
From T. D. Cook + (1979) 2
D. T. CAMPBELL

Quasi-Experimentation

Validity

We shall use the concepts *validity* and *invalidity* to refer to the best available approximation to the truth or falsity of propositions, including propositions about cause. In keeping with the discussion in chapter 1, we should always use the modifier “approximately” when referring to validity, since one can never know what is true. At best, one can know what has not yet been ruled out as false. Hence, when we use the terms valid and invalid in the rest of this book, they should always be understood to be prefaced by the modifiers “approximately” or “tentatively.”

One could invoke many types of validity when trying to develop a framework in which to understand experiments in complex field settings. Campbell and Stanley (1963) invoked two, which they called “internal” and “external” validity. Internal validity refers to the approximate validity with which we infer that a relationship between two variables is causal or that the absence of a relationship implies the absence of cause. External validity refers to the approximate validity with which we can infer that the presumed causal relationship can be generalized to and across alternate measures of the cause and effect and across different types of persons, settings, and times.

For convenience, we shall further subdivide the two validity types of Campbell and Stanley. Covariation is a necessary condition for inferring cause, and practicing scientists begin by asking of their data: “Are the presumed independent and dependent variables related?” Therefore, it is useful to consider the particular reasons why we can draw false conclusions about covariation. We shall call these reasons (which are threats to valid inference-making) threats to *statistical conclusion validity*, for conclusions about covariation are made on the basis of statistical evidence. (This type of validity was listed by Campbell [1969] as a threat to internal validity. It was called “instability” and was concerned with drawing false conclusions about population covariation from unstable sample data. We shall later consider “instability” as one of the major threats to statistical conclusion validity.)

If a decision is made on the basis of sample data that two variables are related, then the practicing researcher's next question is likely to be: "Is there a *causal* relationship from variable *A* to variable *B*, where *A* and *B* are manipulated or measured variables (operations) rather than the theoretical or otherwise generalized constructs they are meant to represent?" To answer this question, the researcher has to rule out a variety of other reasons for the relationship, including the threat that *B* causes *A* and the threat that *C* causes both *A* and *B*. The first of these threats is usually handled easily in experiments, as we shall see later. The latter is not so easily dealt with, especially in quasi-experiments. Much of the researcher's task involves self-consciously thinking through and testing the plausibility of noncausal reasons why the two variables might be related and why "change" might have been observed in the dependent variable even in the absence of any explicit treatment of theoretical or practical significance. We shall use the term *internal validity* to refer to the validity with which statements can be made about whether there is a causal relationship from one variable to another in the form in which the variables were manipulated or measured.

Internal validity has nothing to do with the abstract labeling of a presumed cause or effect; rather, it deals with the *relationship* between the research operations *irrespective of what they theoretically represent*. However, researchers would like to be able to give their presumed cause and effect operations names which refer to theoretical constructs. The need for this is most explicit in theory-testing research where the operations are explicitly derived to represent theoretical notions. But applied researchers also like to give generalized abstract names to their variables, for it is hardly useful to assume that the relationship between the two variables is causal if one cannot summarize these variables other than by describing them in exhaustive operational detail. Whether one wants to test theory about the effects of "dissonance" on "attitude change," or is interested in policy issues relating to "school desegregation" and "academic achievement," one wants to be able to make generalizations about higher-order terms that have a referent in explicit theory or everyday abstract language. Following the lead of Cronbach and Meehl (1955) in the area of measurement, we shall use the term *construct validity of causes or effects* to refer to the approximate validity with which we can make generalizations about higher-order constructs from research operations. Extending their usage, we shall use the term to refer to manipulated independent variables as well as measured traits. We shall base inferences about constructs more on the fit between operations and conceptual definitions than on the fit between obtained data patterns and theoretical predictions about such data patterns—more on what Campbell (1960) called trait validity than what Cronbach and Meehl (1955) termed nomological validity. We shall not ignore nomological validity, however.

The construct validity of causes and effects was listed by Campbell and Stanley (1963) under the heading of "external validity," and it is what experimentalists mean when they refer to inadvertent "confounding."¹ (That is, was the effect

¹Confounding is sometimes done deliberately in more complex experimental designs, e.g., Latin squares or incomplete lattice designs. Such deliberate confounding is meant to achieve efficiency at the cost of reduced interpretability for some carefully chosen interactions that are considered implausible or that are of little theoretical or practical significance.

due to the planned variable X , or was X confounded with “experimenter expectancies” or a “Hawthorne effect,” or was X a “negative incentive” rather than “dissonance arousal”?) As such, construct validity had to do with generalization, with the question: “Can I generalize from this one operation or set of operations to a referent construct?” Given this grounding in the need to generalize, it is not difficult to see why Campbell and Stanley linked generalizing to abstract constructs with generalizing to (and across) populations of persons, settings, and historical moments. Just as one gains more information by knowing that a causal relationship is probably not limited to particular operational representations of a cause and effect, so one gains by knowing that the relationship (1) is not limited to a particular idiosyncratic sample of persons or settings of a given type, and (2) is not limited to a particular population of X s but also holds with populations of Y s and Z s. We shall use the term *external validity* to refer to the approximate validity with which conclusions are drawn about the generalizability of a causal relationship to and across populations of persons, settings, and times.

Our justification for restricting the discussion of validity to these four types is practical only, based on their apparent correspondence to four major decision questions that the practicing researcher faces. These are: (1) Is there a relationship between the two variables? (2) Given that there is a relationship, is it plausibly causal from one operational variable to the other or would the same relationship have been obtained in the absence of any treatment of any kind? (3) Given that the relationship is plausibly causal and is reasonably known to be from one variable to another, what are the particular cause and effect constructs involved in the relationship? and (4) Given that there is probably a causal relationship from construct A to construct B , how generalizable is this relationship across persons, settings, and times? As stated previously, each of these decision questions was implicit in Campbell and Stanley’s explication of validity, with the present statistical conclusion and internal validities being part of internal validity and with the present construct and external validities being part of external validity. All we have done here is to subdivide each validity type and try to make the differences among types explicit. We want to stress that our approach is entirely practical, being derived from our belief that practicing researchers need to answer each of the above questions in their work. There are no totally compelling logical reasons for the distinctions.

STATISTICAL CONCLUSION VALIDITY

Introduction

In evaluating any experiment, three decisions about covariation have to be made with the sample data on hand: (1) Is the study sensitive enough to permit reasonable statements about covariation? (2) If it is sensitive enough, is there any reasonable evidence from which to infer that the presumed cause and effect covary? and (3) if there is such evidence, how strongly do the two variables covary?

The first of these issues concerns statistical power. It is vital in reporting and planning experiments to analyze how much power one has to detect an

effect of a given magnitude with the variances and sample sizes on hand. In planning studies, the power analysis usually consists of discovering the sample size required for detecting an effect of desired magnitude, given the expected variances. The major practical difficulties besetting such an analysis are, first, obtaining agreement as to the magnitude of desired impact, and second, finding acceptable variance estimates (other data sets are often used for the second purpose). Once these difficulties are overcome, the required sample sizes can be computed according to formulae given in most statistical texts.

When an experiment has been completed, power analyses usually have a different function. The known variances and sample sizes are used to compute the magnitude of effect that could have been "reasonably" detected in the study at hand. ("Reasonably" is often taken to mean "with 95% confidence.") If the magnitude of the detectable effect seems low to most persons, one would tentatively conclude that the experiment was powerful and that the null hypothesis might be provisionally accepted. However, if the magnitude estimate seems high, then it is not clear whether the absence of covariation is due to the true absence of a relationship or to the experiment being so weak that reasonably sized true effects could not be detected. Power analyses are desirable in any report of a study where the major research conclusion is that one variable does not cause another.

In research probing causal hypothesis, statistical analyses are primarily used for deciding whether a presumed cause and effect covary. In many studies, a decision about covariation is made by comparing the degree of covariation and random error observed in the sample data to an a priori specified risk of being wrong in concluding that there is covariation. This risk is specified as a probability level (usually 5%), and we speak of setting α at .05. The 5% level is arbitrary, and if we want more protection against being wrong in claiming there is covariation, then one simply lowers the probability level. Observed relationships below the specified probability level are treated as though they are "true," while those above it are treated as though they are "false." However, as statistics texts make clear in their discussions of Type I and Type II error, we can be wrong in concluding that the population means of various treatment groups truly differ even when the obtained probability level is below the specified one. We can also be wrong in concluding that they do not differ relative to the variances when the probability level is above the specified level.

Many traditions have developed in the data analysis field for assessing, not whether covariation can be inferred, but the amount of covariation. When the scale on which the outcome variable is measured has some commonsense referent, effects can be expressed in terms of the treatment causing, say, an average increase in income of \$1,000 per annum per person, or a reduction in prison recidivism of 20% over two years. With other scales, magnitude estimates are more difficult to interpret, and one wonders what an average treatment effect of five points on some academic achievement test "means" when it cannot be validly translated into an estimate of, say, grade level or mental age. This is why some researchers do not like to express the magnitude of impact directly in terms of the original scale but prefer to estimate it indirectly as the additional variance that the treatment permits one to predict in the dependent variable.

Magnitude estimates have a considerable advantage over estimates of whether covariation can reasonably be inferred since they are less dependent on sample size. (Note that very small effects will be statistically significant given a large enough N of units.) This dependence on sample size leads many statisticians to prefer estimates of the magnitude of covariation over inferences about whether it is reasonable to presume any covariation at all. Nonetheless, as we shall use the term, *statistical conclusion validity* refers to inferences about whether it is reasonable to presume covariation given a specified α level and the obtained variances. As such, statistical conclusion validity seems more closely related to tests of statistical significance than to magnitude estimates. Our stress on statistical significance is because decisions about whether a presumed cause and effect covary logically precede decisions about how strongly they covary. Moreover, in most reports where magnitude estimates are given without corresponding statistical significance tests it is usually presumed that the estimates of, say, the difference between two means are statistically significant.

Fortunately, decisions about covariation and its magnitude are not independent, for statistical significance depends on the relationship between a magnitude estimate and a standard error. Consequently, information about whether covariation can be presumed and about how much can be presumed is almost always available somewhere in a research report, though each type of information is not always equally stressed.

We would prefer to see research reported in terms of both statistical significance and magnitude estimates that are bounded by confidence intervals. In this sense, research conclusions might take the following form: "The treatment caused a statistically significant increase in the income of families of four. In tests similar to the one conducted, the effect in 95% of the cases would be an increase of between \$400 and \$1,600 per year." Bounded magnitude estimates realistically reflect the sources of error in social research which limit our confidence in conclusions. If widely used, they would decrease the current dependence on speciously precise point estimates, e.g., "The average increase in income was \$1,000 per year." When sample sizes are small, it is particularly dangerous to rely on statistical significance as the sole editing mechanism that allows us to differentiate between magnitude estimates that are and are not worth treating as different from zero. With small samples, power analyses should be reported to illustrate the magnitude of the effect that could have been detected given the sample sizes, the variances obtained in the study, and the chosen α level.

Some Major Threats to Statistical Conclusion Validity

Below is a list of some threats to drawing valid inferences about whether two variables covary. Later we shall present lists of threats to the other kinds of validity. In listing threats, we have been guided by our own research experience and by our substantive reading about factors that can lead to spurious inferences. No list of threats is the perfect one; and ours outlines forces that we believe plausibly occur in basic or applied research in field settings. But though we consider each threat plausible, we do not believe that each operates

with equal frequency or that each affects outcome variables to the same degree. Indeed, empirical research has been conducted on identified threats to try to establish the conditions under which they are likely to affect responses so that data-based estimates of plausibility can be made. For instance, on the basis of a program of research into how pretest measurement affects behavior in laboratory experiments, Lana (1969) was led to conclude that pretest sensitization is probably less of a threat than was previously feared. In the last analysis, systematic research and carefully considered experience should influence the practicing scientist's concern about the likelihood of each of the threats we have identified and have provisionally labeled as plausible. Moreover, we anticipate that persons with an interest in basic research will see some of the listed threats as more plausible than persons who are interested in more applied work, and vice versa. We also anticipate that the threats we will discuss for statistical conclusion validity and for the three other types of validity will be modified as experience accumulates.

Low Statistical Power

The likelihood of making an incorrect no-difference conclusion (Type II error) increases when sample sizes are small, and α is set low. Moreover, statistical tests differ considerably in power, with some being notably low—for example, tests of the difference between independent correlations (Cronbach and Snow, 1976). Cohen's (1970) book on statistical power gives a preliminary introduction to the topic for social scientists.

Violated Assumptions of Statistical Tests

Most tests of the null hypothesis require that certain assumptions be met if the results of the data analysis are to be meaningfully interpreted. Thus, the particular assumptions of a chosen statistical test have to be known and—where possible—tested in the data on hand. For example, with the analysis of covariance, the regression of the posttest on the first-order covariates should be homogeneous and the groups being compared should be equivalent. Each of these assumptions can be checked to some degree, the former by examining scatterplots and the latter by comparing pretest means. Not all assumptions are equally important. For instance, like other multiple regression techniques, the analysis of variance is robust to violations of normality but is less robust to violations of the assumption of uncorrelated errors (Lindquist, 1953). Since the importance of particular assumptions depends on the test being conducted, it is desirable to consult standard statistical references to check one's understanding about the test's assumptions and to learn how to circumvent any problems with assumptions that might arise with a particular data set.

Fishing and the Error Rate Problem

The likelihood of falsely concluding that covariation exists when it does not (Type I error) increases when multiple comparisons of mean differences are possible and there is no recognition that a certain proportion of the comparisons will be significantly different by chance. Ryan (1959) has distinguished between the error rate per comparison ("the probability that any one of the

comparisons will be incorrectly considered to be significant”), the error rate per experiment (“the expected number of errors per experiment”) and the error rate for experiments in general (“the probability that one or more erroneous conclusions will be drawn in a particular experiment”).

The last two are the most important, and Ryan has illustrated one method of adjusting for the error rate per experiment. This involves computing a new t value which has to be reached before significance at a given α level can be claimed. The new t is obtained by taking the desired α (e.g., .05) and dividing it by the number of possible comparisons so as to give an adjusted proportion (p) that will be lower than .05. Then, the t value corresponding to this adjusted p is looked up in the appropriate tables. It will, of course, be higher than the t values normally associated with $\alpha = .05$. This higher value reflects the stringency required for obtaining a more accurate level of statistical significance when multiple tests are made. A second method for dealing with the error rate problem involves using the conservative multiple comparison tests of Tukey or Scheffé which are discussed in most moderately advanced statistics texts. When there are multiple dependent variables in a factorial experiment, a multivariate analysis of variance strategy can be used for determining whether any of the significant univariate F tests within a particular main or interaction effect are due to chance rather than the manipulations.

The Reliability of Measures

Measures of low reliability (conceptualized either as “stability” or “test-retest”) cannot be depended upon to register true changes. This is because unreliability inflates standard errors of estimates and these standard errors play a crucial role in inferring differences between statistics, such as the means of different treatment groups. Some ways of controlling for unreliability include (1) using longer tests for which items or measures have been carefully selected for their high intercorrelations, or (2) using more aggregated units, e.g., groups instead of individuals, since a group mean will be more stable than individual scores. However, the increase in reliability due to aggregation is counterbalanced by a decrease in the number of degrees of freedom that results. (3) Occasionally the standard corrections for unreliability presented in textbooks can be used, but with great caution—particularly if the reliability estimate is low.

The Reliability of Treatment Implementation

The way a treatment is implemented may differ from one person to another if different persons are responsible for implementing the treatment. There may also be differences from occasion to occasion when the same person implements the treatment. This lack of standardization, both within and between persons, will inflate error variance and decrease the chance of obtaining true differences. The threat—which is pervasive in most kinds of field experiments (see Boruch and Gomez, 1977, for a statistical analysis)—can most obviously be controlled by using all the available opportunities to make the treatment and its implementation as standard as possible across occasions of implementation. In some instances, despite best efforts to standardize, treatments will have con-

siderable unplanned variability in how they are implemented. It is always wise to try and understand and measure this heterogeneity and to use the measures in the data analysis. The advantages and disadvantages of doing this are outlined in chapters 3 and 4.

Random Irrelevancies in the Experimental Setting

Some features of an experimental setting other than the treatment will undoubtedly affect scores on the dependent variable and will inflate error variance. This threat can be most obviously controlled by choosing settings free of extraneous sources of variation or by choosing experimental procedures which force respondents' attention on the treatment and lower the salience of environmental variables. In many complex field settings these suggestions will be very difficult to implement. In such cases the need will be to measure the anticipated sources of extraneous variance which are common to all the treatment groups in *as valid a fashion as possible* in order to introduce the measures into the statistical analysis. Monitoring the experiment in its initial stages will suggest other setting variables which will probably add to the error variance and which, if reliably measured, could be introduced into the analysis to reduce error.

