3

Construct Validity and External Validity

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

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

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

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

CONSTRUCT VALIDITY

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

> In measuring human height (or weight or lung capacity, for example), there is little disagreement about the meaning of the construct being measured, or about the units of measurement (e.g., centimeters, grams, cubic centimeters). . . . Measuring growth in psychological domains (e.g., vocabulary, quantitative reasoning, verbal memory, hand–eye coordination, self-regulation) is more problematic. Disagreement is more likely to arise about the definition of the constructs to be assessed. This occurs, in part, because there are often no natural units of measurement (i.e., nothing comparable to the use of inches when measuring height). (Shonkoff & Phillips, 2000, pp. 82–83)

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

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

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

out of different genera and species of plants and animals, and the discovery of the structure of genes" (Mark, 2000, p. 150)—though such taxonomic work is considerably more difficult in the social sciences, for reasons which we now discuss.

Why Construct Inferences Are a Problem

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

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

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

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

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

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

those elements. But in the Australian aboriginal language called Dyirbal, women, fire, and dangerous things are all part of one category.[1]

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

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

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

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

Assessment of Sampling Particulars

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

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

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

chosen to treat, say, time as a separate feature of each experiment, as we occasionally have in some of our past work. Such additions would not change the key point. Construct validity involves making inferences from assessments of *any* of the sampling particulars in a study to the higher-order constructs they represent.

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

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

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

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

These examples make clear that assessments of study particulars need not be done using formal multi-item scales—though the information obtained would often be better if such scales were used. Rather, assessments include any method for generating data about sampling particulars. They would include archival records, such as patient charts in psychiatric hospitals in which data on diagnosis and symptoms are often recorded by hand or U.S. Census Bureau records in which respondents indicated their racial and ethnic identities by checking a box. They would include qualitative observations, sometimes formal ones such as participant

observation or unstructured interviews conducted by a trained anthropologist but often simply the report of the research team who, say, describe a setting as a "poverty neighborhood" based on their personal observations of it as they drive to and from work each day. Assessments may even include some experimental manipulations that are designed to shed light on the nature of study operations, as when a treatment is compared with a placebo control to clarify the extent to which treatment *is* a placebo.

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

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

Threats to Construct Validity

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

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

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

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

Inadequate Explication of Constructs

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

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

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

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

Construct Confounding

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

Mono-Operation Bias

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

sources, the data can still be combined from all three. Then the investigator can test if expertise is effective despite whatever sources of heterogeneity are contained within the three particular operations.

Monomethod Bias

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

Confounding Constructs with Levels of Constructs

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

Treatment-Sensitive Factorial Structure

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

Reactive Self-Report Changes

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

Reactivity to the Experimental Situation

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

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

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

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

Experimenter Expectancies

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

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

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

Novelty and Disruption Effects

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

Compensatory Equalization

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

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

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

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

Compensatory Rivalry

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

Resentful Demoralization

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

Treatment Diffusion

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

Construct Validity, Preexperimental Tailoring, and Postexperimental Specification

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

of the multiple operations reflects multiple methods so that single-method confounds (e.g., self-report biases) can be better assessed.

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

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

EXTERNAL VALIDITY

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

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

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

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

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

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

On the other hand, it is also wrong to limit external validity to questions about as-yet-unstudied instances. Campbell and Stanley (1963) made no such distinction in their original formulation of external validity as asking the question, "to what populations, settings, treatment variables, and measurement variables

can this effect be generalized?" (p. 5). Indeed, one of the goals of their theory was to point out "numerous ways in which experiments can be made more valid externally" (p. 17). For example, they said that external validity was enhanced in single studies "if in the initial experiment we have demonstrated the phenomenon over a wide variety of conditions" (p. 19) and also enhanced by inducing "maximum similarity of experiments to the conditions of application" (p. 18). This goal of designing experiments to yield inferences that are "more valid externally" is not a novel concept. To the contrary, most experiments already test whether treatment effects hold over several different outcomes. Many also report whether the effect holds over different kinds of persons, although power is reduced as a sample is subdivided. Tests of effect stability over variations in treatments are limited to studies having multiple treatments, but these do occur regularly in the scientific literature (e.g., Wampold et al., 1997). And tests of how well causal relationships hold over settings also occur regularly, for example, in education (Raudenbush & Willms, 1995) and in large multisite medical and public health trials (Ioannidis et al., 1999).

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

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

both men and women. If the original data can be accessed, and if they were coded and stored in such a way that the analysis is possible, the question can be addressed by reanalyzing the data to test this interaction.

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

Threats to External Validity

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

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

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

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

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

Interaction of Causal Relationship with Units

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

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

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

Interaction of Causal Relationship Over Treatment Variations

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

Interaction of Causal Relationship with Outcomes

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

Interaction of Causal Relationship with Settings

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

Context-Dependent Mediation

Causal explanation is one of the five principles of causal generalization in the grounded theory we outlined in Chapter 1. Though we discuss this principle in more detail in Chapter 12, one part of explanation is identification of mediating processes. The idea is that studies of causal mediation identify the essential

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

Constancy of Effect Size Versus Constancy of Causal Direction

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

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

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

Random Sampling and External Validity

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

divided. So, although we argue in Chapters 1 and 11 that random sampling has major practical limitations when combined with experiments, its benefits for external validity are so great that it should be used on those rare occasions when it is feasible.

