CHAPTER ONE

Research Design and Issues of Validity

MARILYNN B. BREWER

Validity refers to "the best available approximation to the truth or falsity of *propositions*" (Cook & Campbell, 1979, p. 37; italics added). In this sense, we cannot speak of the validity or invalidity of research per se. Rather, it is the statements, inferences, or conclusions we wish to draw from the results of empirical research that can be subject to validation. Of course, the way that a research study is designed and conducted has a great deal to do with the validity of the conclusions that can be drawn from the results, but validity must be evaluated in light of the *purposes* for which the research was undertaken in the first place.

RESEARCH PURPOSE AND TYPES OF VALIDITY

There are any number of ways in which the various objectives of research can be classified, but for present purposes the goals of empirical research in social psychology can be differentiated into three broad categories: demonstration, causation, and explanation.

Research undertaken for the purpose of *demonstration* is conducted in order to establish empirically the existence of a phenomenon or relationship. Much demonstration research is intended to be descriptive of the state of the world, including the frequency of occurrence of specified events across time or space (e.g., distribution of forms of cancer, variations in crime rates, probability of intervention in emergency situations, participation in collective demonstrations, etc.) and the assessment of the degree of relationship between specified states or conditions (e.g., the correlation between cigarette smoking and lung cancer, the relationship between ambient temperature and violent crime, the correlation between economic prosperity and collective protest, etc.). Although most descriptive research is conducted in field settings with the purpose of assessing phenomena as they occur "naturally," some demonstration studies are also undertaken in the controlled setting of the psychological laboratory. Studies of gender differences or personality types are often conducted in lab settings. Further, many of the classic studies in social psychological research – including Sherif's (1935) studies of formation of arbitrary group norms, Asch's (1956) original conformity studies, Milgram's (1963) study of obedience to authority, and Tajfel's (1970) initial studies of ingroup favoritism – were essentially demonstrations of social psychological phenomena in the laboratory.

Although establishing that the presence or absence of one event is correlated with the presence or absence of another is often of interest in its own right, most of the time scientists are interested in whether such covariation reflects a causal relationship between the two events. Thus, much research is undertaken not simply to demonstrate that a relationship exists but to establish a cause–effect linkage between specific variables (i.e., testing linkages of the form, if X then Y). For this purpose we are using the concept of causation in the utilitarian sense (Collingwood, 1940; Cook & Campbell, 1979; Gasking, 1955; Mackie, 1974). In this sense, the search for cause–effect relationships is for the purpose of identifying agents that can be controlled or manipulated in order to bring about changes in outcome. In other words, research on causation (cf. West, Biesanz, & Pitts, this volume, Ch. 3) is intended to demonstrate that interventions that produce change in one state of the world will produce subsequent changes in the outcome of interest. For this purpose, the goal of research is to establish causal connections, not to explain how or why they occur (Cook & Shadish, 1994).

When research has the purpose of establishing causal relationships in this sense, the purported causal factor is generally referred to as the "independent variable" and the outcome or effect as the "dependent variable." In fact, the use of these terms in describing a study is effectively a statement of purpose. However, there are important differences across types of research in the meaning of "independent variable" – differences that have to do with how variation in the purported causal variable is produced. When the state of the independent variable is manipulated by interventions under the control of the researcher, we have research that can be defined as an experiment or "quasi-experiment" (Campbell & Stanley, 1963, 1966). In correlational field studies, by contrast, the so-called "independent variable" is not manipulated or controlled but instead variations are assessed as they occur naturally for purposes of establishing the relationship between such variations and subsequent variations in the outcome variable of interest. In such cases, causal inference is usually predicated on temporal precedence, establishing that variations in the purported cause precede variations in the purported effect. Such temporal precedence is a necessary but not sufficient basis for inferring causation. In studies of this type, the independent variable(s) might better be labeled "predictor" variables. As we shall see, the validity of causal inferences can be significantly influenced by differences in how the independent variable is defined and varied.

As a goal of research, utilitarian causation is sufficient for most applied and action research purposes. Knowing that a reliable cause–effect relationship between X and Y exists is a critical step in designing interventions that can bring about desired changes in the outcome, Y. For utilitarian purposes, what "works" is what counts, irrespective of why it works. For basic, theory-testing research purposes, however, knowing that a cause–effect relationship exists is not sufficient. The purpose of this type of research is *explanation*, or establishing the intervening processes that mediate the linkage between variations in X and Y. This reflects the "essentialist" conceptualization of causation to which most scholars now subscribe (Cook & Campbell, 1979). Research undertaken for the purpose of explanation has the goal of determining not only whether causation exists but why and under what conditions.

Although there are many legitimate questions about validity that can be raised in connection with conclusions drawn from demonstration research, the fact is that most of the controversies that arise over validity issues in the social psychological literature revolve around inferences about causation and explanation. It was specifically in connection with research intended to establish cause–effect relationships that Campbell introduced the now-classic distinction between internal validity and external validity (Campbell, 1957; Campbell & Stanley, 1963, 1966).