Random Heterogeneity of Respondents

The respondents in any of the treatment groups of an experiment can differ on factors that are correlated with the major dependent variables. Occasionally certain kinds of respondents will be more affected by a treatment than others, and this—as we shall soon see—is a matter of external validity. At other times respondent variables do not interact with the treatment but are related to outcomes. When this happens, the error variance will be inflated. This threat can obviously be controlled by selecting homogeneous respondent populations, but this is often at some cost to external validity. Alternatively, the relevant respondent characteristics can be reliably measured and, under the appropriate circumstances (see Elashoff, 1969), used for blocking or as covariates. It is also worth noting that the threat is reduced when within-subject error terms are appropriate—that is, when they depend not on differences between persons but on differences between occasions of responding within the same persons. The advantages of within-subject designs can be most simply illustrated in pretest-posttest designs where the reduction in error depends on the correlation between each respondent's pre- and post-scores: the higher the correlation, the greater the reduction in the error term.

The Problem of “Accepting” the Null Hypothesis

It is frequently stated in the literature on experimental design that the null hypothesis cannot logically be proven. There are two reasons for this. First, there is always the possibility, however remote, that statistics have failed to detect a true difference. Second, we cannot know what would have resulted if the treatment had been more powerful, or a statistical test of greater power had been used, or if the statistical analysis had extraneous sources of respondent or setting variance which correlate with the dependent variable. (For a discussion

of the logic of accepting the null hypothesis see Cook, Gruder, Hennigan and Flay, 1979.)

These arguments are logically compelling. But while we cannot prove the null hypothesis, in many practical contexts we have to make decisions and act *as though* the null hypothesis were true. This is especially the case in applied research, where decisions have to be based on imperfect knowledge which only suggests that a treatment has had no detectable effect. The issue then becomes: By what standards should one estimate the confidence that can be placed in "accepting" the null hypothesis, particularly if a decision has to be based on the results of a single experiment?

Situation 1

When an explicit directional hypothesis guides the research, it is sometimes possible to conclude with considerable confidence that the derived effect was not obtained under the conditions in which the testing occurred. This conclusion is easiest to draw when the results are statistically significant and in the opposite direction to that specified in the hypothesis or when the results, though not statistically reliable, are contrary to the derived prediction and are found in a vast majority of reasonably powerful subgroup analyses, e.g., at different sites. But note here that the issue is not acceptance of the hypothesis of no-difference, but acceptance of the hypothesis that a particular predicted effect was not obtained.

Yet even when the results are opposite to what was expected, it is still useful to check whether influential suppressor variables might be obscuring a smaller true effect in the predicted direction. For example, Campbell and Boruch (1975) have maintained that true effects of compensatory education programs have been obscured in much past research because the experimental groups receiving compensatory treatments came from economically poorer homes than did the more affluent control group children whose parents' incomes made them ineligible for the compensatory program. Since the children in experimental groups usually come from homes that are associated with lower achievement levels and slower growth rates over time, Campbell and Boruch argue that two countervailing forces are set up in compensatory education experiments. First there are differences in growth rates, which cause experimentals to gain *less* than controls over time. Second, there are true treatment effects which cause experimentals to gain *more* than would randomized controls. The net effect of these two forces can be the misleading appearance of either no treatment effect or even a harmful effect.

Situation 2

A situation where no-difference findings are generally interpretable is when point estimates of the size of a desired effect are available (e.g., we want an increase of more than 2% on one variable or an increase of at least ten points on some other). This specification, when combined with variance estimates, allows us to compute—before any data are collected—the sample size required for inferring at a given confidence level that a difference of the desired size would have been observed if it had truly occurred. Thus, if the number of observations in a completed study is at least equal to the number specified in the prestudy analysis, and if the variances used for computing the desired sample size are similar to

those that were actually obtained in the study, then one can compute with a known level of confidence whether a specified point standard has been exceeded with the data on hand. This is clearly a desirable situation for any data analyst.

Let us illustrate the above points by describing a section from a report on the effects of manpower training programs on subsequent earnings. Aschenfelter (1974) knew that training costs were about \$1,800 for each trainee. He estimated that a return of at least 10% on this investment (i.e., \$200) would be adequate for declaring the manpower training program a "success." Then, assuming equal numbers of persons in the experimental training group and the no-training control group, and knowing from previous data that the standard deviation in income was about \$2,000 for white males, Aschenfelter calculated that about 1,600 persons would be needed in the experiment if a true effect of at least \$200 was to be detected with 95% certainty. However, Aschenfelter further calculated that if he were to break down the data by two race and two sex groups, he would need a total of 6,400 respondents—1,600 in each subgroup. Knowing this, he was then in a position to assess whether he had the necessary resources to design an experiment of this size or whether he would be better served by using some other technique for trying to evaluate the training program.

Unfortunately, it is rare to have valid variance estimates and a prior point estimate of the size of an expected effect. The problem of specifying expected effect sizes is sometimes political, largely because a publicized point estimate can become a reference against which a social innovation is evaluated. As a result, even if an innovation has had *some* ameliorative effect, it may not be given much credit if it failed to have the *promised* effect. It is no small wonder, therefore, that the managers of programs prefer less specific statements such as "We want to increase achievement," or "We want to reduce inflation" to statements such as "We want to increase achievement by two years for every year of teaching" or "We want to reduce inflation to 5% a year." The problem of specifying magnitudes is also sometimes one of "consciousness," for the issue may simply not be considered in designing the research. Alternatively, it may be silently considered by some persons but never brought to the level of discussion for fear that different parties to the research may disagree on the level of effect required to conclude that a treatment has made a significant, practical difference.

Situation 3

Even when no magnitude-of-effect estimate is available, it is still possible to use information about sample sizes and variances in order to calculate *retrospectively* the size of any effect that could have been detected in a particular experiment with, say, 95% confidence. This magnitude can then be inspected and interpreted. At times it will seem so unreasonably large that the only responsible conclusion is that the experiment was not powerful enough to have detected a true effect. For instance, in the Aschenfelter case a sample size of 400, split equally between experimentals and controls, would have required the experimentals to earn considerably more than \$200 on the average if a true effect were to be detected at the 5% level. How reasonable is it to expect an average increase in earnings over \$200 in the first working year after graduating from a job-training program? The answer to this cannot be definitive since no criteria exist for assess-

ing reasonableness. Nonetheless, the figure seems to us to be very high. We would strongly advise anyone whose research results in a no-difference conclusion to conduct the retrospective analysis indicated above.

Situation 4

When data are first analyzed, it is often the case that the estimate of the treatment effect (say, a difference between sample means) is statistically nonsignificant but in the expected direction. Typically, efforts are then made to "reduce the error term" used for testing the treatment effect—a topic that we shall now discuss.

Obviously, it is desirable to design the research initially so as to minimize this error. For instance, "Student" (1931) reexamined an experiment which compared how four months of free pasteurized milk affected the height and weight of Scottish school children when compared to four months of raw milk. About 5,000 students received each type of milk. "Student" maintained that the same statistical power (and much lower financial costs) would have resulted had only 50 sets of identical twins been used. This is because weight and height are highly correlated for monozygotic twins, leading to lower error terms than those associated with differences between nonrelated children. In light of modern knowledge, we might not want to design the study in the way "Student" suggested because of nonstatistical considerations. For example, would parents seek to supplement one of their twin's diets if they knew that the other was receiving a school-provided supplement, and how generalizable would findings from 50 sets of twins be? Nonetheless, "Student's" point is important, and it suggests designing research *wherever possible* so as to have small error terms, provided that the means of reducing the error do not trivialize the research.

Perhaps the best way of reducing the error due to differences between persons is to match *before* random assignment to treatments. (This, we shall soon see, is quite different from matching as a substitute for randomization. While matching prior to randomization can increase statistical conclusion validity and permit tests to discover in which particular subgroups a treatment effect is obtained, matching as an alternative to randomization often leads to statistical regression artifacts that can masquerade as treatment effects.) The best matching variables are those that are most highly correlated with posttest scores. Normally the pretest is the best single matching variable since it is a proxy for all the social and biological forces that make some individuals or aggregates of individuals different from others. The actual process of matching is simple. One takes all the scores on the matching variable, ranks them, and places them into blocks whose size corresponds to the number of experimental groups (say, three). Then, the three persons in the first block are randomly assigned to one of the experimental groups, the next three in the next block are randomly assigned, and so on until all the cases are assigned. The data that result from such a design can then be analyzed as coming from a Levels \times Treatment design. The same logic basically holds when matching takes place on multiple variables, but the problem of finding matches is harder. (Matching will be discussed in greater detail in several of the chapters to come.)

Given random assignment to treatment conditions, it is nonetheless possible to match retrospectively after all the data have been collected. With large sample

sizes, retrospective matching will result in treatment groups that have comparable proportions of units with the characteristic on which matching takes place. However, the major disadvantage of this technique compared to prospective matching is that subgroups with few members (e.g., blacks in many settings) can be disproportionately represented in each treatment group, with very few persons in one of the groups. This makes it difficult to estimate treatment effects for the subgroups in question. But this problem aside, *ex post facto* blocking can be extremely useful both because effects of the blocking variable can be removed from the error term and because the interaction of the blocking variable with the treatment can be assessed.

When there is no interest in testing how the dependent variables are related to the matching or blocking variable, an alternative method of reducing the error term can be used that loses fewer degrees of freedom. It requires using multiple regression strategies involving variables which are correlated with the dependent variable *within treatment groups* and whose effects are to be partialled out of the dependent variable. Covariance analysis is one such strategy. The extent to which covariance analysis reduces error depends on the correlation between the covariates (the lower the better) and the correlation of each of them with the dependent variable (the higher the better). But two words of caution are required about such multiple regression adjustments. First, important statistical assumptions have to be demonstrably met for the results of the analysis to be meaningful, especially the assumption of homogeneous regression within treatment groups. Second, in experiments with non-comparable groups, the analysis will reduce error but will rarely adjust away all the initial differences between groups. Thus, the function of reducing error—which makes covariance so useful with both randomized experiments and quasi-experiments—should not be confused with the purported function of making groups equivalent. Equivalence is not needed with randomized experiments and is rarely achieved by regression adjustments with quasi-experiments. (For an extended discussion of these last points see chapters 3 and 4.)

Both matching and multiple regression adjustments assume that measures have been made of the variables for which adjustments are to be made. Failure to measure them means that error terms cannot be reduced to reflect the way that person or setting variables are related to the major outcome measures of an experiment. Increasing one's confidence in accepting the null hypothesis demands *valid* measurement of the variables that are most likely to affect posttest scores.

There is little point in reducing the error variance due to differences among persons and settings if the outcome measures are so unreliable that they cannot register true change. Thus, the experimenter has to be certain to begin with reliable measures. Alternatively, the experimenter has to try and develop even more reliable measures after an experiment is under way by adding items to tests, by rescaling, or by aggregating data. But whether or not attempts are undertaken to increase reliability, it is important that internal consistency estimates and test-retest correlations be displayed in a research report. The reader can at least judge for himself the extent to which measures may have been capable of registering true changes.

Statistical procedures exist for correcting for unreliable measurement. This means that analysis of "true" scores should be possible. (Details of this procedure

are available in many standard texts, including McNemar, 1975.) These correctional procedures can often be misleading in practice. First, there are many ways of conceptualizing reliability, each of which implies a different reliability measure and different numerical estimates of the amount of reliability. Second, for any one kind of reliability, its own reliability will not be directly known. And third, reliability-adjusted values do not logically correspond with the values that would have been obtained had there been perfectly reliable measurement. This is perhaps most dramatically illustrated by reliability-adjusted correlations in excess of 1.00, or by the fact that a nonsignificant r of, say, $-.10$ must inevitably result in a higher adjusted value *of the same sign*, whereas the population correlation may have been $+.04$. Great caution must be exercised, therefore, in the use of reliability adjustments. It would be naive to present the results only for adjusted data or, when adjusted results are presented, to use only one estimate of reliability. A range would be better.

Each of the foregoing strategies can reduce error terms. Consequently, it is advisable for purposes of statistical conclusion validity to use as many as possible of the following design features. (1) Each person might be his own control (i.e., serve in more than one experimental group); (2) samples might be selected that are as homogeneous as possible (monozygotic twins are merely the extreme of this); (3) pretest measures should be collected on the same scales that are used for measuring effect; (4) matching might take place, before or after randomization, on variables that are correlated with the posttest; (5) the effects of other variables that are correlated with the posttest might be covaried out; (6) the reliability of dependent variable measures might be increased; or, (7) occasionally the raw scores might be adjusted for unreliability. In addition, (8) estimates of the desired magnitude of effect should be elicited, where possible, before the research begins. Even when no such estimate can be determined, (9) the absolute magnitude of a treatment effect should be presented so that readers can infer for themselves whether a statistically reliable effect is so small as to be practically insignificant or whether a nonreliable effect is so large as to merit further research with more powerful statistical analyses. It should not be forgotten that all of these strategies have negative consequences if uncritically used and that all of them require trade-offs that will become more obvious later. Moreover, most of them are more problematic when analyzing data from quasi-experiments than data from randomized experiments.

Situation 5

Having tried to make the error term as small as possible, the researcher will encounter a problem if the data analysis still fails to result in statistically significant effects. All one can then conclude is that this particular example of this particular treatment contrast had no observable effects. One cannot easily draw conclusions about what would have happened if each treatment had been more homogeneously implemented (i.e., each person or unit in a group had received exactly the same amounts of the treatment) or if the experimental contrast had been larger (i.e., the mean difference between groups had been greater on some measure designed to assess the strength of the treatment implementation).

As we shall see later, quasi-experimental analyses can sometimes be conducted to assess these two possibilities by capitalizing upon the fact that measures

of treatment implementation can be made which estimate presumed differences in the strength of the treatment. Such differences can then be associated with estimates of the magnitude of changes between a pretest and posttest in order to determine if the two are related. While such analyses should definitely be conducted, chapters 3 and 4 will illustrate that great care must be exercised in interpreting the results. This is because individuals will normally have voluntarily chosen to expose themselves to treatments in different amounts, and so the kind of person at one treatment level is likely to be different from a person at another level. Nonetheless, if sophisticated quasi-experimental analyses of the kind in chapters 3 and 4 still fail to result in covariation between the treatment and outcome measures, then the analyst can be all the more confident in accepting the null hypothesis.

INTERNAL VALIDITY

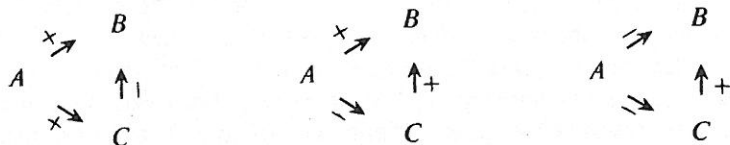
Introduction

Once it has been established that two variables covary, the problem is to decide whether there is any causal relationship between the two and, if there is, to decide whether the direction of causality is from the measured or manipulated A to the measured B , or vice versa.

The task of ascertaining the direction of causality usually depends on knowledge of a time sequence. Such knowledge is usually available for experiments, as opposed to most passive observational studies. In a randomized experiment, the researcher knows that the measurement of possible outcomes takes place after the treatment has been manipulated. In quasi-experiments, most of which require both pretest and posttest measurement, the researcher can relate some measure of pretest-posttest change to differences in treatments.

It is more difficult to assess the possibility that A and B may be related only through some third variable (C). If they were, the causal relationship would have to be described as: $A \rightarrow C \rightarrow B$. This is quite different from the model $A \rightarrow B$ which most clearly implies that A causes B . To conclude that A causes B when in fact the model $A \rightarrow C \rightarrow B$ is true would be to draw a false positive conclusion about cause. Accounting for third-variable alternative interpretations of presumed A - B relationships is the essence of internal validity and is the major focus of this book.

Although in the examples that follow we shall deal primarily with the possibility of false positive findings, it should not be forgotten that third variables can also threaten internal validity by leading to false negatives. The latter occur whenever relationship signs are as below. In the case to the left, an increase in A causes an increase in both B and C , but the increase in C causes a decrease in B . Thus,



the net effect of A and C on B would be to tend to obscure a true $A \rightarrow B$ relationship. In the case depicted in the center, an increase in A would cause an increase in B and a decrease in C , while a decrease in C would cause a decrease in B . Once again, the effects of A and C would tend to cancel each other out. In the final case, an increase in A would cause a decrease in both B and C , and the decrease in C would cause a countervailing increase in B .

Let us give an example of the second of these three relationships. Imagine that A is tutoring and B is academic achievement. Imagine, further, that tutoring is given to the weaker students academically and is withheld from the stronger, this process of selection into the treatment being C . Since tutoring is negatively related to initial achievement, we have $A \Rightarrow C$. Being weaker, the students with tutoring would be expected to gain less over time than their fellow students for a number of reasons that have nothing to do with tutoring (e.g., slower rates of learning from other sources). Hence, $C \Rightarrow B$. Thus, if tutoring did raise achievement ($A \rightarrow B$) but the children who received tutoring were expected to gain less from schooling anyway (that is, $C \Rightarrow B$), then the effects of tutoring and of lower expected growth rates would countervail. In the special case where the two forces were of equal magnitude, they would totally cancel each other out. In cases where one force was stronger than the other, the stronger cause would prevail but its effect would be weakened by the countervailing cause. We hope that our later examples, which emphasize internal validity threats and false positive findings, will not blind readers to the effects that such threats can have in leading to false negative findings because of the operation of suppressor variables.

