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 the-

ory 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.