Internal validity, in Campbell's terms, refers to the truth value that can be assigned to the conclusion that a cause–effect relationship between an independent variable and a dependent variable has been established within the context of the particular research setting. The question here is whether changes in the dependent measure were produced by variations in the independent variable (manipulation, in the case of an experiment) in the sense that the change would not have occurred without that variation. *External validity,* in Campbell's original terminology, referred to the generalizability of the causal finding, that is, whether it can be concluded that the same cause–effect relationship would be obtained across different subjects, settings, and methods.

In a later elaboration of validity theory, Cook and Campbell (1979) differentiated the concept of external validity further. The term *construct validity* was introduced to refer to the extent to which a causal relationship could be generalized from the particular methods and operations of a specific study to the theoretical constructs and processes they were meant to represent. The term external validity was reserved to refer to the generalizability of findings to target populations of persons and settings. It is this tripartite distinction – internal, external, and construct validity – that will provide the basis for organizing the discussion of validity issues in the remainder of this chapter.

INTERNAL VALIDITY: THE THIRD VARIABLE PROBLEM

As stated above, the essence of internal validity is establishing that variation in an effect (dependent variable) has been produced by changes in level or intensity of the independent variable and not by some other causal force (or forces).[1] In notational form, we are interested in the proposition:

$$X \rightarrow Y.$$

Threats to the validity of this proposition come from any plausible claim that the obtained variations in the outcome variable (Y) were actually produced by some third factor which happened to be correlated with the

[1] This does not mean that the independent variable under investigation is assumed to be the only cause of the outcome, but rather that this variable has a causal influence independent of any other causal forces.

variations in level of X. Again in notational terms, the alternative proposition is

$$\begin{tikzcd} C \arrow[r] \arrow[d, leftrightarrow] & Y \\ X. & \end{tikzcd}$$

In this case, the relationship between X and C (the "hidden" third factor) is not a causal one. However, because X and C are correlated, causes of the variation in Y could be misattributed to X when they were actually produced by C. This pattern is referred to as a *spurious* correlation between X and Y.

It is this third-variable causation pattern that is, in part, responsible for the well-known dictum that "correlation does not prove causation." Two variables can be correlated with each other because both are correlates of a third factor, even when there is no direct or indirect causal relationship between the first two. Consider, for example, the results of a hypothetical study of weather conditions and psychological mood. Let's say that the researcher finds that the incidence of depression and suicide is greater on rainy days than on sunny days. Before we could conclude from this relationship that rain causes depression we would have to take into account the fact that there are a number of other weather-related factors that are correlated with the presence or absence of rain – including atmospheric pressure and gray skies – any of which could plausibly be the true causal factor. In this case, the conclusion that rain causes depressed mood would have low internal validity in that we cannot assign it a high truth value with any confidence.

In social psychological research, many potentially problematic "third variables" are those associated with self-selection. If a researcher is concerned with the effects of some environmental variable or treatment (such as interracial contact or presentation of a persuasive message), causal inference is undermined if exposure to different levels of the treatment variable is correlated with differences among people in personality or aptitudes. If persons choose for themselves what experiences they will be exposed to, there may be a relationship between the experience (treatment) and the outcome variable (e.g., persons who engage in intergroup contact are less prejudiced; individuals who listen to a Democratic campaign speech are more likely to vote Democratic). However, we cannot tell whether the outcome was influenced by the treatment or whether it would have occurred because of the correlated individual differences even in the absence of the treatment experience.

Hidden causes are not the only way that unintended third variables can influence the validity of cause–effect inferences. Sometimes causal relationships can be either augmented or blocked by the presence or absence of factors that serve as *moderator variables* (Baron & Kenny, 1986). To take another weather-related illustration, consider the causal relationship between exposure to sun and sunburn. Although there is a well-established cause–effect link here, it can be moderated by a number of factors. For instance, the relationship is much stronger for fair-skinned individuals than for dark-skinned persons. Thus, fair skin is a moderator variable that enhances the causal relationship between sun exposure and burning. However, this does not mean that the sun–sunburn relationship is spurious. The moderator variable (skin pigmentation) does not cause the effect in the absence of the independent variable (sun exposure).

Other moderator variables can reduce or block a causal sequence. For instance, the use of effective suntan lotions literally "blocks" (or at least retards) the causal link between the sun's ultraviolet rays and burning. Thus, a researcher who assesses the correlation between sun exposure and sunburn among a sample of fair-skinned people who never venture outdoors without a thick coat of 30 SPF sunblock would be ill-advised to conclude that the absence of correlation implied the absence of causation.

Like Baron and Kenny (1986), I think it is important here to distinguish between third variables that serve as *moderators* (as the illustration above) and those that serve as *mediators* of a cause–effect relationship (cf. Judd, this volume, Ch. 14). Moderator relationships can be represented notationally as follows:

$$\begin{tikzcd} & C \arrow[d] & \\ X \arrow[r, dash] & {} \arrow[r] & Y. \end{tikzcd}$$

The causal link is actually between X and Y, but the observed relationship between these two variables is qualified by levels of variable C, which either enhances or blocks the causal process.