It is possible for more than one internal validity threat to operate in a given situation. The net bias that the threats cause depends on whether they are similar or different in the direction of bias and on the magnitude of any bias they cause independently. Clearly, false causal inferences are more likely the more numerous and powerful the validity threats and the more homogeneous the direction of their effects. Though our discussion will be largely in terms of threats *taken singly*, this should not blind readers to the possibility that multiple internal validity threats can operate in cumulative or countervailing fashion in a single study.

Threats to Internal Validity

Bearing this brief introduction in mind, we want to define some specific threats to internal validity.

History

"History" is a threat when an observed effect might be due to an event which takes place between the pretest and the posttest, when this event is not the treatment of research interest. In much laboratory research the threat is controlled by *insulating* respondents from outside influences (e.g., in a quiet laboratory) or by *choosing dependent variables* that could not plausibly have been affected by outside forces (e.g., the learning of nonsense syllables). Unfortunately, these techniques are rarely available to the field researcher.

Maturation

This is a threat when an observed effect might be due to the respondent's growing older, wiser, stronger, more experienced, and the like between pretest and posttest and when this maturation is not the treatment of research interest.

Testing

This is a threat when an effect might be due to the number of times particular responses are measured. In particular, familiarity with a test can sometimes enhance performance because items and error responses are more likely to be remembered at later testing sessions.

Instrumentation

This is a threat when an effect might be due to a change in the measuring instrument between pretest and posttest and not to the treatment's differential impact at each time interval. Thus, instrumentation is involved when human observers become more experienced between a pretest and posttest or when a test shifts in metric at different points. The latter can happen, for instance, if intervals are narrower at the ends of a scale than at the midpoint, resulting in so-called ceiling or basement effects. (Basement effects are also called "floor" effects.)

Statistical Regression

This is a threat when an effect might be due to respondents' being classified into experimental groups at, say, the pretest on the basis of pretest scores or correlates of pretest scores. When this happens and measures are unreliable, high pretest scorers will score relatively lower at the posttest and low pretest scorers will score higher. It would be wrong to attribute such differential "change" to a treatment because it might be due to statistical regression.

Statistical regression is not an easy concept to grasp intuitively. It might help you understand it if you think of your own academic test taking. You may sometimes have surprised yourself by doing worse than you expected, perhaps because you didn't sleep well the night before, you read the questions too quickly and misunderstood them, there may have been someone with an infuriating cough in front of you, or because the test just happened to have had a disproportionately high number of items from a part of the curriculum that you had not studied in detail. Any or all of these factors could have depressed your scores, and they can be conceptualized as error factors that do not reflect true ability. Consequently, the next time you took a test on the same or similar subject matter your scores would probably be higher and would more accurately reflect your ability. This is because, all things being equal, you will be less likely to have been deprived of sleep, less likely to have read the questions too quickly, less likely to have had someone with a cough sit in front of you, and less likely to have received questions from parts of the curriculum that you had studied the least.

Viewed more generally, statistical regression (1) operates to increase obtained pretest-posttest gain scores among low pretest scores, since this group's pretest scores are more likely to have been depressed by error; (2) operates to decrease obtained change scores among persons with high pretest scores since their pretest scores are likely to have been inflated by error; and (3) does not affect obtained

change scores among scorers at the center of the pretest distribution since the group is likely to contain as many units whose pretest scores are inflated by error as units whose pretest scores are deflated by it. Regression is always to the population mean of a group. Its magnitude depends both on the test-retest reliability of a measure and on the difference between the mean of a deliberately selected subgroup and the mean of the population from which the subgroup was chosen. The higher the reliability and the smaller the difference, the less will be the regression.

Selection

This is a threat when an effect may be due to the difference between the kinds of people in one experimental group as opposed to another. Selection is therefore pervasive in quasi-experimental research, which is defined in terms of different groups receiving different treatments as opposed to probabilistically equivalent groups receiving treatments as in the randomized experiment.

Mortality

This is a threat when an effect may be due to the different kinds of persons who dropped out of a particular treatment group during the course of an experiment. This results in a selection artifact, since the experimental groups are then composed of different kinds of persons at the posttest.

Interactions with Selection

Many of the foregoing threats to internal validity can interact with selection to produce forces that might spuriously appear as treatment effects. Among these are selection-maturation, selection-history, and selection-instrumentation. Selection-maturation results when experimental groups are maturing at different speeds. Such group differences in growth rates typically occur, for example, when middle-class and lower-class children are compared at two different time intervals on some test of cognitive knowledge. In this situation, the children from more affluent backgrounds tend to gain at a faster rate than the others. Selection-history (or local history) results from the various treatment groups coming from different settings so that each group could experience a unique local history that might affect outcome variables. (The interaction can also occur with randomized experiments if a treatment is only implemented at one or two sessions—usually with large groups of respondents. In such cases, the treatment will be associated with any unique events that happened during the few sessions which provided all the data about a particular treatment's effects.) Selection-instrumentation occurs when different groups score at different mean positions on a test whose intervals are not equal. The best known examples of this occur when there are differential "ceiling" and "floor" effects, the former being when an instrument cannot register any more true gain in one of the groups, and the latter when more scores from one group than another are clustered at the lower end of a scale.

Ambiguity About the Direction of Causal Influence

It is possible to imagine a situation in which all plausible third-variable explanations of an *A-B* relationship have been ruled out and where it is not clear

whether *A* causes *B* or *B* causes *A*. This is an especially salient threat to internal validity in simple correlational studies where it will often not be clear whether, for example, less foreman supervision causes higher productivity or whether higher productivity causes less supervision. This particular threat is not salient in most experiments since the order of the temporal precedence is clear. Nor is it a problem in those correlational studies where one direction of causal influence is relatively implausible (e.g., it is more plausible to infer that a decrease in the environmental temperature causes an increase in fuel consumption than it is to infer that an increase in fuel consumption causes a decrease in outside temperature). Nor is it necessarily a problem in correlational studies when the data are collected at more than one time interval, for then one knows something about temporal antecedence. However, ambiguity about the direction of causal influence is a problem in many correlational studies that are cross-sectional.

Diffusion or Imitation of Treatments

When treatments involve informational programs and when the various experimental (and control) groups can communicate with each other, respondents in one treatment group may learn the information intended for others. The experiment may, therefore, become invalid because there are no planned differences between experimental and control groups. This problem may be particularly acute in quasi-experiments where the desired similarity of experimental units may be accompanied by a physical closeness that permits the groups to communicate. For example, if one of the New England states were used as a control group to study the effects of changes in the New York abortion law, any true effects of the law would be obscured if New Englanders went freely to New York for abortions.

Compensatory Equalization of Treatments

When the experimental treatment provides goods or services generally believed to be desirable, there may emerge administrative and constituency reluctance to tolerate the focused inequality that results. Thus, in nationwide educational experiments such as Follow Through, the control schools, particularly if equally needy, tended to be given Title I funds earmarked for disadvantaged children. Since these funds were given to the supposed "no-treatment controls" in amounts approximately equivalent to those coming to the experimental schools, the planned contrast obviously broke down. Several other experimental evaluations of compensatory education have encountered the same problem. It exemplifies a problem of administrative equity that must certainly occur elsewhere, including among units of industrial organizations. Such focused inequities may explain some administrators' reluctance to employ random assignment to treatments which their constituencies consider valuable.

Compensatory Rivalry by Respondents Receiving Less Desirable Treatments

Where the assignment of persons or organizational units to experimental and control conditions is made public (as it frequently must be), conditions of social competition may be generated. The control group, as the natural underdog, may be motivated to reduce or reverse the expected difference. This result is particularly likely where intact units (such as departments, plants, work crews, and the

like) are assigned to treatments, or if members of the control group will be at a disadvantage if a treatment is successful. As an example, Saretsky (1972) has pointed out that the success of performance contracting—paying commercial contractors according to the size of learning gains made by students—would threaten the job security of school teachers. Many would think that they could be replaced by contractors or that their professional role might be redefined because of the contractors. Given this threat, Saretsky has suggested that the academic performance of children taught by teachers in the control groups of the OEO Performance Contracting Experiment could have been better during the experiment than it had been in past years. The net effect of atypically high learning gains by controls, if it occurred, would be to diminish the difference in learning between control students taught by their regular teachers and experimental children taught by outside contractors. Saretsky (1972) describes other examples of the phenomenon of special effort by the controls. He calls this effort a "John Henry effect" in honor of the steel driver who, when he knew his output was to be compared to that of a steam drill, worked so hard that he outperformed the drill and died of overexertion. (Compensatory rivalry is very like compensatory equalization in that each is a response to the focused inequity that inevitably results when treatments differ in desirability, as they do if valuable resources are being distributed. However, compensatory equalization is a response of administrators and compensatory rivalry is a response of those in the less desirable treatment groups.)

Resentful Demoralization of Respondents Receiving Less Desirable Treatments

When an experiment is obtrusive, the reaction of a no-treatment control group or groups receiving less desirable treatments can be associated with resentment and demoralization, as well as with compensatory rivalry. This is because persons in the less desirable treatment groups are often relatively deprived when compared to others. In an industrial setting the persons experiencing the less desirable treatments might retaliate by lowering productivity and company profits, while in an educational setting, teachers or students could "lose heart" or become angry and "act up." Any of these forces could lead to a posttest difference between treatment and no-treatment groups, and it would be quite wrong to attribute the difference to the planned treatment. Cause would not be from the planned cause, *A*, given to a treatment group. Rather, it would be from the inadvertent resentful demoralization experienced by the no-treatment controls.

Estimating Internal Validity in Randomized Experiments and Quasi-Experiments

Estimating the internal validity of a relationship is a deductive process in which the investigator has to systematically think through how each of the internal validity threats may have influenced the data. Then, the investigator has to examine the data to test which relevant threats can be ruled out. In all of this process, the researcher has to be his or her own best critic, trenchantly examining all of the threats he or she can imagine. When all of the threats can plausibly be eliminated, it is possible to make confident conclusions about whether a relationship is probably causal. When all of them cannot, perhaps because the appropriate data are not available or because the data indicate that a particular threat may indeed have

operated, then the investigator has to conclude that a demonstrated relationship between two variables may or may not be causal. Sometimes the alternative interpretations may seem implausible enough to be ignored and the investigator will be inclined to dismiss them. They can be dismissed with a special degree of confidence when the alternative interpretations seem unlikely on the basis of findings from a research tradition with a large number of relevant and replicated findings.

Invoking plausibility has its pitfalls, since it may often be difficult to obtain high inter-judge agreement about the plausibility of a particular alternative interpretation. Moreover, theory testers place great emphasis on testing theoretical predictions that seem so implausible that neither common sense nor other theories would make the same prediction. There is in this an implied confession that the "implausible" is sometimes true. Thus, "implausible" alternative interpretations should reduce, but not eliminate, our doubt about whether relationships are causal.

When respondents are randomly assigned to treatment groups, each group is similarly constituted on the average (no selection, maturation, or selection-maturation problems). Each experiences the same testing conditions and research instruments (no testing or instrumentation problems). No deliberate selection is made of high and low scorers on any tests except under conditions where respondents are first matched according to, say, pretest scores and are then randomly assigned to treatment conditions (no statistical regression problem). Each group experiences the same global pattern of history (no history problem). And if there are treatment-related differences in who drops out of the experiment, this is interpretable as a consequence of the treatment. Thus, randomization takes care of many threats to internal validity.

With quasi-experimental groups, the situation is quite different. Instead of relying on randomization to rule out most internal validity threats, the investigator has to make all the threats explicit and then rule them out one by one. His task is, therefore, more laborious. It is also less enviable since his final causal inference will not be as strong as if he had conducted a randomized experiment. The principle reason for choosing to conduct randomized experiments over other types of research design is that they make causal inference easier.

Threats to Internal Validity That Randomization Does Not Rule Out

Though randomization conveniently rules out many threats to internal validity, it does not rule out all of them. In particular, imitation of treatments, compensatory equalization, compensatory rivalry, and demoralization in groups receiving less desirable treatments can each threaten internal validity even when randomization has been successfully implemented and maintained over time. Some of these threats will usually cause spurious differences (e.g., demoralization in the controls). However, other threats will tend to obscure true differences, especially by making no-treatment control groups perform atypically. This last happens with the imitation of treatments, compensatory equalization, and compensatory rivalry. We want to make clear that, while randomized experiments are superior to quasi-experiments with respect to internal validity, they are not perfect.

Most of the threats that randomization does not rule out result from the focused inequities that inevitably accompany experimentation because some peo-

ple receive one treatment and others receive different treatments or no treatment at all. Since much social experimentation is ameliorative, treatments have to differ in desirability by virtue of the very research problem (e.g., the different payment levels in a compensatory education or an income supplement program, or the different amounts of time that can be spent away from cell-block confinement in a prison experiment on "rehabilitation"). Obviously, individual respondents want to receive the more desirable treatments. In the same vein, officials want to avoid salient inequities which can lead to charges that they favored some respondents over others in distributing treatments.

It is rare in our society to have valuable resources distributed on a random basis. *Instead, we expect them to be distributed according to need, merit, seniority or on a "first come, first served" basis.* The point is that distributing resources by lottery violates the meritocratic and/or social responsibility norms which regulate and justify most differences in rewards and opportunities in the United States. This is not to say that lotteries are never used in resource distribution. They seem to be convenient, for instance, in distributing sudden "windfalls" or universally undesired resources (e.g., a lottery was used for inducting young men into the U.S. armed services after 1969). Nonetheless, distribution by merit or need is more common than distribution by chance, and the latter often violates expectations about what is "just." It is this which leads to randomization exacerbating some internal validity threats.

The extent of an administrator's apprehension about randomization probably depends on four subjective estimates: (1) the differences in desirability between treatments; (2) the probability that individuals will learn of treatment differences; (3) the probability that organized constituencies will learn of these differences; and (4) how much the various constituencies will feel that their interests are affected by the most likely research outcomes. Some research questions make it difficult to rule out all of an administrator's apprehensions since, first, they absolutely require treatments that differ in desirability (e.g., what is the effect of extra payments to schools?). Second, scarce research resources require geographical contiguity (e.g., we can only do the study in one school district). Third, it seems to be part of an administrator's job to consider how various constituencies might react to focused inequities and to fear the worst (e.g., what will the teachers' union or the PTA think if resources are distributed by chance instead of by need or merit?). And fourth, administrators know that constituency representatives want to get the best possible advantages for their charges and want to avoid any potential harm to them (e.g., a teachers' union official might think: If performance contracting works in schools, then the role of the classroom teacher could be reduced in scope and importance—do we want that?). Such considerations highlight both the difficulties of gaining permission to randomize and of ruling out the threat of compensatory equalization when randomization has taken place.

The only other internal validity threat that can operate in a randomized experiment is differential mortality from the treatment groups. While such differences can be interpreted as a consequence of the treatment—and as a result will often be very important—they have the undesirable side effect of obscuring the interpretation of other results. This is because the units remaining in one treatment group may not be comparable on the average to those in another group. Thus, if there

were differential attrition from, say, an experiment on the effect of income supplements on the motivation to find work, we would not be sure if a relationship between the dollar value of a supplement and the number of days worked was due to the supplement reducing the number of days worked or to selection differences associated with the kinds of persons who remained in each treatment condition for the entire experiment. Treatment-correlated attrition leads to the possibility of a selection confound. We might readily surmise that such attrition is all the more likely the more the treatments differ in desirability.

With the exception of differential mortality and the selection problems that follow from it, the threats to internal validity which random assignment does not rule out are caused by atypical behavior on the part of persons in no-treatment control groups or groups that receive less desirable treatments. Such behavior represents an unplanned but nonetheless causal consequence of the planned experimental contrast. Even when there is a valid causal relationship at the operational level, one may wonder how differences in *B* can be interpreted as the result of threats to internal validity. Internal validity is, after all, concerned with threats that cast doubt on whether there is a valid *causal* connection, and the threats we are discussing do not deny the validity of a causal connection. The answer is in one sense simple. Internal validity refers to doubt about whether there is a causal connection from *A*-as-manipulated (or measured) to *B*-as-measured; on the other hand, the threats to internal validity which we are discussing (e.g., resentful demoralization of the controls) cast doubt on whether the causal connection is from *A* to *B* or is from *A*'s comparison group to *B*. (In another sense, this issue is academic, for causal inference always depends on the *contrast* between *A* and *A*'s comparison, irrespective of whether *A* or the comparison causes the observed changes in the dependent variable. Given our emphasis on the desirability of identifying *active* causal agents, it is important to identify whether *A* or its comparison accounts for change, since knowing the active causal agent allows one to know what to manipulate. This is why we specify internal validity in terms of the pattern of influence from *A* to *B* rather than in terms of the pattern of influence from the contrast between *A* and its comparison to *B*).

Assessing the Plausibility of Internal Validity Threats If a Randomized Experiment Has Been Implemented

The possibility of a selection artifact resulting from differential attrition can best be empirically assessed in two ways. First, an analysis is called for of the proportion of respondents, originally assigned to each experimental condition, who actually provide posttest data. Differences in this proportion across treatments indicate a differential dropout. Second, an analysis is called for of the pretest scores in each treatment group computed on the basis of all those who provided posttest data. This gives a preliminary indication of whether the dropouts differed across groups on the background characteristics that are most likely to affect posttest scores (i.e., those that are highly correlated with pretest scores on the same test). We will deal with these points in greater detail in chapter 8.