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

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

Purposive Sampling and External Validity

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

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

MORE ON RELATIONSHIPS, TRADEOFFS, AND PRIORITIES

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

The Relationship Between Construct Validity and External Validity

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

these common cases, knowledge of construct validity provides only weak evidence about external validity. We provide some examples of how this occurs in Chapters 11 through 13.

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

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

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

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

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

The Relationship Between Internal Validity and Construct Validity

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

Tradeoffs and Priorities

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

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

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

Internal Validity: A Sine Qua Non?

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

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

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

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

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

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

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

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

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

Is Descriptive Causation a Priority?

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

Can Nonexperimental Methods Give a Satisfactory Answer?

Even if a descriptive causal inference has been well justified as a high priority, experimental methods are still not the only choice. Descriptive causal questions can be studied nonexperimentally. This happens with correlational path analysis in sociology (e.g., Wright, 1921, 1934), with case-control studies in epidemiology (e.g., Ahlbom & Norell, 1990), or with qualitative methods such as case studies (e.g., Campbell, 1975). The decision to investigate a descriptive causal question using such methods depends on many factors. Partly these reflect disciplinary traditions that developed for either good or poor reasons. Some phenomena are simply not amenable to the manipulation that experimental work requires, and at other times manipulation may be undesirable for ethical reasons or for fear of changing the phenomenon being studied in undesirable ways. Sometimes the cause of interest is not yet sufficiently clear, so that interest is more in exploring a range of possible causes than in zeroing in on one or two of them. Sometimes the investment of time and resources that experiments may require is premature, perhaps because insufficient pilot work has been done to develop a treatment in terms of its theoretical fidelity and practical implementability, because crucial aspects of experimental procedures such as outcome measurement are underdeveloped, or because results are needed more quickly than an experiment can provide. Premature experimental work is a common research sin.

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

The Weak and Strong Senses of Sine Qua Non

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

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

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

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

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

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

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

Relatively few programs of research are this systematic. However, one might view the four validity types as a loose guide to programmatic experimentation, instructing the researcher to iterate back and forth among them as comparative weaknesses in generalized causal knowledge of one kind or another become apparent. For example, many researchers start a program of research by noticing an interesting relationship between two variables (e.g., McGuire, 1997). They may do further studies to confirm the size and dependability of the relationship (statistical conclusion validity), then study whether the relationship is causal (internal validity), then try to characterize it more precisely (construct validity) and to specify its boundaries (external validity). Sometimes, the phenomenon that piques an experimenter's curiosity already has considerable external validity; for instance, the covariation between smoking and lung cancer across different kinds of people in different settings and at different times led to a program of research designed to determine if the relationship was causal, to characterize its size and dependability,

and then to explain it. Other times, the construct validity of the variables has already been subject to much attention, but the question of a causal relationship between them suddenly attracts notice. For instance, the construct validity of both race and intelligence had already been extensively studied when a controversy arose in the 1990s over the possibility of a causal relationship between them (Devlin, 1997; Herrnstein & Murray, 1994). Programs of experimental research can start at many different points, with existing knowledge lending strength to different kinds of inferences and with need to repair knowledge weaknesses of many different kinds. Across a program of research, all validity types are a high priority. By the end of a program of research, each validity type should have had its turn in the spotlight.

SUMMARY

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

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

4

Quasi-Experimental Designs That Either Lack a Control Group or Lack Pretest Observations on the Outcome

Qua·si (kwā′zī′, sī′, kwä′zī, sī′): [Middle English as if, from Old French from Latin *quasi: quam,* as; see *kwo-* in Indo-European Roots + *s,* if; see *swo-* in Indo-European Roots.] adj. Having a likeness to something; resembling: *a quasi success.*

WE BEGIN this chapter (and subsequent ones) with a brief example that illustrates the kind of design being discussed. In 1966, the Canadian Province of Ontario began a program to screen and treat infants born with phenylketonuria (PKU) in order to prevent PKU-based retardation. An evaluation found that, after the program was completed, 44 infants born with PKU experienced no retardation, whereas only 3 infants showed evidence of retardation—and of these three, two had been missed by the screening program (Webb et al., 1973). Statistics from prior years showed a higher rate of retardation due to PKU. Although the methodology in this study was quite primitive, particularly because it lacked a control group[1] that did not receive treatment, the authors concluded that the program successfully prevented PKU-based retardation. Subsequently, such programs were widely adopted in Canada and the United States and are still seen as extremely effective. How did this study achieve such a clear, correct, and useful conclusion when it used neither a control group nor random assignment? That is the topic of this chapter.

1. The term *control group* usually refers to a group that does not receive treatment; the more general term *comparison group* may include both control groups and alternative treatment groups.