A mediational relation, on the other hand, is represented as follows:

$$X \rightarrow C \rightarrow Y.$$

In this case, the presence of C is necessary to complete the causal process that links X and Y. In effect, varying X causes variations in C, which in turn causes changes in Y. To return to our weather example, the effect of rain on depression may be mediated by social factors. Rain causes people to stay indoors or to hide behind big umbrellas, hence reducing social contact. Social isolation may, in turn, produce depression. However, rain may not be the only cause of social isolation. In this case, rain as an independent variable is a sufficient, but not necessary, cause in its link to depression. To demonstrate that X causes Y only if C occurs does not invalidate the claim that X and Y have a causal relationship; it only explicates the causal chain involved.

In order to establish unequivocally the causal relationship between two variables, variation in the causal factor has to be produced or observed under conditions that are isolated from third factors that may produce a spurious correlation. These third variables must be either held constant or uncorrelated with variations in X. This is the essence of the logic of good experimental design. In addition to control over variation in the independent variable, random assignment of subjects to different levels of the manipulated factor serves to rule out many potential third-variable threats to causal inference. Without random assignment, manipulating the independent variable is not sufficient to achieve the internal validity of a true experiment. This does not mean that correlational or quasi-experimental studies in field settings can never lead to justified causal inferences. However, many potential threats to the internal validity of such inferences have to be ruled out one by one, whereas random assignment rules out whole classes of potential spurious causal variables in one operation (cf. West et al., this volume, Ch. 3).

By way of caveat, it should be noted that even true experimental design does not always guarantee internal validity. Ill-conceived or poorly executed experimental studies can introduce many potential third-variable artifacts that undermine causal inference. A fuller treatment of the relationship between the design and conduct of laboratory experiments and internal validity is provided in the following chapter (Smith, this volume, Ch. 2).

FROM CONSTRUCT TO OPERATION AND BACK AGAIN

For many applied research purposes it is sufficient to know that a specific intervention (e.g., passage of a particular gun control law) produces a specific outcome (e.g., reduction in violent crime). Most social psychological research, however, is inspired not by such specific action-oriented questions but by general theories about the interrelationships among cognition, affect, and social behavior. Theories are stated in terms of abstract concepts and hypothetical constructs that cannot be directly observed or measured. In order to be subject to empirical testing, theoretical constructs must be "translated" from the abstract to the more concrete, from concepts to operations that can be observed and replicated.

Most social psychological researchers accept the philosophy that the specific operations and measures employed in a given research study are only partial "representations" of the theoretical constructs of interest – and imperfect representations at that. Hence, the conduct of theory-testing research has a cyclical nature of the form

$$\text{Construct}_1 - \text{Operations} - \text{Construct}_2,$$

where the first link refers to the stage of translating initial theoretical concepts into empirically testable hypotheses and specific operations, and the second link refers to the process of inference from empirical results back to the level of theoretical concepts, which then become the starting point for the next cycle of research.[2] Construct validity refers to inferences made at both stages of research linking concepts to operations. At the front end, we can ask how valid are the specific operations and measures used in a research project as representations or manifestations of the theoretical constructs to be tested; in other words, how good is the logic of translation from concept to operation? At the last stage, inference goes in the other direction (from empirical operations to hypothetical constructs and processes), and the question becomes how justified is the researcher in drawing conclusions from the concrete findings to the level of theory.

The validity issues that mark the initial, operationalization stage of research have been represented as in Figure 1.1 by Rakover (1981). LH-1 refers to the inferential link between the operational definition of the independent variable in an experiment and the corresponding causal concept at the theoretical level. LH-2 refers to the analogous link between the hypothetical effect and the actual response measure assessed in an experiment. (To this system I would add an LH-3 referring to the linkage between direct and indirect assessment of mediating variables and the hypothetical mediational processes, because measures of process are now common in social psychological research.)

Rakover claimed that both links are problematic in social psychological research because there are no standardized operations that correspond closely to our

[2] Because abstract definitions and theory are rarely unaffected by the process and outcomes of empirical research, we assume here that Construct$_1$ and Construct$_2$ are not necessarily conceptually equivalent.

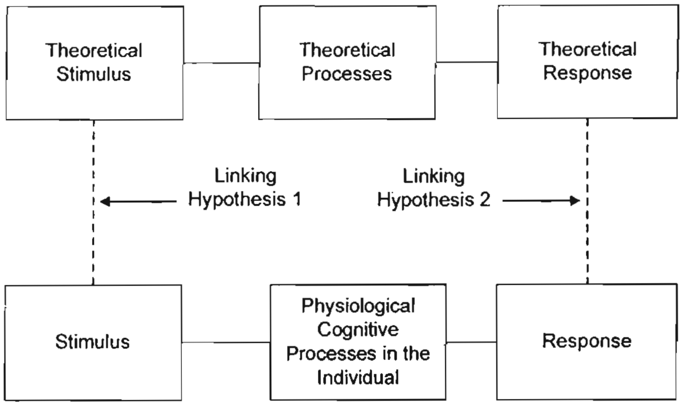

Figure 1.1. Constructs and Operations. From Rakover (1981). Social psychology theory and falsification. *Personality and Social Psychology Bulletin, 7,* p. 125. Copyright SAGE Publications, Inc. and Society for Personality and Social Psychology. Reprinted with permission.