An assessment of imitation, compensatory equalization, or compensatory rivalry can often be made by direct measures in the experimental and control groups of the process that the independent variable was meant to affect. Thus, if a

treatment were meant to provide money to some schools but not others, the finances of both kinds of schools would need examining. If a treatment was expected to make experimental children view an education television program, it would be necessary to measure how often they watch the show and how often the controls watch it. A small or nonexistent experimental contrast would suggest that imitation, compensatory equalization, or compensatory rivalry may have occurred. Thus, measures of the exact nature of the treatment in *all* treatment and control groups are absolutely vital in any experiment. The sooner such measurements are taken, the easier it will be to detect unexpected patterns of behavior in the experiment and control groups and the easier it will be to take corrective action.

It will normally be easy to use background information to find out if controls had contact with experimentals and copied them or to find out if administrators provided additional resources to some units from nonexperimental sources. It will normally not be as easy to assess whether compensatory rivalry took place, though direct measures of verbal expressions of such rivalry by the controls can give a lead, as can indications of whether control group performance is greater than would be expected. Saretsky (1972), it will be remembered, tried to determine this in the performance level in past years in the same classes, but he probably ran into a regression problem. Nonetheless, if used with care, the use of secondary data from past classes can be useful for attempting to assess the magnitude of any compensatory rivalry. Such data could also be useful for assessing resentful demoralization, because this threat leads to the testable prediction that performance should be atypically low in the control group during the experiment.

CONSTRUCT VALIDITY OF PUTATIVE CAUSES AND EFFECTS

Introduction

Construct validity is what experimental psychologists are concerned with when they worry about "confounding." This refers to the possibility that the operations which are meant to represent a particular cause or effect construct can be construed in terms of more than one construct, each of which is stated at the same level of reduction. Confounding means that what one investigator interprets as a causal relationship between theoretical constructs labeled *A* and *B*, another investigator might interpret as a causal relationship between constructs *A* and *Y* or between *X* and *B* or even between *X* and *Y*.

In the discussion that follows we shall restrict ourselves to the construct validity of presumed causes and effects, since these play an especially crucial role in experiments whose *raison d'être* is to test causal propositions. But it should be clearly noted that construct validity concerns are not limited to cause and effect constructs. All aspects of the research require naming samples in generalizable terms, including samples of people and settings as well as samples of measures or manipulations. Even with internal validity and statistical conclusion validity, inferences have to be made about abstract constructs: viz "cause" and "reliable change" or "reliable differences."

The reference to the level of reduction in the definition of "confounding" is important, because it is always possible to "translate" sociological terms

into psychological ones, or psychological terms into biological ones. For example, participative decision making could become conformity to membership group norms on one level, or some correlate of, say, the ascending reticular activating system on another. Each of these levels of reduction is useful in different ways and none is more legitimate than any other. But such "translations" from one level to another do not involve the confounding of rival explanations that is at issue here.

Before we continue our abstract characterization of construct validity, some concrete examples of well-known construct validity concerns may help. In earlier medical experiments on drugs, the psychotherapeutic effect of the doctor's helpful concern was confounded with the chemical action of the pill. So, too, were the doctor's and the patient's belief that the pill should have helped. To circumvent these problems and to increase confidence that any observed effects could be attributed to the chemical action of the pill *alone*, the placebo control group and the double-blind experimental design were introduced. (The first of these involves giving a chemically inert substance to respondents, and the second requires that neither the person prescribing the pill nor the person evaluating its effects knows the experimental condition to which the patient has been assigned.)

In industrial relations research, the Hawthorne effect is another confound which causes uncertainty about how operations should be labeled. If we assume for the moment that productivity was increased in the original Hawthorne studies by the planned experimental intervention, the issue for construct validity purposes is: Was the increase due to shifts in illumination (the planned treatment) or to the demonstrated administrative concern over improved working conditions (the "Hawthorne effect") or to telling the women how well they were doing their work (an inadvertent correlate of increasing the illumination)?

Construct validity concerns begin to surface at the planning and pilot-testing stages of an experiment when attempts are made to fit the anticipated cause and effect operations to their referent constructs, whether these are derived from formal social science theory or from policy considerations. Such "fitting" to the construct of interest is best achieved (1) by the careful preexperimental explication of constructs so that definitions are *clear* and in conformity with public understanding of the words being used, and (2) by data analyses directed at some of the four following points, preferably all of them.

First, a test should be made of the extent to which the independent variables alter what they are meant to alter. This is done by assessing whether the treatment manipulation is related to direct measures of the process designed to be affected by the treatment. (This is called "assessing the 'take' of the independent variable.") Second, a test should be conducted to assess whether an independent variable does not vary with measures of related but different constructs. For instance, a manipulation of "communicator expertise" should be correlated with reports from respondents about the communicator's level of knowledge, but it should not be correlated with attributions about cognate constructs, such as trustworthiness, congeniality, or power. If there are such correlations, it is difficult to differentiate effects due to expertise from those due to the other variables. Third, the proposed dependent variables should tap into the

factors they are meant to measure. Normally, some form of inter-item correlation can suggest this. And fourth, the dependent variables should not be dominated by irrelevant factors that make them measures of more or less than was intended. Thus, the outcome construct, like the treatment construct, has to be differentiated from its particular cognates.

As we have detailed the procedure, assessing construct validity depends on two processes: first, testing for a *convergence* across *different* measures or manipulations of the same "thing" and, second, testing for a *divergence* between measures and manipulations of related but conceptually distinct "things." Our position should not be interpreted to imply that construct validity absolutely depends on having information about both convergences and divergences, for it is clearly desirable to have information about convergences even when nothing is known directly about divergences. Indeed, other discussions of construct validity have restricted themselves to convergences, even while noting that a close correspondence between different types of measures of the same thing is less meaningful if there are similar measurement irrelevancies associated with each measure, as when only paper-and-pencil or observational measures of the same construct are made—see Campbell and Tyler, 1957; Cronbach and Meehl, 1955; Cronbach, Glesser, Nanda, and Rajaratnam, 1972. However, as Campbell and Fiske (1959) suggest, a construct should be differentiated from related theoretical constructs as well as from methodological irrelevancies. (For an example of differentiation from other theoretical constructs in basic research, see Cook, Crosby and Hennigan, 1977; and for an example in applied research, see the differentiation of viewing "Sesame Street" from "being encouraged to view 'Sesame Street' by paid professionals," Cook et al., 1975.)

We can illustrate these points by considering a possible experiment on the effects of supervisory distance. Suppose we operationalized "supervision" as a foreman standing within comfortable speaking distance of workers (e.g., ten feet). This particular operationalization would exclude distances that were beyond speaking but not beyond seeing, and the treatment might be more exactly characterized as "supervision from speaking distances." It would be dangerous to generalize from such a specific treatment to the general "supervision" construct, especially if supervision has different consequences when it comes from shorter and longer distances. To lessen this possibility, it would be useful if supervisory distance were systematically varied by means of planned manipulations. That is not always possible. However, it would still be useful if supervision *inadvertently* varied across a wide range of distances because foremen differed in their behavior from day to day. Careful analysis of the effects of spontaneous variation in distance would then allow us to test whether we can generalize from one supervisory distance to another. If we can, we can generalize with greater confidence to the general construct of "supervision," whereas if we cannot, we would like to restrict our generalization to "supervision from ten feet or less."

The foremen might also differ from each other, or might themselves differ from day to day, in whether they supervise with a smile or in an officious manner. Neither the smile nor the officiousness would seem to be necessary components of most definitions of "supervision." Hence, the researcher might

hope that such irrelevant behaviors would occur with different frequency across instances where supervisory distance was manipulated, and that data analyses could be conducted to differentiate the effects of supervision and, say, smiling. If the effects of supervision depended on whether the supervisor did his work with a smile or officiously, important contingencies would be specified that determine the particular type of supervision that causes a particular effect. Such contingency-specifying restrictions to generality are very important. They more accurately delimit the causal construct, which might be "supervision with a smile" rather than the more general but less accurate "supervision."

The kind of specification we have just been discussing concerns variables that are inadvertently manipulated at the same time as the intended treatment or that are inadvertently measured as part of an effect construct. It is more difficult to spell out the implications for construct validity of "developmental sequences," which are the processes that causally follow from the treatment and mediate its consequences. For example, close supervision by a foreman might mean that workers can ask for, and receive, task-relevant feedback that increases the quality or quantity of their performance. Alternatively, workers might feel resentment that their freedom is being curtailed by the supervision and might work less. The feedback and resentment process are consequences that presumably depend on who the foreman or worker is, how past relations have been in the particular work environment, and so forth. The researcher, therefore, faces the following dilemma: Should the treatment be specified as, say, "closeness of supervision," or "closeness of supervision which leads to task-relevant feedback"? The latter is probably the construct that led to the observed effects even though it was not the planned causal construct. We are presently inclined not to include developmental sequences under the heading of construct validity since they do not have to do with the correspondence between an operational treatment or measure and its referent construct. However, developmental sequences are very important in their own right. They help specify why a particular treatment is effective and so contribute to developing theory about the conditions under which the treatment will or will not be expected to have the observed effect. It will not have the effect, for instance, in any setting where manipulating the treatment fails to elicit a change in the developmental process that mediates the effect.

Our discussion of construct validity thus far has implied that a common definition exists of all the constructs about which we would like to test propositions. This is manifestly untrue. It is difficult enough to get an accepted formal definition of, say, being black in the U.S. today (How much African blood makes someone black? Is blackness sociological rather than biological?), let alone a widely accepted formal definition of a less grounded construct such as aggression. (Does aggression of necessity entail "intent to harm"? Does it include verbal as well as physical acts?)

Practically speaking, it would be much simpler if there were accepted definitions; but from a certain philosophical perspective, it is fortunate that we cannot in reality achieve widely accepted definitions of most constructs. This is because propositions about constructs are more reliable if they have been successfully tested, not only across many overlapping operational representations

of a single definition of a construct, but also across representations of many overlapping definitions of the *same* construct. Think how much utility there is in knowing that for many propositions about aggression it is irrelevant whether or not one defines aggression to include "intent to harm," for the same relationships hold with or without the inclusion of intent. And think how accurate and specific our propositions would be if we had information allowing us to differentiate between propositions that are valid for intentional but not unintentional aggression. Our advocacy of multiple operationalism overlaps, therefore, with an advocacy of multiple formal definitionism, provided that all the definitions seem reasonable to most members of a given language group even though not necessarily accepted by all the members of that group.

Though the construct validity of causes and effects has to do with theoretical concepts, it would be a mistake to think that construct validity should only be a concern of the theoretician. First, many treatments in applied research are complex packages of variables rather than indicators of apparently unidimensional constructs. Consequently, it will often be difficult to describe and reproduce the total package, making replication more difficult than if the causal components of the package had been well specified and their independent contributions had been explored. Second, if one knows which components of a treatment are most responsible for an effect, it would be more efficient to reproduce just these features than it would be to reimplement the more expensive total package. For instance, "Sesame Street" was evaluated using a treatment where children and their parents were encouraged to watch the show on a weekly or monthly basis by paid professionals who left toys, books, and games about the show in the home. This total effort increased viewing but cost between \$100 and \$200 per child per viewing season over and above the costs of producing and distributing the show. Since viewing the show without such encouragement costs \$1 to \$2 per child per viewing season, would it not be useful to know whether the educational impact observed because of encouragement might be due to viewing the show (\$1 to \$2) as opposed to factors associated with being encouraged that have nothing to do with viewing—factors that cost between \$100 and \$200 per child? Third, since construct validity involves the fit between operations and referent constructs, it requires a rigorous definition of the referent construct. In policy research, this means being highly explicit about the nature of the problem under investigation and thereby reducing the chances of conducting "irrelevant" research. For instance, what is the hoped-for causal construct to which we want to make generalizations—viewing "Sesame Street" under the conditions most children spontaneously view it, or viewing "Sesame Street" in a context of visits from paid professionals when "Sesame Street"-related artifacts, not paid for by the parents, are in the home?

It is our distinct impression that most applied experimental research is much more oriented toward high construct validity of effects than of causes. This is entirely understandable, for what one wants to see is evidence that the social problem being addressed is at least partially ameliorated—not any problem, but *the* major problem as generally conceived. Thus, great care goes into measuring outcomes, for unless a rigorous measure of "recidivism" or "employ-

ment" or "academic achievement" is used which most competent persons believe to be reasonable, the research is likely to be seen as "irrelevant." While a focus on impact is entirely reasonable in applied research, this focus is often accompanied by a restricted interest in understanding the contingencies on which impact depends. More frequent measurement of the dimensions of the multivariate treatment and more frequent internal data analyses aimed at assessing the contribution of each component to any observed effect would greatly improve applied social research that uses experiments.

It is also our impression that applied researchers are more concerned than basic researchers with the range of cause and, especially, effect constructs over which a relationship can be generalized. In studies of school desegregation, for example, the implications of finding in an urban school district that the achievement of minority school children increased to a small degree would possibly be different if researchers had asked questions about "white flight" from the city than if such questions had not been raised. The whole concern with the unanticipated side-effects of innovations reflects a realistic understanding of the utility of having a net of dependent variable measures that tap into many constructs, some of which have been developed after explicit attempts to think through what may be unexpected effects. In our treatment here, construct validity will deal largely with attempts to generalize from operations to constructs but generalizing across constructs will be very briefly considered.

List of Threats to the Construct Validity of Putative Causes and Effects

Here is our list of some threats to construct validity. They all have to do either with the operations failing to incorporate all the dimensions of the construct, which we might call "construct underrepresentation," or with the operations containing dimensions that are irrelevant to the target constructs, which we might call "surplus construct irrelevancies." The list concentrates mostly on the fit between constructs and the way that the research problem is conceptualized, and devotes less attention to generalizing across constructs. Getting the initial question "right" is not as important a construct validity issue as getting one's operations to reflect one's research constructs. The list that follows is about the latter.

Inadequate Preoperational Explication of Constructs

The choice of operations should depend on the result of a conceptual analysis of the essential features of a construct. For instance, by consulting dictionaries (social science or otherwise) and the past literature on a topic, one would find that "attitude" is usually defined as a stable predisposition to respond. This stability is understood either (a) as a consistency across modes of responding to an attitude object (affective, cognitive, and behavioral), or (b) as a consistency in individual responses across time (i.e., as a positive correlation between responses to the same measure given at different time intervals). Such an analysis suggests the inadequacy of the usual procedure of measuring preferences or beliefs at a single time and then calling these responses "attitude." (For an extended discussion of this subject, see Cook and Flay, 1978.)

To give another example, many definitions of aggression include both the intent to harm others and the fact that harm results from actions. This is to distinguish between (a) the black eye one boy gives another as they collide coming round a blind bend, (b) the black eye that one boy gives another to get his candy (instrumental aggression) or to harm him (noninstrumental), and (c) the verbal threat by one child to another that he will give him a black eye unless the other boy gives him some candy.

Since intent and physical harm are stressed in the definition above—which is not the only one possible—only (b) above is adequate as an example of the construct “aggression,” though it will not be adequate for the minority of persons who prefer some other definition of the term. A precise explication of constructs is vital for high construct validity since it permits tailoring the manipulations and measures to whichever definitions emerge from the explication. Sometimes, several formal definitions are reasonable. Resources and the extent to which one formal definition is preferred over others in the local language community will then play important roles in determining the formal definitions used in the research.

Mono-Operation Bias

Many experiments are designed to have only one exemplar of a particular possible cause, and some have just one measure to represent each of the possible effect constructs. Since single operations both underrepresent constructs and contain irrelevancies, construct validity will be lower in single exemplar research than in research where each construct is multiply operationalized in order to triangulate on the referent. There is rarely an adequate excuse for single operations of effect constructs, since it is not costly to gather additional data from alternative measures of the targets. There is more excuse for having only one manipulation of a possible causal construct. This is because increasing the total number of treatments in a factorial design can lead either to very large sample research or to small sizes within each cell of the design should it not be possible to increase the total sample size. Moreover, if one lets irrelevancies in the treatment presentation vary spontaneously from occasion to occasion, this threatens statistical conclusion validity, even though any treatment effects that emerge despite the inflated error are presumably not due to those irrelevancies. Nonetheless, there is really no substitute for deliberately varying two or three exemplars of a treatment, *where possible*. Hence, if one were interested in the expertise of a communicator, one might use, say, three fictitious sources: a distinguished male professor from a well-known university, a distinguished female research scientist from a prestigious research center, and a famous science journalist from West Germany. Then, the variance due to the difference between these sources can be examined to test whether the different combinations of irrelevancies (sex, affiliation, nationality, or academic standing) differently affected responses and whether each expert singly—and the three combined—caused the expected outcome. If sample size did not permit analyzing separately by source, the data could be combined from all three. The investigator could then test whether expertise was effective despite the irrelevant sources of heterogeneity.

