D. J. (1996). What counts as an experiment?A transdisciplinary analysis of textbooks, 1930–1970. *American Journal of Psychology, 109*, 599–616.

Winston, P. H., & Horn, B. K. P. (1989). *LISP* (3rd ed.). Reading, MA: Addison-Wesley.

Witte, J. F. (1998). The Milwaukee voucher experiment. *Educational Evaluation and Policy Analysis, 20*, 229–251.

Witte, J. F. (1999). The Milwaukee voucher experiment: The good, the bad, and the ugly. *Phi Delta Kappan, 81*, 59–64.

Witte, J. F. (2000). Selective reading is hardly evidence. *Phi Delta Kappan, 81*, 391.

Wolchik, S. A., West, S. G., Westover, S., Sandler, I. N., Martin, A., Lustig, J., Tein, J.-Y., & Fisher, J. (1993). The Children of Divorce Parenting Intervention: Outcome evaluation of an empirically based program. *American Journal of Community Psychology, 21*, 293–331.

Wolcott, H. F. (1990). On seeking—and rejecting—validity in qualitative research. In E. W. Eisner & A. Peshkin (Eds.), *Qualitative inquiry in education: The continuing debate* (pp. 121–152). New York: Teachers College Press.

Wolf, F. M. (1986). *Meta-analysis: Quantitative methods for research synthesis.* Thousand Oaks, CA: Sage.

Wood, G. (1978). The knew-it-all-along effect. *Journal of Experimental Psychology: Human Perception and Performance, 4,* 345–353.

Woodworth, G. (1994). Managing meta-analytic data bases. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 177–189). New York: Russell Sage Foundation.

World Medical Association. (2000). Declaration of Helsink: Ethical principles for medical research involving human subjects. *Journal of the American Medical Association, 284,* 3043–3045.

Wortman, C. B., & Rabinowitz, V. C. (1979). Random assignment: The fairest of them all. In L. Sechrest, S. G. West, M. A. Phillips, R. Redner, & W. Yeaton (Eds.), *Evaluation studies review annual* (Vol. 4, pp. 177–184). Beverly Hills, CA: Sage.

Wortman, P. M. (1992). Lessons from the meta-analysis of quasi-experiments. In F. B. Bryant, J. Edwards, R. S. Tindale, E. J. Posavac, L. Heath, E. Henderson, & Y. Suarez-Balcazar (Eds.), *Methodological issues in applied social psychology* (pp. 65–81). New York: Plenum Press.

Wortman, P. M. (1994). Judging research quality. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 97–109). New York: Russell Sage Foundation.

Wortman, P. M., Reichardt, C. S., & St. Pierre, R. G. (1978). The first year of the education voucher demonstration. *Evaluation Quarterly, 2,* 193–214.

Wortman, P. M., Smyth, J. M., Langenbrunner, J. C., & Yeaton, W. H. (1998). Consensus among experts and research synthesis. *International Journal of Technology Assessment in Health Care, 14,* 109–122.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research, 20,* 557–585.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics, 5,* 161–215.

Yaremko, R. M., Harari, H., Harrison, R. C., & Lynn, E. (1986). *Handbook of research and quantitative methods in psychology for students and professionals.* Hillsdale, NJ: Erlbaum.

Yeaton, W. H., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology, 49,* 156–167.

Yeaton, W. H., & Sechrest, L. (1986). Use and misuse of no-difference findings in eliminating threats to validity. *Evaluation Review, 10,* 836–852.

Yeaton, W. H., & Wortman, P. M. (1993). On the reliability of meta-analytic reviews. *Evaluation Review, 17,* 292–309.

Yeaton, W. H., Wortman, P. M., & Langberg, N. (1983). Differential attrition: Estimating the effect of crossovers on the evaluation of a medical technology. *Evaluation Review, 7,* 831–840.

Yinger, J. (1995). *Closed doors, opportunities lost.* New York: Russell Sage Foundation.

Yu, J., & Cooper, H. (1983). A quantitative review of research design effects on response rates to questionnaires. *Journal of Marketing Research, 20,* 36–44.

Zabin, L. S., Hirsch, M. B., & Emerson, M. R. (1989). When urban adolescents choose abortion: Effects on education, psychological status, and subsequent pregnancy. *Family Planning Perspectives, 21,* 248–255.

Zadeh, L. A. (1987). *Fuzzy sets and applications.* New York: Wiley.

Zajonc, R. B., & Markus, H. (1975). Birth order and intellectual development. *Psychological Review, 82,* 74–88.

Zeisel, H. (1973). Reflections on experimental technique in the law. *Journal of Legal Studies, 2,* 107–124.

Zelen, M. (1974). The randomization and stratification of patients to clinical trials. *Journal of Chronic Diseases, 27,* 365–375.

Zelen, M. (1979). A new design for randomized clinical trials? *New England Journal of Medicine, 300,* 1242–1245.

Zelen, M. (1990). Randomized consent designs for clinical trials: An update. *Statistics in Medicine, 9,* 645–656.

Zhu, S. (1999). A method to obtain a randomized control group where it seems impossible. *Evaluation Review, 23,* 363–377.

Zigler, E., & Weikart, D. P. (1993). Reply to Spitz's comments. *American Psychologist, 48,* 915–916.

Zigulich, J. A. (1977). *A comparison of elementary school environments: Magnet schools versus traditional schools.* Unpublished doctoral dissertation, Northwestern University, Evanston, IL.

Zucker, D. M., Lakatos, E., Webber, L. S., Murray, D. M., McKinlay, S. M., Feldman, H. A., Kelder, S. H., & Nader, P. R. (1995). Statistical design of the child and adolescent trial for cardiovascular health (CATCH): Implications of cluster randomization. *Controlled Clinical Trials, 16,* 96–118.

Subject Index

Subject Index