hypothetical constructs, and the inferential steps between concept and operation are often quite remote. The LH-1 and LH-2 links represent little more than "intuited causal relationships" (Rakover, 1981, p. 125). More specifically he identified four major difficulties in connecting theory and data: the stimulus and response validity problems and the "unknown range of stimulus variation" and "unknown range of measurement" problems. The stimulus and response validity problems are the standard construct validity questions of whether the stimulus variations and response measures of the empirical research actually reflect variation in the corresponding theoretical states (cf. John & Benet-Martinez, this volume, Ch. 13). The unknown range problems refer to the failure to specify precisely what levels of the independent variable are expected to be causally significant, and over what range of outcomes. Because of these problems, it is difficult to determine whether a failure to confirm a predicted causal or explanatory relationship represents a failure of theory or a failure of operation. The hypothetized relationship could be true at the conceptual level but go undemonstrated because operations were unrepresentative or failed to capture the effective range within which the causal process operates.

Causes and Confounds

Criticisms of construct validity often revolve around the meaning of the independent variable as operationalized (the LH-1 link in Rakover's, 1981, model). Even when the causal efficacy of the independent variable is not in question, there can be questions about the conceptual causal process that is actually operating to produce the observed effect.

In any research study, the operations that are meant to represent a particular causal construct can be construed in multiple ways. Any particular operation (manipulation of the independent variable) may be associated with variation in more than one hypothetical state, any one of which may be the "true" causal variable. This is what experimentalists are often referring to when they talk about "confounding" the independent variable; something about the independent variable is causing the outcome of interest, but it is not clear what.

For instance, a researcher may be interested in the effects of social isolation on susceptibility to influence. An independent variable is designed to produce variations in feelings of social isolation (e.g., waiting in a room with others present, waiting alone for a short time, waiting alone for an extended period of time), but these experimental conditions may also be producing variations in other subjective states, such as fear of the unknown or cognitive rumination. Any causal effects of this "treatment" may be attributable to social deprivation (as intended by the researcher), but could also be due to these other factors that are confounded with isolation in this particular operation.

This type of confounding should be distinguished from threats to internal validity because they are inherent in the independent variable itself. The causal effect (in the utilitarian sense) of the treatment is not threatened in this case, but the validity of the explanation of the effect is in question. Internal validity is threatened when the independent variable covaries with other variables that are correlated with but separate from (or extraneous to) the treatment itself. Self-selection, for example, undermines internal validity because individual personality differences have effects that are independent of any effects associated with variations in the intended independent variable. Construct confounds, on the other hand, are causal factors that are intrinsic rather than extrinsic to the independent variable as operationalized.

Potential threats to internal validity can be evaluated or ruled out by examining whether the variations in the independent variable are inadvertently correlated with variations in extraneous variables. Threats to construct validity cannot be so readily disentangled. Nonetheless, there are ways of planning and designing research operations so that the number of potentially confounding factors associated with the independent variable can be reduced.

Many potential confounds arise from the general "reactivity" of social psychological research (Cook & Campbell, 1979) that exists because such research involves social interaction, and subjects are usually aware that they are participants in a research study. Reactivity effects include "demand characteristics" (Orne, 1962), "experimenter expectancies" (Rosenthal, 1966), and "evaluation apprehension" (Rosenberg, 1969). All of these effects derive from the fact that alert, aware participants are actively seeking cues in the research setting to inform them of what they are expected to do or what they should do in order to present themselves in a favorable light. Different levels of the independent variable may contain different cues that influence participants' guesses about what the research study is really about or what constitutes a proper response. When experimental treatments and demand characteristics are confounded in this way, the construct validity of the independent variable is compromised.

We can use the concept of demand characteristics to illustrate the difference between methodological *confounds* (which affect construct validity) and methodological *artifacts* (which are threats to internal validity). Demand characteristics confound the conceptual interpretation of the causal effect of an independent variable when the cues are inherent in the experimental manipulations themselves. To take an example from classic dissonance research, the amount of money participants are offered to write a counterattitudinal essay is intended to manipulate the presence of high or low external justification for engaging in an attitude-discrepant behavior. However, offering a participant as much as $20 for the favor requested by the experimenter may also carry extraneous cues to the participant that convey the idea that the requested behavior must be either unpleasant or immoral to be worth paying so much. In this case, the "message" is implicit in the independent variable itself; $20 carries a different cue or message than an offer of $5 or $1. As a consequence, when participants show less attitude change under the high payment condition than under the low payment condition, we cannot be sure whether this is due to the external justification provided by the money offered (the theoretical construct of interest) or to the demand characteristic inherent in the manipulation itself.

Contrast the above example with another case illustration in which demand characteristics are created by experimenter expectancy effects. Because the researcher may be biased or predisposed to elicit different responses in different experimental conditions, he or she may deliver the experimental instructions in ways that vary systematically across treatment conditions. For instance, the $20 offer may be delivered in a different tone of voice or with different nonverbal cues than the $5 condition. Such experimental behaviors are extraneous to the independent variable itself, but if they are correlated with the differences in instructional conditions, they are procedural artifacts that threaten the internal validity of any causal interpretations of the effects of the independent variable. This is an illustration of how poor procedural controls can undermine the internal validity of even a true experiment with random assignment to treatment conditions.