To have more than one operational representation of a construct does not necessarily imply that all irrelevancies have been made heterogeneous. Indeed, when all the manipulations are presented the same way, or all the measures use the same means of recording responses, then the method is itself an irrelevancy whose influence cannot be dissociated from the influence of the target construct. Thus, if all the experts in the previous hypothetical example had been presented to respondents in writing, it would not logically be possible to generalize to experts who are seen or heard. Thus it would be more accurate to label the treatment as "experts presented in writing." To cite another example, attitude scales are often presented to respondents without apparent thought to (a) using methods of recording other than paper-and-pencil, (b) varying whether the attitude statements are positively or negatively worded, or (c) varying whether the positive or negative end of the response scale appears on the right or left of the page. On these three points depends whether one can test if "personal private attitude" has been measured as opposed to "paper-and-pencil nonaccountable responses," or "acquiescence," or "response bias."

Hypothesis-Guessing Within Experimental Conditions

The internal validity threats called "resentful demoralization" and "compensation rivalry" were assumed to result because persons who received less desirable treatments compared themselves to persons who received more desirable treatments, making it unclear whether treatment effects of any kind occurred in the treatment group. Reactive research may not only obscure true treatment effects, but also result in effects of diminished interpretability. This is especially true if it is suspected that persons in one treatment group compared themselves to persons in other groups and guessed how the experimenters expected them to behave. Indeed, in many situations it is not difficult to guess what the experimenters hope for, especially in education or industrial organizations. Hypothesis-guessing can occur without social comparison processes, as when respondents know only about their own treatment but persist in trying to discover what the experimenters want to learn from the research.

The problem of hypothesis-guessing can best be avoided by making hypotheses (if present) hard to guess, by decreasing the general level of reactivity in the experiment, or by deliberately giving different hypotheses to different respondents. But these solutions are at best partial, since respondents are not passive and can always generate their own treatment-related hypotheses which may or may not be the same as the experimenters'. Learning an hypothesis does not necessarily imply either the motivation or the ability to alter one's behavior because of the hypothesis. Despite the widespread discussion of treatment confounds that are presumed to result from wanting to give data that will please the researcher—which we suspect is a result of discussions of the Hawthorne effect—there is neither widespread evidence of the Hawthorne effect in field experiments (see reviews by D. Cook, 1967; Diamond, 1974), nor is there evidence of a similar orientation in laboratory contexts (Weber and Cook, 1972). However, we still lack a sophisticated and empirically corroborated theory of the conditions under which hypothesis-guessing (a) occurs, (b) is

treatment specific, and (c) is translated into behavior that (d) could lead to erroneous conclusions about the nature of a treatment construct when (e) the research takes place in a field setting

Evaluation Apprehension

Rosenberg (1969) has reviewed considerable evidence from laboratory experiments which indicates that respondents are apprehensive about being evaluated by persons who are experts in personality adjustment or the assessment of human skills. In such cases respondents attempt to present themselves to such persons as both competent and psychologically healthy. It is not clear how widespread such an orientation is in social science experiments in field settings, especially when treatments last a long time and populations do not especially value the way that social scientists or their sponsors evaluate them. Nonetheless, it is possible that some past treatment effects were due to respondents being willing to present themselves to experimenters in ways that would lead to a favorable personal evaluation. Being evaluated favorably by experimenters is rarely the target construct around which experiments are designed. It is a confound.

Experimenter Expectancies

There is some literature (Rosenthal, 1972) which indicates that an experimenter's expectancies can bias the data obtained. When this happens, it will not be clear whether the causal treatment is the treatment-as-labeled or the expectations of the persons who deliver the treatments to respondents. This threat can be decreased by employing experimenters who have no expectations or have false expectations, or by analyzing the data separately for persons who deliver the treatments and have different kinds or levels of expectancy. Experimenter expectancies are thus a special case of treatment-correlated irrelevancy, and they may well operate in some (but certainly not all) field settings.

Confounding Constructs and Levels of Constructs

Experiments can involve the manipulation of several discrete levels of an independent variable that is continuous. Thus, one might conclude from an experiment that *A* does not affect *B* when in fact *A*-at-level-one does not affect *B*, whereas *A*-at-level-four might well have affected *B* if *A* had been manipulated as far as level four. This threat is a problem when *A* and *B* are not linearly related along the whole continuum of *A*; and it is especially prevalent, we assume, when treatments have only a weak impact. If they do, because low levels of *A* are manipulated, and if conclusions are drawn about *A* without any qualifications concerning the strength of the manipulation, then misleading negative conclusions can be drawn. The best control for this threat is to conduct parametric research in which many levels of *A* are varied and many levels of *B* are measured.

Interaction of Different Treatments

This threat occurs if respondents experience more than one treatment which is common in laboratory research but quite rare in field settings. We do not

know in such an instance whether we could generalize any findings to the situation where respondents received only a single treatment. More importantly, we would not be able to unconfound the effects of the treatment from the effects of the context of several treatments. The solution to this problem is either to give only one treatment to respondents or, wherever possible, to conduct separate analyses of the first and succeeding treatments which respondents received.

Interaction of Testing and Treatment

To which kinds of testing situations can a cause-effect relationship be generalized? In particular, can it be generalized beyond the testing conditions that were originally used to probe the hypothesized cause-effect relationship? The latter is an especially important question when the pretesting of respondents is involved and might condition the reception of the experimental stimulus, although the previously cited work of Lana (1969) suggests that pretest sensitization is far from omnipresent. We would want to know whether the same result would have been obtained without a pretest, and a posttest-only control group is necessary for this. Similarly, if repeated posttest measurements are made, we would want to know whether the same results would be obtained if respondents were posttested once rather than at each delay interval. We would want to know whether the effect does or does not have to be specified as including the frequency of posttest measurement. The recommended solution to this problem is to have independent experimental groups at each delayed-test session.

Restricted Generalizability Across Constructs

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—punctuality, the ability to follow orders, and so on?" When such questions can be answered, we have a fuller picture of a treatment's total impact and are more likely to gain a comprehensive assessment of the program. Sometimes treatments will affect dependent variables quite differently, implying a positive effect on some construct and an unintended negative effect on another. While it is impossible to measure all the constructs that a particular treatment could affect, it is useful to explore with other persons how a treatment might influence constructs other than those that first come to mind in the original formulation of the research question. Particularly in the program evaluation area, we could cite many studies where the guiding research questions were not well explored and where it would have been feasible to collect more outcome measures, making the research more useful.

Construct Validity, Preexperimental Tailoring, and Postexperimental Specification

Our presentation of the construct validity of putative causes and effects has thus far emphasized the researcher critically (a) thinking through how a construct should be defined, (b) isolating the cognate constructs from which any particular construct has to be differentiated, and (c) deciding which measures or manipula-

tions he can use to index the particular hypothetical construct of interest. Then, we emphasized both (d) the need to have multiple measures or manipulations wherever possible. This need does not deny that some measures are better than others but merely indicates that no single measure is perfect and also indicates (e) the need to present the manipulations or measures in multiple delivery modes. All of these points are geared toward helping the researcher answer the major conceptual questions guiding the research, whether the questions are theoretical or applied.

Data analyses do not always produce the desired results that suggest high construct validity. Consider, first, direct measures which are collected to test whether the treatment varied what it should have varied and did not vary what it was not supposed to have varied. If a reliable measure of, say, communicator credibility suggests that a communicator was not perceived to be more credible in one experimental group than another, then it is not easy to say that credibility caused any effects that may have been inferred from the outcome data. The investigator is then forced to become a detective whose goal is to use whatever means are available to specify what might have caused the observed effects if credibility did not.

Next, consider what might happen if the data indicate that a manipulation affected two reliably measured exemplars of a particular construct but not three others that were equally well measured. How is the effect to be labeled in this case, since the planned label does not fit all the results and so seems inappropriate? Feldman's (1968) experiment in Boston, Athens, and Paris offers a concrete example of this. He used five measures of "cooperation" in an effort 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 foregoing chances to cheat in ways that would be to one's advantage. Thus, the data forced Feldman to specify two kinds of "cooperation" (involving low-cost favors versus foregoing one's own financial advantage) where initially he had tailored his measures to reflect what he had hoped was the unitary construct of cooperation. Moreover, since his respecification of the constructs came after the data were received we can place less confidence in them than might otherwise have been warranted. This is not to downplay Feldman's research, which was exemplary given his research question. If he had not had the five measures, a much less differentiated—and hence less accurate—picture would have emerged of the differences in help given to compatriots and foreigners.

The important point is that construct validity consists of more than merely assessing the fit between planned constructs and the operations that were tailored to these constructs. One can use the obtained pattern of data to edit one's thinking about both the cause and effect constructs, and one can suggest, *after the fact*, other constructs that might fit the data better than those with which the experiment began. Often, the data force one to be more specific in one's labeling than originally planned, as in the Feldman example or with the research of Parker (1963), who set out to test whether the introduction of television caused a decrease in per capita library circulation. He finally concluded that it did for the circulation of

fiction books but not for the circulation of factual ones. The process of hypothesizing constructs and testing how well treatment and outcome operations fit these constructs is similar whether it occurs before the research begins or after the data are received. The major difference is that in the later stage one specifies constructs that fit the data, whereas in the earlier stage one derives operations from constructs.

In their pathfinding discussion of construct validity, Cronbach and Meehl (1955) stressed the utility of drawing inferences about constructs from the fit between patterns of data that would be predicted if a particular theoretical construct was operating and the multivariate pattern of data was actually obtained in the research. They used the term "nomological net" to refer to the predicted pattern of relationships that would permit naming a construct. For instance, a current version of dissonance theory predicts that being underpaid for a counterattitudinal advocacy will result in greater belief change than being overpaid, provided that the individual who makes the advocacy thinks he has a free choice to refuse to perform the advocacy. The construct "dissonance" would therefore be partially validated if experimental data showed that underpayment caused more belief change than overpayment but only under free choice conditions. However, the fit between the complex prediction and the complex data only facilitates belief in "dissonance" to the extent that other theoretical constructs could not explain this same data pattern. Bem (1972) obviously believes that reinforcement constructs do as good a job of complex prediction in this case as "dissonance."

We have implicitly used the "nomological net" idea in our presentation of construct validity. First, we discussed the usefulness—for labeling the treatment—of examining whether the planned treatment is related to direct measures of the treatment process and is not related to cognate processes. Second, we discussed the advantages of determining in what ways the outcome variables are related to treatments and the type of treatment that could have resulted in such a differentiated impact. For instance, if the introduction of television decreases the circulation of fiction but not fact books, one can hypothesize that the causal impact is mediated by television taking time away from activities that are functionally similar—such as fantasy amusement—but not from functionally dissimilar activities—such as learning specific facts. However, our emphasis has differed slightly from that of Cronbach and Meehl (1955) inasmuch as we are more interested in fitting cause and effect operations to a generalizable construct (see Campbell, 1960—the discussion of "trait validity") than we are in using complex predictions and data patterns to validate entirely hypothetical scientific constructs like "anxiety," "intelligence" or "dissonance." However, we readily acknowledge that the way the data turn out in experiments helps us edit the constructs we deal with, as when we find that a foreman's "supervision" has different consequences from less than ten feet as opposed to more than ten feet.

EXTERNAL VALIDITY

Introduction

Under external validity, Campbell and Stanley originally listed the threat of not being able to generalize across exemplars of a particular presumed cause or effect construct. We have obviously chosen to incorporate this feature under con-

struct validity as "mono-operation bias." The reason for listing this threat differently from Campbell and Stanley is not fundamental. Rather it is meant to emphasize that most researchers want to draw conclusions about constructs, but the Campbell and Stanley discussion had a flavor of definitional operationalism, although a *multiple* definitional operationalism. We have tried to avoid this flavor by invoking construct validity to replace generalizing across cause and effect exemplars. The other features of Campbell and Stanley's conceptualization of external validity are preserved here and elaborated upon. They have to do with (1) generalizing to particular target persons, settings, and times, and (2) generalizing across types of persons, settings, and times.

Bracht and Glass (1968) have succinctly explicated external validity, pointing out that a two-stage process is involved: a target population of persons, settings, or times has first to be defined and then samples are drawn to represent these populations. Very occasionally, the samples are drawn from the populations with known probabilities, thereby maximizing the final representativeness discussed in textbooks on sampling theory (e.g., Kish, 1965). But usually the samples cannot be drawn so systematically and are drawn instead because they are convenient and give an intuitive impression of representativeness, even if it is only the representativeness entailed by class membership (e.g., I want to generalize to Englishmen and the people I found on streetcorners in Birkenhead, England, belong to the class called Englishmen). Accidental sampling, as it is technically labeled, gives us no guarantee that the achieved population (a subset of Englishmen who hang around streetcorners in Birkenhead) is representative of the target population of which they are members. Consequently, we find it useful to distinguish among (1) target populations, (2) formally representative samples that correspond to known populations, (3) samples actually achieved in field research, and (4) achieved populations.

One of many examples that could be cited to illustrate these last points concerns the design of the first negative income tax experiment. Practical administrative considerations led to the study being conducted in a few localities within New Jersey and in one city in neighboring Pennsylvania. Since the basic question guiding the research did not require such a restricted geographical location, the New Jersey-Pennsylvania setting must be considered a limitation which reduces one's ability to generalize to the implicit target population of the whole United States. (To criticize the study because the achieved sample of settings was not formally representative of the target population may appear unduly harsh in light of the fact that financial and logistical resources for the experiment were limited, and so sampling was conducted for convenience rather than formal representativeness. We shall return to this point later. *For the present, however, it is worth noting that accidental samples of convenience do not make it easy to infer the target population, nor is it clear what population is actually achieved.*)

Generalizing to well-explicated target populations should be clearly distinguished from generalizing *across* populations. Each is germane to external validity: the former is crucial for ascertaining whether any research goals that specified populations have been met, and the latter is crucial for ascertaining which different populations (or subpopulations) have been affected by a treatment, i.e., for assessing how far one can generalize. Let us give an example.

Suppose a new television show were introduced that was aimed at teaching basic arithmetic to seven-year-olds in the United States. Suppose, further, that one could somehow draw a random sample of all seven-year-olds to give a representative national sample within known limits of sampling error. Suppose, further, that one could then randomly assign each of the children to watching or not watching the television show. This would result in two randomly formed, and thus equivalent, experimental groups which were representative of all seven-year-olds in the United States. Imagine, now, that the data analysis indicated that the average child in the viewing group gained more than the average child in the nonviewing group. One could generalize such a finding to the average seven-year-old in the nation, the target population of interest.

This is equivalent to saying that the results were obtained *despite possible variations in how much different kinds of children in the experimental viewing group may have gained from the show*. A more differentiated data analysis might show that the boys gained more than the girls (or even that only the boys gained), or the analysis might show that children with certain kinds of home background gained while children from different backgrounds did not. Such differentiated findings indicate that the effects of the televised arithmetic show could not be generalized *across* all subpopulations of seven-year-old viewers, even though they could be generalized *to* the population of seven-year-old viewers in the United States.

To generalize across subpopulations like boys and girls logically presupposes being able to generalize to boys and girls. Thus, the logical distinction between generalizing to and across should not be overstressed. The distinction is most useful for its practical implications insofar as many researchers who are concerned about generalizing *across* populations are usually not as concerned with careful samplings as are persons who want to generalize *to* target populations. Many researchers with the former focus would be happy to conclude that a treatment had a specific effect with the particular achieved sample of boys or girls in the study, irrespective of how well the achieved population of boys or girls can be specified.

The distinction between generalizing to target populations and across multiple populations or subpopulations is also useful because commentators on external validity have often implicitly stressed one over the other. For instance, some persons discuss external validity as though it were only about estimating limits of generalizability, as is evidenced by comments such as: "Sure, the treatment affected seven-year-olds in Tucson, Arizona, and that was your target group. But what about children of different ages in other areas of the United States?" Other commentators discuss external validity exclusively in terms of the fit between samples and target populations, as is evidenced by comments such as: "I'm not sure whether the treatment is really effective with children who have learning disabilities, for if you look at the pretest achievement means for the groups in your experiment, you'll see that they are as high as the test publisher quotes for the national average. How could children with learning disabilities have scored so high? I doubt that the research really involved the kind of child you said it did."

Finally, we make the distinction between generalizing to and across in order to emphasize the greater stress that we shall place in this presentation on generalizing

across. The rationale for this is that formal random sampling for representativeness is rare in field research, so that strict generalizing to targets of external validity is rare. Instead, the practice is more one of generalizing across haphazard instances where similar-appearing treatments are implemented. Any inferences about the targets to which one can generalize from these instances are necessarily fallible and their validity is only haphazardly checked by examining the instances in question and any new instances that might later be experimented upon. It is also worth noting that the formal generalization to target populations of persons is often associated with large-scale experiments. These are often difficult to administer both in terms of treatment implementation and securing high-quality measurement. Moreover, attrition is almost inevitable, and so the sample with which one finishes the research may not represent the same population with which one began the research. A case can be made, therefore, that external validity is enhanced more by a number of smaller studies with haphazard samples than by a single study with initially representative samples if the latter could be implemented. Of course, it should not be forgotten that all the haphazard instances of persons and settings that are examined can belong to the class of persons or settings to which one would like to be able to generalize research findings. Indeed, they should belong to such a class.

List of Threats to External Validity

Tests of the extent to which one can generalize across various kinds of persons, settings, and times are, in essence, tests of statistical interactions. If there is an interaction 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. Where effects of different magnitude exist, we must then specify where the effect does and does not hold and, hopefully, begin to explore why these differences exist. Since the method we prefer of conceptualizing external validity involves generalizing across achieved populations, however unclearly defined, we have chosen to list all of the threats to external validity in terms of statistical interaction effects.

Interaction of Selection and Treatment

In which categories of persons can a cause-effect relationship be generalized? Can it be generalized beyond the groups used to establish the initial relationship—to various racial, social, geographical, age, sex, or personality groups? Even when respondents belong to a target class of interest, systematic recruitment factors lead to findings that are only applicable to volunteers, exhibitionists, hypochondriacs, scientific do-gooders, those who have nothing else to do, and so forth. One feasible way of reducing this bias is to make cooperation in the experiment as convenient as possible. For example, volunteers in a television-radio audience experiment who have to come downtown to participate are much more likely to be atypical than are volunteers in an experiment carried door-to-door. An experiment involving executives is more likely to be ungeneralizable if it takes a day's time than if it takes only ten minutes, for only the latter experiment is likely to include those people who have little free time.

Interaction of Setting and Treatment

Can a causal relationship obtained in a factory be obtained in a bureaucracy, in a military camp, or on a university campus? The solution here is to vary settings and to analyze for a causal relationship within each. This threat is of particular relevance to organizational psychology since its settings are on such disparate levels as the organization, the small group, and the individual. When can we generalize from any one of these units to the others? The threat is also relevant because of the volunteer bias as to which organizations cooperate. The refusal rate in getting the cooperation of industrial organizations, school systems, and the like must be nearer 75% than 25%, especially if we include those that were never contacted because it was considered certain they would refuse. The volunteering organizations will often be the most progressive, proud, and institutionally exhibitionist. For example, Campbell (1956), although working with Office of Naval Research funds, could not get access to destroyer crews and had to settle for high-morale submarine crews. Can we extrapolate from such situations to those where morale, exhibitionism, pride, or self-improvement needs are lower?

Interaction of History and Treatment

To which periods in the past and future can a particular causal relationship be generalized? Sometimes an experiment takes place on a very special day (e.g., when a president dies), and the researcher is left wondering whether he would have obtained the same cause-effect relationship under more mundane circumstances. Even when circumstances are relatively more mundane, we still cannot logically extrapolate findings from the present to the future. Yet, while logic can never be satisfied, "commonsense" solutions for short-term historical effects lie either in replicating the experiment at different times (for other advantages of consecutive replication, see Cook, 1974a) or in conducting a literature review to see if prior evidence exists which does not refute the causal relationship.

Models to Be Followed in Increasing External Validity

In many instances researchers know that they want to generalize to specific target populations of persons, settings, or times. This is particularly the case in much applied research, although it is also found among basic researchers interested in contingency theories (e.g., a theory of schizophrenia, or of behavior in street settings which require the ability to make references about schizophrenics and street settings, however these are defined). Clearly, when target populations are specified, it is necessary that the research samples be "representative" in some way.

In other instances, the researchers may not have specific populations in mind. This is most likely to be the case with someone developing a general theory, but it is also sometimes appropriate in developing more limited theories or conducting applied research. For instance, the applied researcher in education may have fourth-grade inner-city children as the primary intended target population. But he or she may not have a specific target group of persons in mind for giving the achievement tests. Yet if all the posttest measurement is conducted by middle-class testers hired for the particular project, the researcher cannot extrapolate

beyond such testers. In a sense, he or she has drawn an unintended secondary sample with an unclear population referent that has no intrinsic interest, and without further evidence no generalization beyond such testers is warranted. How much better it would be if the irrelevant factor of tester social status were not fixed but varied. Then, one could analyze the data to test whether similar effects were obtained despite background differences among testers—that is, one could test whether it is possible to generalize *across* factors like tester status that are irrelevant to major research goals.

When a target population has been specified, it is appropriate—where possible—to draw up a sampling frame and select instances so that the sample is representative of the population within known limits of sampling error. Many textbooks on sampling theory exist and are informative about the advantages and disadvantages of drawing samples in different ways. Formally speaking, the most representative samples will be those that are randomly chosen from the population, and it is possible for these randomly selected units to be randomly assigned to various experimental groups. We might label the first stage in such a two-stage randomization process as following the *random sampling for representativeness model*.

It is probably only feasible to follow this model when sampling intended primary targets of persons, the more so if generalization to a limited setting is required (e.g., to residents of Detroit, rather than the whole United States). However, random sampling for representativeness is theoretically possible on a larger scale, particularly if multistage area sampling of, say, the whole nation is undertaken. But studies on this scale require considerable resources. Moreover, while it is clear that the model can be followed for some issues where it is important to generalize to particular target populations of persons, it is less clear whether it is often feasible to generalize to target settings, except where these are highly restricted. For instance, by selecting a representative national sample of persons, one should be able to generalize to various geographical settings (i.e., cities, towns, and the like). But regions do not exhaustively define settings, and the nationwide representative experiments of which we are aware—all of which embed treatments within polling studies—take place in the respondents' homes rather than in the street or in factories. While a restriction to living rooms is desirable for anyone interested in generalizing to settings where opinion polls typically take place, it is less desirable for the majority of researchers who have no such particular target setting in mind. The point to be noted is that the model of random sampling for representativeness requires considerable resources which are probably more readily available for sampling target populations of persons than of settings or historical times and which are probably more available for restricted populations of persons (e.g., inhabitants of Detroit) than for the United States at large.

A second model for increasing external validity is the *model of deliberate sampling for heterogeneity*. Here the concern is to define target classes of persons, settings, and times and to ensure that a wide range of instances from within each class is represented in the design. Thus, a general educational experiment might be designed to include boys and girls from cities, towns, and rural settings who differ widely in aptitude and in the value placed on achievement in their home settings. The task would then be to test whether an educational innovation has

comparable effects in each of the subgroups of children and settings. If the achieved sample sizes do not permit this, then the task would be to test whether the innovation has observable effects *despite* differences between kinds of children and kinds of settings. The first task involves an obvious attempt at multiple replication, either by testing for interactions of the treatment and student characteristics or by statistical tests of whether treatment has any observed effects within each group. The second task involves testing whether a treatment effect is obtained even though differences between persons and settings are not taken into account in the data analysis and are inflating the error terms that are used for testing treatment effects.

Deliberate sampling for heterogeneity does not require random sampling at any stage in the sampling design. Hence one cannot—technically speaking—generalize from the achieved samples to any formally meaningful populations. All one has are purposive quotas of persons with specified attributes. These quotas permit one to conclude that an effect has or has not been obtained across the particular variety of samples of persons, settings, and times that were under study, which is like saying: "We tried to have children of Types I and II in the experiment in order to see if the effect would hold with each of them. It did. We're not sure how well one can generalize from our particular achieved samples of children to children of Type I and Type II in general, but at least we learned that the effect holds with at least one sample of Type I children and at least one sample of Type II children. What we cannot do with any confidence is specify the populations of children involved." To have a sample of persons in an experiment with Type I characteristics is not at all sufficient for formally concluding that we can generalize any findings to the average Type I persons.

When one samples nonrandomly, it is usually advantageous to obtain opportunistic samples that differ as widely as possible from each other. Thus, if it were possible, one might choose to implement a treatment both in a "Magnet School," that is, a school established to exemplify teaching conditions at their presumed best, and also in one of the city's worst problem schools. If each instance produced comparable effects, then one might begin to suspect that the effect would hold in many other kinds of schools. However, there is a real danger in having only extreme instances at each end of some implicit, impressionistic continuum. This can best be highlighted by asking: "What would you conclude about external validity if an effect were obtained at one school but not the other?" In this case, one would be hard pressed to conclude anything about the effects of the innovation in the majority of schools between the extremes. For this reason, it is especially advantageous if deliberate sampling for heterogeneity results in at least one instance of the impressionistic mode of the class under investigation as well as instances at each extreme. In other words, at least one instance should be representative of the "typical school" of a particular city (or nation), and at least one instance representing the best and worst schools.

The model of deliberate sampling for heterogeneity is especially useful in avoiding the pitfall of restricted inference that results from the failure to consider sampling questions about secondary targets of inference (e.g., the social class of educational testers as opposed to the social class of school children). Unless one has good reasons for matching the class of testers and children, the model based

on seeking heterogeneity indicates that it would be unwise to sample from a homogeneous group of testers with a common background. Comparable background does not mean identical testers, of course, for testers of any one class differ from each other in a multitude of ways. Nonetheless, social class is relatively homogeneous, should plausibly affect test scores, and is an irrelevant source of homogeneity that can often be made heterogeneous at little or no extra cost.

Deliberate purposive sampling for heterogeneity is usually more feasible than random sampling for representativeness. Imagine conducting an experiment in a school district to which you want to generalize. You could draw up a list of schools and randomly select a number of them in order to generalize with confidence. But resources and politics often prevent working with so many schools. Instead, the researcher is often lucky if he can afford (or be granted) access to more than one or two schools—an achieved sample of convenience. This being so, the researcher should seek convenient samples which differ considerably on attributes that he or she especially wants to generalize across and should take care not to be inadvertently restricted to populations, particularly those of secondary interest.

A third model for extending external validity is the *impressionistic modal instance model*. Here, the concern is to explicate the kinds of persons, settings, or times to which one most wants to generalize and then to select at least one instance of each class that is impressionistically similar to the class mode. We alluded to this strategy earlier in detailing the desirability of having at least one school similar to the average school in a district. To achieve this aim is simple. Where comprehensive records exist, one can detail the average size of schools, average achievement levels, average per capita expenditure, and so forth, and choose one or more schools that most closely approximate the modal school characteristics that have been "drawn up." Should there be no obvious single mode, one can then define the multiple modes and try to obtain at least one sample of each. Thus, in many urban school districts, one might find three modes corresponding to all-black, all-white, and heavily desegregated schools. Then a choice of one group from each class would be called for. Where no suitable archive measures exist, it should nonetheless be possible for the researcher to sample the opinions of experts and interested parties to obtain their impression of what the average school or student is like. A composite impression is then derived for all the single impressions, and this composite forms the framework for deciding the order in which potential respondents (or which access-granting authorities) should be approached for permission to do the study in their locale.

The definition and selection of modal instances is probably most easy in consultant work or project evaluation where very limited generalization is required. For instance, an industrial manager knows that he or she wants to generalize to the present work force in its current setting carrying out its present tasks the effectiveness of which is measured by means of locally established indicators of productivity, profitability, absenteeism, lateness, and the like. The consultant or evaluator then knows that he or she has to select respondents and settings to reflect these circumscribed targets. A feasible method is to concentrate on sampling impressionistically modal instances if sampling has to be carried out at all. (The evaluator might also do well to select out exemplary instances in order to gain a

preliminary understanding of what a business or project is capable of. But that is another matter.)

The determination of modal instances is more difficult the closer one comes to theoretical research. This is because target populations are less likely to be specified. For instance, in testing propositions about "helping" behavior, it is not desirable to generalize only to workers who are presently employed in a particular factory, working at a particular task, and producing a particular product. The persons, the settings, the task, and the product would be irrelevant to any helping theory. Yet—logically speaking—the factors incorporated into a particular test of a proposition about helping determine the external validity of the findings, and the researcher presumably does not welcome this restriction. Instead, he or she would like to generalize to all persons (in the United States? beyond our shores?), all settings (the street, the home, the factory?), and all tasks (helping someone who has fainted, helping the permanently disabled?). The feasibility of confident generalizations of such breadth is low, and the most that the basic researcher can do is to attempt to replicate his or her original findings across settings with different restrictions or to wait until others have conducted the replications. Sampling for heterogeneity is at issue here rather than sampling to obtain impressionistically modal instances that the researcher cannot convincingly define.

It should be clear by now that, where targets are specified, the model of random sampling for representativeness is the most powerful model for generalizing and that the model of impressionistic modal instances is the least powerful. The model of heterogeneous instances lies between the two. However, the last model has advantages over the other two in that it can be used when no targets are specified and the major concern is not to be limited in one's generalizations. Moreover, it can be used with small numbers of samples of convenience. In many cases the random selection of instances results in generalizing to targets that are of minimal significance for persons whose interests differ from those of the original researcher's. For instance, to be able to generalize to all whites living in the Detroit area, while of interest for some purposes, is generally of little interest to most people. However, it is worth noting that whites in Detroit differ in age, SES, intelligence, and the like so that it is possible to test whether a particular treatment can have similar effects *despite* such differences. In addition, subgroup analyses can be conducted to examine generality across subpopulations. In other words, formal randomization from populations of low interest can be used to test causal relationships across heterogeneous subpopulations. In other words, an important function of random samples is to permit examining the data for differential effects on a variety of subpopulations. Given the negative relationships between "inferential" power and feasibility, the model of heterogeneous instances would seem most useful, particularly if great care is made to include impressionistically modal instances among the heterogeneous ones.

In the last analysis, external validity—like construct validity—is a matter of replication. It is worth noting that one can have multiple replication both *within* a single study—subgroup analyses exemplify this—and also *across* studies—as when one investigator is intrigued by a pattern of findings and tries to replicate them using his or her own procedures or procedures that have been closely modeled on the original investigators'.

Three dimensions of replication are worth noting. First, is the simultaneous or consecutive replication dimension, with the latter being preferable since it offers some test, however restricted, of whether a causal relationship can be corroborated at two different times. (Generalizing across times is necessarily more difficult than generalizing across persons or settings.) Second is the independent or nonindependent investigator dimension, with the former being more important, especially if the independent investigators have different expectations about how an experiment will turn out. Third is the dimension of demonstrated or assumed replication. The former is assessed by explicit comparisons among different types of persons and settings where some persons did or did not receive a particular treatment. The latter is inferred from treatment effects that are obtained with heterogeneous samples, but no explicit statistical cognizance is taken of the differences among persons, settings, and times. Demonstrated replication is clearly more informative than assumed, for to obtain an effect with a mixed sample of, say, boys and girls, does not logically entail that the effect could be obtained separately for both boys and girls. It only entails that the effect was obtained despite any differences between boys and girls in how they reacted to the treatment.

The difficulties associated with external validity should not blind experimenters to practical steps that can be taken to increase generalizability. For instance, one can often deliberately choose to perform an experiment at three or more sites where different kinds of persons live or work. Or, if one can randomly sample, it is useful to do so even if the population involved is not meaningful, for random sampling ensures heterogeneity. Thus, in their experiment on the relationship between beliefs and behavior about open housing, Brannon et al. (1973) chose a random sample of all white households in the metropolitan Detroit area. While few of us are interested in generalizing to such a population, the sample was nonetheless considerably more heterogeneous than that used in most research, despite the homogeneity on the attributes of race and geographical residence.

In addition, our three models for increasing external validity can be used in combination, as has been achieved in some survey research experiments on improving survey research procedures (Schuman and Duncan, 1974). Usually, random samples of respondents are chosen in such experiments, but the interviewers are not randomly chosen; they are merely impressionistically modal of all experienced interviewers. Moreover, the physical setting of the research is limited to one target setting that is of little interest to anyone who is not a survey researcher—the respondent's living room—and the range of outcome variables is usually limited to those that survey researchers typically study—that is, those that can be assessed using paper and pencil. However, great care is normally taken that these questions cover a wide range of possible effects, thereby ensuring considerable heterogeneity in the effect constructs studied.

Our pessimism about external validity should not be overgeneralized. An awareness of targets of generalization, of the kinds of settings in which a target class of behaviors most frequently occurs, and of the kinds of persons who most often experience particular kinds of natural treatments will, at the very least, prevent the designing of experiments that many persons shrug off willy-nilly as "irrelevant." Also, it is frequently possible to conduct multiple replications of an experiment at different times, in different settings, and with different kinds of

experimenters and respondents. Indeed, a strong case can be made that external validity is enhanced more by many heterogeneous small experiments than by one or two large experiments, for with the latter one runs the risks of having heterogeneous treatment, measures that are not as reliable as they could be, and measures that do not reflect the unique nature of the treatment at different sites. Many small-scale experiments with local control and choice of measures is in many ways preferable to giant national experiments with a promised standardization that is neither feasible nor even desirable from the standpoint of making irrelevancies heterogeneous.

RELATIONSHIPS AMONG THE FOUR KINDS OF VALIDITY

Internal Validity and Statistical Conclusion Validity

Drawing false positive or false negative conclusions about causal hypotheses is the essence of internal validity. This was a major justification for Campbell (1969) adding "instability" to his list of threats to internal validity. "Instability" was defined as "unreliability of measures, fluctuations in sampling persons or components, autonomous instability of repeated or equivalent measures," all of which are threats to drawing correct conclusions about covariation and hence about a treatment's effect. (What precipitated the need for this additional threat was the viewpoint of some sociologists who had argued against using tests of significance unless the comparison followed random assignment to treatments. See Winch and Campbell, 1969, for further details.)

The status of statistical conclusion validity as a special case of internal validity can be further illustrated by considering the distinction between bias and error. Bias refers to factors which systematically affect the value of means; error refers to factors which increase variability and decrease the chance of obtaining statistically significant effects. If we erroneously conclude from a quasi-experiment that *A* causes *B*, this might either be because threats to internal validity bias the relevant means or because, for a specifiable percentage of possible comparisons, sample differences as large as those found in a study would be obtained by chance. If we erroneously conclude that *A* does not affect *B* (or cannot be demonstrated to affect *B*), this can either be because threats to internal validity bias means and obscure true differences or because the uncontrolled variability obscures true differences. Statistical conclusion validity is concerned not with sources of systematic bias but with sources of random error and with the appropriate use of statistics and statistical tests.