Construct Validity and Conceptual Replications

Apart from methodological confounds, research operations are subject to multiple theoretical interpretations. Many interesting controversies in the social psychological literature have been fueled by disagreements over the correct theoretical interpretation of a particular phenomenon. Such debates require conceptual replication of the phenomenon, using different operations that are intended to represent the same causal construct.

Consider, for example, the classic study by Aronson and Mills (1959), in which cognitive dissonance was induced by having female participants read aloud some embarassing, obscene passages in the guise of an "initiation" test for admission to a discussion group. The intended conceptual independent variable here was a state of dissonance associated with inconsistency between the participant's behavior (going through high embarassment in order to join the group) and any negative perceptions of the group. But when participants recite a list of obscene words and then listen to a boring group discussion, one cannot be sure that this represents an empirical realization of the intended conceptual variable and nothing else. The complex social situation used by Aronson and Mills has many potential interpretations, including the possibility that reading obscene materials generated a state of sexual arousal that carried over to reactions to the group discussion. If that were the case, it could be that transfer of arousal, rather than dissonance accounted for attraction to the group.

A conceptual replication of the initiation experiment by Gerard and Mathewson (1966) was undertaken to rule out this alternative interpretation. Their experiment was constructed so as to differ from the Aronson and Mills (1959) study in many respects. For example, Gerard and Mathewson used electric shocks instead of the reading of obscene words as their empirical realization of severe initiation (and the dissonance it produced), the shocks were justified as a test of "emotionality" rather than as a test of embarrassment, and the group discussion that participants listened to was about cheating rather than sex. Thus sexual arousal was eliminated as a concomitant of the experimental operationalization of the independent variable. The results confirmed the original findings: People who underwent painful electric shocks in order to become members of a dull group found that group to be more attractive than did people who underwent mild shocks. Such a confirmation of the basic initiation effect under quite different experimental operations supported the contention that it was cognitive dissonance produced by a severe initiation, and not some other conceptual variable, that was responsible for the results in the original experiment. A considerable amount of research in social psychology has been motivated by similar controversies over the valid interpretation of results obtained with complex experimental procedures. Designing conceptual replications to assess threats to constuct validity of the causal variable is both challenging and important to the theoretical development of our field.

Multiple Operations: Convergent and Discriminant Validity

These early dissonance experiments illustrate a general principle underlying the idea of construct validity as originally defined by Cook and Campbell (1979). According to Cook and Campbell, the most serious threat to construct validity of any program of research comes from a "mono-operation bias," that is, the tendency to use only a single operation or measure to represent a particular theoretical construct. Because any one operation invariably underrepresents the construct of interest, and embodies potentially irrelevant constructs as well, the conceptual interpretation of single operations can always be challenged. It takes conceptual replication across multiple different operationalizations of the same construct to establish construct validity.

Ideally, multiple operations will allow for testing both convergent and discriminant validity of the construct being studied (Cook & Campbell, 1979). *Convergent* validity is established when different operations representing the same underlying theoretical construct produce essentially the same results (as in the Gerard and Mathewson, 1966, experiment described previously). Equally important, however, is establishing that operations that represent the construct of interest show the predicted effect, whereas other operations which do not reflect the theoretical construct do not have similar effects. If Gerard and Mathewson had demonstrated that dissonance aroused by tolerating electric shock had produced attraction to the discussion group, whereas sexual arousal alone (without dissonance) did not produce attraction, they would have gone further in establishing that the dissonance explanation had discriminant validity. That is, dissonance arousal would have been demonstrated to produce effects that differentiate it from other types of arousal.

Measures of dependent variables can also be subjected to the tests of convergent and discriminant validity. In order to establish measurement construct validity, it is necessary to demonstrate that a particular measure correlates positively with different ways of measuring the same construct and does not correlate as strongly with other measures that use similar methods but are intended to assess a different construct. This is the logic behind the use of the "multitrait multimethod matrix" to establish construct validity of psychological instruments (Campbell & Fiske, 1959; John & Benet-Martinez, this volume, Ch. 13). The multitrait–multimethod procedure involves measuring more than one theoretical construct using more than one method for each construct. If a measure has construct validity, two different measures of the same trait should be more highly related than two different traits assessed by the same method. At a broader level, this logic can be generalized to testing the theoretical framework within which a construct is embedded. Theoretical validity is established when measured constructs prove to be related to theoretically relevant variables and not to theoretically irrelevant ones. Ultimately, then, construct validity is equivalent to theoretical validity.

Causal Processes and Mediational Analyses

Some theoretical debates do not revolve around conceptual interpretation of the operations themselves but are about the intervening processes that mediate the link between the causal variable and its effects. To return to Figure 1.1, these debates over theoretical processes cannot be resolved by examining the construct validity of the independent variable (LH-1) or the dependent variable (LH-2) alone. Theoretical controversies at this level require operations that tap into the intervening physiological, cognitive, and affective processes themselves.