An important caveat has to be added to the preceding statement that random errors reduce the risk of statistically corroborating true differences. This does not imply that random errors invariably inflate standard errors or that they never lead to false positive conclusions about covariation. Let us try to illustrate these points. Imagine multiple replications of an unbiased experiment where the treatment had no effect. The distribution of sample mean differences should be normal with a mean of zero. However, many individual sample mean differences will not be zero. Some will inevitably be larger or smaller than zero, even to a statistically significant degree.

Imagine, now, the same assumptions except that bias is operating. Because of the bias, the distribution of sample mean differences will no longer have a mean of zero, and the difference from zero indicates the magnitude of the bias. However, the point to be emphasized is that some sample mean differences will be as large when there is bias as when there is not, although the proportion of differences reaching the specified magnitude will vary between the bias and nonbias cases depending on the direction and magnitude of bias. Since sampling error, which is one kind of random error, affects both sample means and variances, it can lead to both false positive and false negative differences. In this respect, sampling error is like internal validity. But it is unlike internal validity in that it cannot affect population means. Only sources of bias—threats to internal validity—can do the latter.

Construct Validity and External Validity

Making generalizations is the essence of both construct and external validity. It is instructive, we think, to analyze the similarities and differences between the two types of validity. The major similarity can perhaps best be summarized in terms of the notion of statistical interaction—that is, the sign or direction of a treatment effect differs across populations. It is easy to see how person, setting, and time variables can moderate the effectiveness of a treatment. It is probably also easy to see how an estimate of the effect may depend on such threats to construct validity as the number of treatments a respondent receives or the frequency with which outcomes are measured. It may be less easy to see how a treatment effect can interact with (i.e., depend on) the particular method used for collecting data (mono-method bias), or the expectancies of the persons implementing a treatment (experimenter expectancies), or the guesses that respondents make about how they are supposed to behave (hypothesis-guessing). But in all these instances an internally valid effect can be obtained under one condition (say, when paper-and-pencil measures of attitude are used) and a different, but still valid, effect may result when attitude is measured some other way.

Specifying the factors that codetermine the direction and size of a particular cause-effect relationship is useful for inferring cause and effect constructs. This is because some of the causes or effects that might explain a particular relationship observed under one condition may not be able to explain why there are different causal relationships under other conditions. It should especially be noted that specifying the populations of persons, settings, and times over which a relationship holds can also clarify construct validity issues. For instance, suppose a negative income tax causes more married women than men to withdraw their labor from the labor market (see the summary statements of the four negative income tax experiments in Cook, Del Rosario, Hennigan, Mark, and Trochim, 1978). Such an action might suggest that the causal treatment can be understood, not just in monetary terms but also in terms of a possible shift in economic risks (i.e., where the family breadwinner is guaranteed an income, the withdrawal of his or her labor could have extremely serious consequences if the income guarantee were withdrawn or if the guaranteed sum failed to rise with inflation. But where a source of more marginal income is involved—as

with some married women—the withdrawal of their labor is less critical since the family is not so heavily dependent on that one source of income.) Other interpretations of why men and women are affected differently are also possible. Their existence highlights the difficulty of inferring causal constructs where the clarifying inference is indirect, being based on differences in responding across populations rather than on attempts to refine the causal operations directly so that they better fit a planned construct. The major point to be noted, however, is that both external and construct validity are concerned with specifying the contingencies on which a causal relationship depends and all such specifications have important implications for the generalizability and nature of causal relationships. Indeed, external validity and construct validity are so highly related that it was difficult for us to clarify some of the threats as belonging to one validity type or another. In fact, two of them are differently placed in this book than in Cook and Campbell (1976). These are “the interaction of treatments” and “the interaction of testing and treatment.” They were formerly included as threats to external validity on grounds that the number of treatments and testings were part of the research setting. On reflection, however, we think they are more useful for specifying cause and effect constructs than for delimiting the settings under which a causal relationship holds, though they obviously can serve both purposes.

The major difference between external and construct validity has to do with the extent to which real target populations are available. In the case of external validity the researcher often wants to generalize to specific populations of persons, settings, and times that have a grounded existence, even if he or she can only accomplish this by impressionistically examining data patterns across accidental samples. However, with cause and effect constructs it is more difficult to specify a particular construct—what, for instance, *is* aggression? Any definitions would be disputed and would not have the independent existence of, say, the population of American citizens over 18 years of age. Even though the latter is a theoretical construct, it is obviously more grounded in reality than such constructs as “attitude towards authority” or “a negative income tax.”

Issues of Priority Among Validity Types

Some ways of increasing one kind of validity will probably decrease another kind. For instance, internal validity is best served by carrying out randomized experiments, but the organizations willing to tolerate these are probably less representative than organizations willing to tolerate passive measurement. Second, statistical conclusion validity is increased if the experimenter can rigidly control the stimuli impinging on respondents, but this procedure can decrease both external and construct validity. And third, increasing the construct validity of effects by multiply operationalizing each of them is likely to increase the tedium of measurement and to cause either attrition from the experiment or lower reliability for individual measures.

These countervailing relationships—and there are many others—suggest how crucial it is to be explicit about the priority ordering among validity types in planning any experiment. Means have to be developed for avoiding all unnecessary trade-offs between one kind of validity and another, and to minimize the

loss entailed by the necessary trade-offs. However, since some trade-offs are inevitable, we think it unrealistic to expect that a single piece of research will effectively answer all of the validity questions surrounding even the simplest causal relationship.

The priority among validity types varies with the kind of research being conducted. For persons interested in theory testing it is almost as important to show that the variables involved in the research are constructs *A* and *B* (construct validity) as it is to show that the relationship is causal and goes from one variable to the other (internal validity). Few theories specify crucial target settings, populations, or times to or across which generalization is desired. Consequently, external validity is of relatively little importance. In practice, it is often sacrificed for the greater statistical power that comes through having isolated settings, standardized procedures, and homogeneous respondent populations. For investigators with theoretical interests our estimate is that the types of validity, in order of importance, are probably internal, construct, statistical conclusion, and external validity.

We also estimate that the construct validity of causes may be more important for such researchers than the construct validity of effects, particularly in psychology. Think, for example, of how simplistically "attitude" is operationalized in many persuasion experiments, or "cooperation" in bargaining studies, or "aggression" in studies of interpersonal violence. Think, on the other hand, about how much care goes into demonstrating that a particular manipulation varied "cognitive dissonance" and not reactance, communicator expertise and not experimenter expectancies or evaluation apprehension. Might not the construct validity of effects rank lower than statistical conclusion validity for most theory-testing researchers? If it does, this would be ironic since multiple operationalism makes it easier to achieve higher construct validity of effects than of causes.

Much applied research has a different set of priorities. It is concerned with testing whether a particular problem has been alleviated by a treatment—recidivism in criminal justice settings, achievement in education, or productivity in industry (high internal validity and high construct validity of the effect). It is also crucial that any demonstration of change in the indicator be made in a context which permits either wide generalization or generalization to the specific target settings or persons in whom the researcher or his clients are particularly interested (high interest in external validity). The researcher is relatively less concerned with determining the causally efficacious components of a complex treatment package, for the major issue is whether the treatment as implemented caused the desired change (low interest in construct validity of the cause). The priority ordering for many applied researchers is something like internal validity, external validity, construct validity of the effect, statistical conclusion validity, and construct validity of the cause.

For the kinds of causal problems we have been discussing, the primacy of internal validity should be noted for both basic and applied researchers. The reasons for this will be given below, and they relate to the often considerable costs of being wrong about the magnitude and direction of causal relations, and the often minimal gains in external validity that are achieved in moving from

initial accidental samples of convenience that belong in the class to which generalization is desired to other types of samples. Consequently, jeopardizing internal validity for the sake of increasing external validity usually entails a minimal gain for a considerable loss.

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 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*. The relative desirability of randomized experiments over quasi-experiments becomes even clearer in this context, for the former allows stronger tests of causal hypotheses than the latter. This is not to say that the randomized experiment guarantees a perfect test of internal validity. Far from it. However, it usually allows a stronger test than most quasi-experiments; and most of the quasi-experiments we discuss in chapters 3 and 5 of this volume permit stronger tests than the nonexperiments we shall discuss in chapter 7.

Though experiments are designed to test causal hypotheses, and internal validity is the *sine qua non* of causal inference, there are contexts where it would not be advisable to subordinate too much to internal validity. Someone commissioning research to improve the efficiency of his own organization might not take kindly to the idea of testing a proposed improvement in a laboratory setting with sophomore respondents. A necessary condition for meeting such a client's needs is that he can generalize any findings to his own organization and to the indicators of efficiency that he regularly uses for monitoring performance. Indeed, his need in this respect may be so great that he is prepared to sacrifice some gains in internal validity for a necessary minimum of external validity. We would tend to agree with him if increasing internal validity meant going outside his organization or organizations like his own into some completely different type of setting, e.g., the psychological laboratory. In most cases, the desirable minimum of external validity would be that the achieved samples of persons, settings, and measures belong to the specified target "populations," however accidental the samples finally achieved happened to be. However, we would be less inclined to agree with him if class membership were not enough and he insisted on, say, the formal random sampling of respondents when this type of selection precluded random assignment to treatments, which it might if it were feared that many of the potential respondents would refuse to be in the study if random assignment to treatments took place. In this last case, the gain in external validity in moving from accidental samples to samples that were *initially* formally random would not usually seem worth the loss in internal validity that is associated with going from random to systematic assignment to treatments.

Many basic researchers specify target populations when they formulate their guiding research questions, and they want to test causal theories about specific classes of persons (e.g., alcoholics) or settings (e.g., urban ghettos), for their research would be trivialized by any procedures that increased internal validity through conducting research with groups other than alcoholics or in settings other than ghettos. Thus, when targets of generalization are specified in guiding research questions, cognizance has to be taken of this in designing an experiment, and instances should be chosen that at least belong in the class to which generalization is desired. Unfortunately, being a member of a class does not necessarily imply being representative of that class.

SOME OBJECTIONS TO OUR VALIDITY DISTINCTIONS

A number of criticisms of the original Campbell and Stanley distinction between internal and external validity have recently appeared, and we wish to discuss them here (Gadenne, 1976; Kruglanski and Kroy, 1975; Hultsch and Hickey, 1978; Cronbach, in preparation). These critics make partially overlapping but also independent criticisms which we shall address one by one.

The first objection is to the claim that random assignment rules out all threats to internal validity. The argument is made that it is *in principle* impossible to rule out all validity threats because the true cause of an observed effect may be either the planned treatment, or procedural correlates of the treatment, or the interaction of the treatment and the procedures in which the treatment is embedded. The critics cite Rosenthal's work on experimenter expectancies to support this point, arguing that such expectancies are just one of many conceivable, and some as yet inconceivable, forces that operate in an experiment and can be treatment related.

On one level this is an important point, highlighting theorists' concerns with generalizing from operationalized independent variables to theoretical causal constructs. But the objection does not take into account the fact that Campbell and Stanley conceived of procedural variables, like experimenter expectancies, as threats to external and not internal validity. This is because such threats cast doubt on whether a causal relationship can be generalized beyond particular settings (e.g., where the experimenter had an hypothesis about the outcome of the study). They do not cast doubt on whether there was a causal relationship from the independent-variable-as-manipulated to the dependent-variable-as-measured. Indeed, the critics seem quite prepared to acknowledge that causal inference is involved in studies where the experimenters' expectancies may play a role, and their concern is with how the treatment should be labeled. Is it an effect of a particular theory-relevant construct or of the theoretical irrelevancy of "experimenter expectancies"? Such an issue of "confounding" has been discussed in this chapter as an issue of construct validity, while in Campbell and Stanley it was an issue of external validity and was never an issue of internal validity.

Nonetheless, the critics do perform an essential service, for it is indeed false to claim that randomization controls for all threats to internal validity. For instance, one can set up a randomized experiment but still have systematic

selection because of differential attrition. Moreover, the process of distributing valued resources on a random basis (instead of by need or merit, say) can lead to the operation of threats like compensatory rivalry or compensatory equalization. These, in their turn, can lead to false inferences about the effects of a treatment. While randomization is the best single means of increasing our confidence in causal inferences, it is not a panacea. Indeed, a book devoted to quasi-experiments implies that randomized experiments are not achievable at will. In the chapter devoted to randomized experiments we will stress the need to design interpretable quasi-experiments along with randomized experiments so the researcher has strong alternative designs should the initial random assignment to treatments break down. Though this book advocates random assignment, it does so in a more explicitly qualified manner than its predecessors.

The second objection made by the critics is that Campbell and Stanley, while explicitly rejecting inductive inference, nonetheless base their concept of external validity on inductive inference—going from samples to populations. At first glance this seems to be a telling criticism, for the language of external validity is the language of generalizing from samples of persons and settings to populations of persons and settings. However, it should be noted that in previous discussions, Campbell (1969b) has stressed that, because of the problem of induction, all generalization in the social sciences is particularly presumptive and that external validity is inherently more problematic than even internal validity whose bases are more obviously deductive.

It should also be noted that the relationship of samples to populations can be specified in deductive terms that permit falsification. For instance, if one conducts an experiment with a random sample drawn from a well-designated universe (say, the city of Detroit), one can rule out the threat that the universe is biased towards white or male or upper-income inhabitants of the city either by understanding what random selection is and then examining how it was implemented in the experiment in question, or by comparing background characteristics of the sample with (hopefully, recent) census data on the population. Stated more formally, the threat to be ruled out is: There is a race bias in the study. One deduces from this that the percentages of persons from different races who are in the sample should not differ from the percentages in the population to a greater degree than is warranted by sampling error. This deduction is testable by collecting data on the race of persons in the sample. If the percentages differ from what is expected on the basis of the (hopefully, recent) census, it would not be false to say that there is probably a race bias. However, if the percentages are as expected from the census, then it would be false to say there is a race bias though there may be other sources of bias.

External validity can also be deductively tested when sampling is carried out to achieve heterogeneity rather than formal representativeness. The postulate is, say: The treatment does not affect black inhabitants of Detroit. The deduction is: The effect will not be observed among blacks. The falsifying test of this is to examine empirically whether there is a causal relationship among blacks. If there is, one cannot say that the causal relationship generalizes to all blacks, but one can at least say that the relationship is not false when tested with a particular biased sample of blacks. If there is no causal relationship

among blacks, one can confidently conclude that the effect does not hold with all black inhabitants of Detroit. It is simply not the case, therefore, that external validity rests on a base of inductive inference that flies in the face of the acknowledged limitations of inductive inference.

Campbell and Stanley included under external validity threats having to do with generalizing from manipulations and measures to target constructs. Their discussion of these issues was explicitly nonpositivist, espousing multiple operationalism. Nonetheless, there was a flavor of positivism in that the inductive framework may have encouraged readers to think that the operations "some-how" were the constructs, that an observed response really was "aggression" or "love." The present book, in distinguishing between external and construct validity, has been written to avoid such a positivist flavor and to stress that constructs are hypothetical entities not "corporeally" represented by samples of operations.

A third objection the critics make is to note that all guiding research propositions must be couched in general/universal terms whereas internal validity is couched in the language of causal connections involving research operations. The critics wonder how one can have any validity internal to an experiment when the propositions whose validity is being tested are phrased in general terms external to the experiment (e.g., *A* is "causally" related to *B*). The concern with the universal nature of research propositions goes beyond internal validity, of course, as can be seen in the fact that the guiding research propositions are likely to be phrased as: "What is the causal effect of school desegregation on the academic achievement of children in the public schools of Evanston in 1969?" Or, "What was the causal effect of a guaranteed income on the labor force participation of working poor persons in New Jersey and Scranton, Pennsylvania, between 1969 and 1972?" Given the universal terms in these propositions, critics point out that validity depends on the fit between research operations and referent constructs (construct validity) or populations (i.e., external validity). Most critics invoke Brunswik at one point or another and call for a "representative" social science in which (1) the target populations and constructs are clearly stated and sampling takes place so as to represent these populations in the research operations; and (2) the targets are not conceived solely in terms of respondents—the representativeness of settings and procedures is also crucial for Brunswik and his followers.

We have a great deal of sympathy with the position that all aspects of research design test propositions of a general and universal nature and that sampling is the means by which one approximates representing general constructs about causes, effects, types of people, or the like. However, it is easier to conceive of the representativeness of constructs and populations than of relationships among variables. How does one, for instance, sample to represent "causality"? We find it difficult to imagine representative samples of causal instances. Instead, we think that testing the nature of an observed relationship between an independent and dependent variable has to revolve around the particularities of a single study—around details concerning covariation, temporal precedence, and the ruling out of alternative interpretations about the nature of the relationship in the experiment on hand. Of course, we do not deny that the

notion of "cause" is an abstract one and that the single study only approximates causal knowledge. But we believe it is confusing to insist that internal validity is a contradiction in terms because all validity is external, referring to abstract concepts beyond a study and not to concrete research operations within a study. It is confusing because the choice of populations and the fit between samples and populations determines representativeness, whereas neither populations nor samples are necessary for inferring cause.

Nonetheless, the critics make a very useful point, for if the goals of a research project are formulated well enough to permit specifying target constructs and populations, then the research operations have to represent these targets if the research is to be relevant either to theory or to policy. Moreover, a focus on representativeness has historically entailed a heightened sensitivity to unplanned and irrelevant targets that unnecessarily limit generalizability, as when all the persons who collect posttest achievement data in an early childhood experiment with economically disadvantaged children are of the same social class. Clearly, relevant research demands representativeness where target constructs or populations are specified. It also demands heterogeneity where irrelevant populations could limit the applicability of the research. Though we advocate putting considerable resources into the preexperimental explication of relevant theory or policy questions—and hence targets—this should not be interpreted in any way as an exclusive focus. As we tried to demonstrate in the discussion of both construct and external validity, it is sometimes the case that the data, once collected and analyzed, force us to restrict (or extend) generalizability beyond the scope of the original formulation of target constructs and populations. The data edit the kinds of general statements we can make.

For instance, in his experiment on the help given to compatriots and foreigners, Feldman (1968) wanted to generalize to "cooperation." He deduced that if his independent variable affected cooperation, he would find five dependent variable measures related to his treatment. But only two were related, and the data outcomes forced him to conclude tentatively that his treatment was differently related to two kinds of cooperation. Similarly, the designers of the New Jersey Negative Income Tax Experiment wanted to generalize to working poor persons, but the data forced them tentatively to conclude that working poor blacks responded one way to the treatments, working poor persons who were Spanish speaking reacted another way, and working poor whites probably did not respond to the treatments at all. The point is this: While it is laudable to sample for representativeness when targets of generalization are specified in advance—and we heartily endorse such sampling—in the last analysis it is the patterning of data outcomes which determines the range of constructs and populations over which one can claim a treatment effect was obtained. One has always to be alert to the data demanding a respecification of the affected populations and constructs and to the possibility that the affected populations and constructs will not be the same as those originally specified.

A fourth objection has been directed towards Campbell and Stanley's stress on the primacy of internal over external validity. The critics argue that no kind of validity can logically have precedence over another. Of what use, critics say, is a theory-testing experiment if the true causal variable is not what the

researchers say it is; or of what use is a policy experiment about the effects of school desegregation if it involves a school in rural Mississippi when most desegregation is in large, northern cities where white children have fewer alternatives to public schools than in the deep South? This point of view has been best expressed by Snow (1974). He uses the term "referent validity" to designate the extent to which research operations correspond to their referent terms in research propositions of the form: "Counseling for pregnant teenagers improves their mental health" or "The introduction of national health insurance causes an increase in the use of outpatient services." Without using our terminology, Snow notes that such propositions usually contain implicit or explicit references to populations, settings, and times (external validity), to the nature of the presumed cause and effect (construct validity), to whether the operations representing the cause and effect covary (statistical conclusion validity), and to whether this covariation is plausibly the result of causal forces (internal validity).

For a study to be useful, the argument goes, each part of the proposition should be given approximately equal weight. There is no need to stress the causality term over any other. Other critics (Hultsch and Hickey, 1978; Cronbach, in preparation) take the argument one step further and stress the primacy of external over internal validity. Hultsch claims that if we have a target population of special interest—for example, the educable mentally retarded—then it is better to test causal propositions about this group on representative samples. He maintains this should be done even if less rigorous means have to be used for testing causal propositions than would be the case if a study was restricted to easily available but nonrepresentative subgroups of the educable mentally retarded or to children who were not educable and retarded. Cronbach (in preparation) echoes this argument and adds, first, that in much applied social research the results are needed quickly and, second, that a high degree of confidence about causal attribution is less important in the decisions of policy-makers (broadly conceived) than is confidence in knowing that one is working with formally or impressionistically representative samples. Consequently, Cronbach contends that the time demands of experiments with experimenter-controlled manipulanda and the reality of how research is (and is not) used in decision making suggest a higher priority for speedy research using available data sets, simple one-wave measurement studies or qualitative studies as opposed to studies which, like quasi-experiments, take more time and explicitly stress internal validity.

It is in some ways ironic that the charge of neglecting external validity should be leveled against one of the persons who invented the construct and elevated its importance in the eyes of those who restricted experimentation to laboratory settings and who wrote about experimentation without formally mentioning the special problems that arise in field settings. But this aside, we have no quarrel *in the abstract* with the point of view that, where causal propositions include references to populations of persons and settings and to constructs about cause and effect, each should be equally weighted in empirical tests of these propositions. The real difficulty comes *in particular instances* of research design and implementation where very often the investigator is forced to make undesirable choices between internal and external validity. Gaining a representative sample of educable, mentally retarded students across the whole nation demands considerable resources.

Even gaining such a sample in a few cities located more closely together is difficult, requiring resources for implementing a treatment, ensuring its consistent delivery, collecting the required pretest and posttest data, and doing the necessary public relations work. Without such resources, one runs the risk of a large, poorly implemented study with a representative sample or of a smaller, better implemented study where the small sample size limits our confidence in generalizing.

Since random sampling is so rare for purposes of achieving representativeness, it is useful to consider the trade-off between internal and external validity when heterogeneous but unrepresentative sampling is used or when impressionistically modal but unrepresentative instances are selected that at least belong in the general class to which generalization is desired. Samples selected this way will have unknown initial biases, since not all schools will volunteer to permit measurement, even fewer schools will agree to deliberate manipulation of any kind, and the sample of schools that *will* agree to randomized manipulation will probably be even more circumscribed than the sample of schools that agrees to measurement with or without quasi-experimentation. The crucial issue is this: Would one do better to work with the initially more representative sample of schools in a particular geographical area that volunteered to permit measurement, even though no deliberate manipulation took place? Or would one rather work with a less representative sample of schools where both measurement and deliberate manipulation took place?

Solving this problem boils down, we think, to asking whether the internal validity costs of eschewing deliberate manipulation and more confident causal inferences are worth the gains for external validity of having an initially more representative sample from which bias-inducing attrition will nonetheless take place. Any resolution must also consider two other factors. First, the study which stresses internal validity has at least to take place in a setting and with persons who belong in the class to which generalization is desired, however formally unrepresentative of the class they might be. Second, the study which stresses external validity and has apparently more representative samples of settings and persons will result in less confident causal conclusions because more powerful techniques of field experimentation were not used or were not used as well as they might have been under other circumstances.

The art of designing causal studies is to minimize the need for trade-offs and to try to estimate in any particular instance the size of the gains and losses in internal and external validity that are involved in different trade-off options. Scholars differ considerably in their estimate of gains and losses. Cronbach maintains that timely, representative, but less rigorous studies can still lead to reasonable approximations to causal inference, even if the studies are nonexperimental and of the kind we shall discuss—somewhat pessimistically—in chapter 7. Campbell and Boruch (1975), on the other hand, maintain that causal inference is problematic with nonexperimental and single-wave quasi-experiments because of the many threats to internal validity that remain unexamined or have to be ruled out by fiat rather than through direct design or measurement. The issue involves estimating how to balance timeliness and the quality of causal inference, whether the costs of being wrong in one's causal inference are not greater than the costs of being late with the results.

Consider two cases of timely research aimed at answering causal questions which used manifestly inadequate experimental procedures. Head Start was evaluated by Ohio-Westinghouse (Cicirelli, 1969) in a design with only one wave of measurement of academic achievement. The conclusion—Head Start was harmful. Analysis of the same data using different statistical models appeared to corroborate this conclusion (Barnow, 1973); to reverse it completely, making Head Start appear helpful (Magidson, 1977); or to result in no-difference findings (Bentler and Woodward, 1978). Since we do not know the effects of Head Start, any timely decisions based on the data would have been premature and perhaps harmful. The second example worth citing is the Coleman Report (Coleman et al., 1966). In this large-scale, one-wave study it was concluded that black children gained more in achievement the higher the percentage of white children in their classes. This finding was used to justify school desegregation. However, better designed subsequent research has shown that if blacks gain at all because of desegregation (which is not clear), they gain much less than was originally claimed. It is important, we feel, not to underestimate the costs of producing timely results about cause, particularly its direction, which turn out to be wrong. Clearly, the chances of being wrong about cause are higher the more one deviates from an experimental model and conducts nonexperimental research using primitive one-wave quasi-experiments.

Since timeliness is important in policy research—though less so for basic researchers for whom this book is also intended—we shall devote some of the next chapter to quasi-experimental designs that do not require pretests and to ways in which archives can be used for rigorous and timely causal analysis. In the end, however, each investigator has to try to design research which maximizes all kinds of validity and, if he or she decides to place a primacy on internal validity, this cannot be allowed to trivialize the research.


We have not tried to place internal validity above other forms of validity. Rather, we wanted to outline the issues. In a sense, by writing a book about experimentation in field settings, we are assuming that readers already believe that internal validity is of great importance, for the *raison d'être* of experiments is to facilitate causal inference. Other forms of knowledge about the social world one more accurately or more efficiently gained through other means—e.g., surveys or participant observation. Our aim differs, therefore, from that of the last critics we discussed. They argue that experimentation is not necessary for causal inference or that it is harmful to the pursuit of knowledge which will be useful in policy formulation. We assume that readers believe causal inference is important and that experimentation is one of the most useful, if not *the* most useful, way of gaining knowledge about cause.

SOME OBJECTIONS TO OUR TENTATIVE PHILOSOPHY OF THE SOCIAL SCIENCES


Protests against “scientism” have been prominent in recent commentaries on the theory of conducting social science. Such protests focus on inappropriate and blind efforts to apply “the scientific method” to the social sciences. Critics argue that quantification, random assignment, control groups and the deliberate intrusion

of treatments—all techniques borrowed from the physical sciences—distort the context in which social research takes place. Their protest against scientism is often linked with the now-pervasive rejection of the logical positivist philosophy of science and is frequently accompanied by a greater emphasis on humanistic and qualitative research methods such as ethnography, participation observation, and ethno-methodology. Critics also point to the irreducibly judgmental and subjective components in all social science research and to the pretensions to scientific precision found in many current studies.

We agree with much of this criticism and have addressed the issue in our previous work (Campbell, 1966, 1974, 1975; Cook, 1974a; Cook and Cook, 1977; Cook and Gruder, 1978; Cook and Reichardt, in press). However, some of the critics of scientism (Guttentag, 1971, 1973; Weiss and Rein, 1970; Hultsch and Hickey, 1978; Mitroff and Bonoma, 1978; Mitroff and Kilman, 1978; Cronbach, in preparation) have cited Campbell and Stanley (1966) and Cook and Campbell (1976) as prime examples of the scientific norm to which they object. While the identification of our previous work with scientism oversimplifies and blurs the issues, we acknowledge that in this volume, as in the past, we advocate using the methods of experiments and quantitative science that are shared *in part* with the physical sciences. We cannot here comment extensively on these criticisms of our background assumptions, which go beyond criticisms of causation issues alone. But we can indicate in broad terms the approach we would take in responding to these objections.



First, we of course agree with the critics of logical positivism. The philosophy was wrong in describing how physical science achieved its degree of validity, which was not through descriptive best-fit theories and definitional operationalism. Although the error did not have much impact on the practice of physics, its effect on social science methods was disastrous. We join in the criticism of positivist social science when *positivist* is used in this technical sense rather than as a synonym for "science." We do not join critics when they advocate giving up the search for objective, intersubjectively verifiable knowledge. Instead we advocate substituting a critical-realist philosophy of science, which will help us understand the success of the physical sciences and guide our efforts to achieve a more valid social science. Critical realists (Mandelbaum, 1964) or "metaphysical realists" (Popper, 1972), "structural realists" (Maxwell, 1972), or "logical realists" (Northrop, 1959; Northrop and Livingston, 1964) are among the most vigorous modern critics of logical positivism. Critical realists particularly concerned with the social sciences identify their position with Marx's materialist criticism of idealism and positivism, e.g., Bhaskar, 1975, 1978; Keat and Urry, 1975.



Second, it is generally agreed that the social disciplines, pure or applied, are not truly successful as sciences. In fact, they may never have the predictive and explanatory power of the physical sciences—a pessimistic conclusion that merits serious debate (Herskovits, 1972; Campbell, 1972). This book, with its many categories of threats to validity and its general tone of modesty and caution in making causal inferences, supports such pessimism and underscores the equivocal nature of our conclusions. However, it is sometimes forgotten that these threats are not limited to quantitative or deliberately experimental studies. They also arise in *less formal, more commonsense, humanistic, global, contextual integrative and*

qualitative approaches to knowledge. Even the "regression artifacts," identified with measurement error, are an observational-inferential illusion that occurs in ordinary cognition (see Tversky and Kahnman, 1974, and Fischhoff, 1975).

We feel that those who advocate qualitative methods for social science research are at their best when they expose the blindness and gullibility of specious quantitative studies. Field experimentation should always include qualitative research to describe and illuminate the context and conditions under which research is conducted. These efforts often may uncover important site-specific threats to validity and contribute to valid explanations of experimental results in general and of perplexing or unexpected outcomes in particular. We also believe, along with many critics, that quantitative researchers in the past have used poorly framed questions to generate quantitative scores and that these scores have then been applied uncritically to a variety of situations. (Chapters 4 and 7, in particular, highlight some of the abuses associated with traditions of quantitative data analysis which have probably led to many specious findings.) In uncritical quantitative research, measurement has been viewed as an essential first step in the research process, whereas in physics the routine measures are the products of past crucial experiments and elegant theories, not the essential first steps. Also, the definitional operationalism of logical positivists has supported the uncritical reification of measures and has encouraged research practitioners to overlook the measures' inevitable shortcomings and the consequences of these shortcomings. A fundamental oversight of uncritical quantifiers has been to misinterpret quantifications as replacing rather than depending upon ordinary perception and judgment, even though quantification at its best goes beyond these factors (Campbell, 1966, 1974, 1975). Experimental and quantitative social scientists have often used tests of significance as though they were the sole and final proof of their conclusions. From our perspective, tests of significance render implausible only *one* of the many plausible threats to validity that are continually arising. Naïve social quantifiers continue to overlook the presumptive, qualitatively judgmental nature of *all* science. In contrast, the tradition we represent, with its heavy use of the word "plausible," stresses that the scientist must continually judge whether a given rival hypothesis will explain the data. Qualitative contextual information (as well as quantitative evidence on tangential variables) has long been recognized as relevant to such judgments.

Valid as these criticisms are, it is not enough merely to point out the limitations of our methods. Critics should be able to offer viable alternatives.* To be superior to the techniques described in the next six chapters, however, the proposed qualitative methods would have to eliminate more of the threats to validity listed in this chapter than do the quantitative methods. In this regard, it is refreshing to note that our humanistic colleague, Howard Becker (1978), has tried to rule out some of the validity threats in research which uses photographs either as evidence or as the means of presenting final research results, no quantification having intervened. Others conducting qualitative research under nonlaboratory conditions also recognize the equivocal nature of any causal inferences drawn from their observations. Many sociologists, anthropologists and historians have attempted to avoid causal explanations, aiming instead for uninterpreted description. Yet careful linguistic analysis of their reports shows that they are rarely successful. Their

understandings, insights, meanings, analysis of intentions and the like are strongly colored by causal conclusions even when the terms "effects," "gains," "benefits" and "results" are carefully avoided.

Most of the critics also write as if they were advocating an alternative social science approach that, when implemented, would remove the equivocality of inference plaguing our research efforts. Examples of such new approaches are rare and, when offered, are not impressive. The best of qualitative social science joins with us and much of the significant literature in recognizing and emphasizing the ambiguity of social science situations. Such an emphasis should help us relinquish the hope for a simple social science rather than try to provide a "superior route" to such a dubious goal.

The methodology presented in this book is a product of a critical dialectic that has continued over the past several decades. In this dialectic, social scientists have criticized each other's attempts to draw causal inferences; in answer to these criticisms, practitioners have collected additional data or conducted secondary analyses. We believe that the threats to validity and methodological arrangements we present in this book would arise from such a critical exchange even without the physical science paradigm. If the starting point for investigating an issue was qualitative humanistic scholarship, the results of the scholarship would, we believe, be criticized in terms of specific rival alternative interpretations like those we have defined. Answering these criticisms would lead, we also believe, to reinventing quantification, control groups, the arbitrary intrusions of experimental manipulations, and randomization, since recognizing threats would lead to a need for mechanisms that rule out such threats. Yet in most real world settings, even these ingenious devices will not be enough; and confident estimates of effects, gains, and the like must be tentative at best and subject to correction. We hope that our long list of threats will encourage such modesty and caution in drawing conclusions regardless of whether the approach is humanistic or scientific.

However, the list of threats to validity should not leave the readers with the belief that reasonably dependable causal knowledge cannot be achieved in the social sciences. Indeed, we hope that this book will help guide the search for and recognition of such dependable connections. The techniques presented in the following chapters should be of some assistance in ruling out as many threats to validity as possible and reducing the number of plausible alternative interpretations of the data. But we wish to stress that causal inferences will never be proved with certainty since the inferences we make depend upon many assumptions that cannot be directly verified. This predicament should not discourage us since it is not unique to social science and is shared in varying degrees by the successful physical sciences (Kuhn, 1961; Toulmin, 1972).