The long-standing debate between alternative explanations of the counterattitudinal advocacy effect derived from dissonance theory and self-perception theory provides a case in point (Greenwald, 1975). In this controversy, the validity of the basic empirical finding and the research operations were not in doubt. Theorists on both sides acknowledged that a causal relationship exists between the presence or absence of external incentives (e.g., monetary payment) and the resulting consistency between behavior and expressed attitudes. What was at issue was the nature of the mediating processes that underlie the relationship between induced behaviors and subsequent attitudes. Self-perceptions theorists held that the effect was mediated by cognitive, self-attribution processes, whereas the dissonance theory explanation rested on motivational processes. As in many cases in social psychological research, the efforts to establish construct validity helped to refine and clarify the theory itself.

Years of attempts to resolve the debate through "critical experiments" were of no avail (Greenwald, 1975). In each case, the same experimental operations could be interpreted as consistent with either theoretical construct. It wasn't until a clever experiment was designed to assess directly the mediating role of motivational arousal that the deadlock was effectively broken. Zanna and Cooper (1974) used a mediational design to demonstrate that the presence of arousal was necessary to produce the attitude change effect and that when the motivational effects of arousal were blocked (through misattribution), attitude change following counterattitudinal behavior did not occur. These findings regarding process were more consistent with the dissonance interpretation of the phenomenon than with the self-perception interpretation, although they established that both motivational and cognitive processes were essential mediating factors.

THE MANY FACES OF EXTERNAL VALIDITY

Construct validity represents one form of generalizing from the observed results of an empirical study to conclusions that go beyond the results themselves. Another form of generalizability has to do with the empirical replicability of the phenomenon under study. External validity refers to the question of whether an effect (and its underlying processes) that has been demonstrated in one research setting would be obtained in other settings, with different research participants and different research procedures.

Actually, external validity is not a single construct but represents a whole set of questions about generalizability, each with somewhat different implications for the interpretation and extension of research findings. The sections that follow discuss three of the most important forms of external validity – robustness, ecological validity, and relevance. Each of these raises somewhat different questions about where, when, and to whom the results of a particular research study can be generalized.

Robustness: Can it Be Replicated?

The robustness issue refers to whether a particular finding is replicable across a variety of settings, persons, and historical contexts. In its most narrow sense, the question is whether an effect obtained in one laboratory can be exactly replicated in another laboratory with different researchers. More broadly, the question is whether the general effect holds up in the face of wide variations in subject populations and settings. Some findings appear to be very fragile, obtainable only under highly controlled conditions in a specific context; other findings prove to hold up despite significant variations in conditions under which they are tested.

Technically, robustness would be demonstrated if a particular research study were conducted with a random sample of participants from a broadly defined population in a random sampling of settings. This approach to external validity implies that the researcher must have theoretically defined the populations and settings to which the effect of interest is to be generalized and then must develop a complete listing of the populations and settings from which a sample is drawn. Such designs, however, are usually impractical and not cost-effective. More often, this form of generalizability is established by repeated replications in systematically sampled settings and types of research participants. For instance, a finding initially demonstrated in a social psychology laboratory with college students from an eastern college in the United States may later be replicated with high school students in the Midwest and among members of a community organization in New England. Such replication strategies are not only more practical but they also have potential advantages for theory-testing purposes. If findings do not replicate in systematically selected cases, we sometimes gain clues as to what factors may be important moderators of the effect in question (Petty & Cacioppo, 1996).

Generalizability across multiple populations and settings should be distinguished from generalizability to a particular population. A phenomenon that is robust in the sense that it holds up for the population at large may not be obtained for a specific subpopulation or in a particular context. If the question of generalizability is specific to a particular target population (say, from college students to the elderly), then replication must be undertaken within that population and not through random sampling.

Generalizability from one subject population or research setting to others is probably the most frequently raised issue of external validity for experimental studies conducted in laboratory settings. Sears (1986) provided what is probably the most cogent arguments about the limitations of laboratory experimentation with college student participants. A review of research articles published in the major social psychology journals in 1985 revealed that 74% were conducted with undergraduate student participants, and 78% were conducted in a laboratory setting. According to Sears, this restriction of populations and settings means that social psychology has a "narrow data base" on which to draw conclusions about human nature and social behavior.

It is important to point out here that Sears (1986) was not claiming that college students or psychology laboratories are any less generalizable to the world at large than any other specific type of persons or settings. Just because an effect has been demonstrated in a particular field setting rather than a lab does not automatically render it more externally valid. What Sears was criticizing is the overrepresentation of a specific type of subject and setting across a large number of studies, all of which then share the same limitations on external validity.

Before we can conclude, however, that the oversampling of college student participants actually limits the external validity of our findings and interpretations, we have to specify in what ways undergraduate students differ from other populations and how these differences might alter the effects we observe. Drawing on research in cognitive and social development, Sears (1986) suggested that there are several distinguishing characteristics of college students that may be relevant to social psychological findings. Compared with the general population, undergraduates are likely to have stronger cognitive skills, less well-formulated or crystallized attitudes and self-concepts, and less stable group identities – differences that are likely to be exacerbated when studies are conducted in academic laboratories with academic-like tasks. Do these differences make a difference? Sears contended that because of these characteristics of our subject population and setting, results of our research may exaggerate the magnitude of effects of situational influences and cognitive processes on social attitudes and behavior.

To argue that characteristics of the setting or subject population qualify the conclusions that can be drawn about cause–effect relationships is, in effect, to hypothesize that the cause interacts with (i.e., is moderated by) the characteristics of the population or context to produce the effect in question. To translate Sears's (1986) arguments into these terms, he is postulating that manipulations of the type of influence used interact with participant characteristics to determine amount of attitude change. For instance, the magnitude of the effect of influence attempts that rely on cognitive elaboration would be expected to differ depending on whether the effect is tested with college students or with older, nonstudent populations. In this case, age is expected to moderate the causal effect of treatments that require cognitive elaboration.

External validity is related to settings as well as to participant populations. The external validity of a finding is challenged if the relationship between independent and dependent variables is altered when essentially the same research procedures are conducted in a different laboratory or field setting or under the influence of different experimenter characteristics. For example, Milgram's (1963) initial studies of obedience were conducted in a research laboratory at Yale University, but used participants recruited from the community of New Haven. Even though these experiments were conducted with a nonstudent sample, a legitimate question is the extent to which his findings would generalize to other settings. Because participants were drawn from outside the university and because many had no previous experience with college, the prestige and respect associated with a research laboratory at Yale may have made the participants more susceptible to the demands for compliance that the experiment entailed than they would have been in other settings.

To address this issue, Milgram undertook a replication of his experiment in a very different physical setting. By moving the research operation to a "seedy" office in the industrial town of Bridgeport, Connecticut and adopting a fictitious identity as a psychological research firm, Milgram hoped to minimize the reputational factors inherent in the Yale setting. In comparison with data obtained in the original study, the Bridgeport replication resulted in slightly lower but still dramatic rates of compliance to the experimenter. Thus, setting could be identified as a contributing but not crucial factor to the basic findings of the research.

Cook and Campbell (1979) made it clear that questions of external validity, or generalizability, are implicitly questions about interactions between the independent variable (treatment) and contextual variables such as subject selection, history, and research setting. In other words, the quest for external validity is essentially a search for moderators that limit or qualify the cause–effect relationship under investigation. As the Milgram experiments illustrate, once one has identified what the potential moderators are, the robustness of an effect can be tested empirically by varying those factors systematically and determining whether the effect is or is not altered.

Ecological Validity: Is It Representative?

The question of whether an effect holds up across a wide variety of people or settings is somewhat different than asking whether the effect is representative of what happens in everyday life. This is the essence of *ecological validity* – whether an effect has been demonstrated to occur under conditions that are typical for the population at large. The concept of ecological validity derives from Brunswik's (1956) advocacy of "representative design," in which research is conducted with probabilistic samplings of subjects and situations.

Representativeness is not the same as robustness. Generalizability in the robustness sense asks whether an effect can occur across different settings and people; ecological validity asks whether it does occur in the world as is. In the Brunswikian sense, findings obtained with atypical populations (e.g., college students) in atypical settings (e.g., the laboratory) never have ecological validity until they are demonstrated to occur naturally in more representative circumstances.

Many researchers (e.g., Berkowitz & Donnerstein, 1982; Mook, 1983; Petty & Cacioppo, 1996) take issue with the idea that the purpose of most research is to demonstrate that events actually do occur in a particular population. Testing a causal hypothesis requires demonstrating only that manipulating the cause can alter the effect. Even most applied researchers are more interested in questions of what interventions could change outcomes rather than what does happen under existing conditions. Thus, for most social psychologists, ecological validity is too restrictive a conceptualization of generalizability for research that is designed to test causal hypotheses. Ecological validity is, however, crucial for research that is undertaken for descriptive or demonstration purposes.

Further, the setting in which a causal principle is demonstrated does not necessarily have to physically resemble the settings in which that principle operates in real life for the demonstration to be valid. As Aronson, Wilson, and Brewer (1998) put it, most social psychology researchers are aiming for "psychological realism," rather than "mundane realism," in their experiments. *Mundane realism* refers to the extent to which the research setting and operations resemble events in normal, everyday life. *Psychological realism* is the extent to which the psychological processes that occur in an experiment are the same as psychological processes that occur in everyday life. An experimental setting may have little mundane realism but still capture processes that are highly representative of those that underlie events in the real world.

Relevance: Does It Matter?

In a sense, the question of ecological validity is also a question of relevance – is the finding related to events or phenomena that actually occur in the real world? However, relevance also has a broader meaning of whether findings are potentially useful or applicable to solving problems or improving quality of life. Again, relevance in this latter sense does not necessarily depend on the physical resemblance between the research setting in which an effect is demonstrated and the setting in which it is ultimately applied. Perceptual research on eye–hand coordination conducted in tightly controlled, artificial laboratory settings has proved valuable to the design of instrument panels in airplanes even though the laboratory didn't look anything like a cockpit.

Relevance is the ultimate form of generalization, and differences among research studies in attention to relevance is primarily a matter of degree rather than of kind. All social psychological research is motivated ultimately by a desire to understand real and meaningful social behavior. But the connections between basic research findings and application are often indirect and cumulative rather than immediate. Relevance is a matter of social process, that is, the process of how research results are transmitted and used rather than what the research results are (Brewer, 1997).

Is External Validity Important?

External validity, like other validity issues, must be evaluated with respect to the purposes for which research is being conducted. When the research agenda is essentially descriptive, ecological validity may be essential. When the purpose is utilitarian, robustness of an effect is particularly critical. The fragility and nongeneralizability of a finding may be a fatal flaw if one's goal is to design an intervention to solve some applied problem. On the other hand, it may not be so critical if the purpose of the research is testing explanatory theory, in which case construct validity is more important than other forms of external validity.

In the field of physics, for example, many phenomena can only be demonstrated empirically in a vacuum or with the aid of supercolliders. Nonetheless, the findings from these methods are often considered extremely important for understanding basic principles and ultimate application of the science. Mook (1983) argued compellingly that the importance of external validity has been exaggerated in the psychological sciences. Most experimental research, he contended, is not intended to generalize directly from the artificial setting of the laboratory to "real life," but to test predictions based on theory. He drew an important distinction between "generality of findings" and "generality of conclusions" and held that the latter purpose does not require that the conditions of testing resemble those of real life. It is the understanding of the processes themselves, not the specific findings, which has external validity.

In effect, Mook (1983) argued that construct validity is more important than other forms of external validity when we are conducting theory-testing research. Nonetheless, the need for conceptual replication to establish construct validity requires a degree of robustness across research operations and settings that is very similar to the requirements for establishing external validity. The kind of systematic, programmatic research that accompanies the search for external validity inevitably contributes to the refinement and elaboration of theory as well.

OPTIMIZING TYPES OF VALIDITY

Among research methodologists, controversies have ensued about the relative importance of different validity concerns, ever since Campbell and Stanley (1963) took the position that internal validity is the sine qua non of experimental research and takes precedence over questions of external validity (e.g., Cook & Shadish, 1994; Cronbach, 1982). The debate includes discussions of whether there are necessary trade-offs among the various aspects of validity or whether it is possible to demand that research maximize internal, external, and construct validity simultaneously.

It is possible to conduct a single research study with the goal of maximizing internal validity. Questions of external validity and construct validity, however, can rarely be addressed within the context of a single research design and require systematic, programmatic research studies that address a particular question across different operations and research settings. Thus, it is patently unfair to expect that any particular piece of research have high internal, external, and construct validity all at the same time. It is more appropriate to require that programs of research be designed in a way that addresses all types of validity issues (see also Smith, this volume, Ch. 2).

In order to design such a research program it is important to recognize the ways in which efforts to maximize one form of validity may reduce or jeopardize other types, hence the need for a diversity of methods as represented throughout this volume. By understanding such trade-offs, we can plan research projects in which the strengths and weaknesses of different studies are complementary. Thus, this chapter will close with a brief discussion of some of these important complementarities.

Setting: Lab Versus Field

It is a common assumption that laboratory research achieves high internal validity at the expense of external validity, whereas research conducted in natural field settings is associated with greater external validity albeit at the cost of more threats to internal validity. There is some basis for this implied association between research setting and types of validity. The laboratory does often permit a degree of control of the causal variable that maximizes internal validity to an extent that is difficult to achieve in "noisy" real-world contexts. And to the extent that natural settings reduce the reactivity that is characteristic of laboratory-based research, one threat to external validity is reduced.

It is hoped, however, that the earlier discussion of the different types of validity has made it clear that there is no invariable association between the setting in which research is conducted and its degree of internal, external, or construct validity. Tightly controlled experimental interventions can be introduced in field settings (and, conversely, laboratory studies can be poorly controlled). Most important, conducting research in a naturalistic context does not by itself confer external validity. Any specific context has limited generalizability. Even if the setting has been chosen to be highly representative or typical of naturally occurring situations, ecological validity is suspect if the research introduces conditions into that setting that do not occur spontaneously.

Establishing either construct validity or external validity requires that the conclusions drawn from research hold up across variation in context. Thus, it is the complementarity of field and lab as research settings that contributes to validity, not the characteristics of either setting alone. One good illustration of the use of selected field sites in conjunction with laboratory research comes from the literature on mood and altruism. A variety of mood-induction manipulations have been developed in laboratory settings, such as having participants read affectively positive or negative passages (e.g., Aderman, 1972). After the mood state induction, participants are given an opportunity to exhibit generosity by donating money or helping an experimental accomplice. Results generally show that positive mood induction elevates helping behavior but depressed mood inhibits helping.

Despite multiple replications of this effect in different laboratories with different investigators, the validity of these findings has been challenged both because of the artificiality of the setting in which altruism is assessed and because of the potential demand characteristics associated with the rather unusual mood-induction experience. To counter these criticisms, researchers in the area took advantage of a natural mood-induction situation based on the emotional impact of selected motion pictures (Underwood et al., 1977).

After some pilot research, in which ratings were obtained from moviegoers, a double feature consisting of *Lady Sings the Blues* and *The Sterile Cuckoo* was selected for its negative-affect-inducing qualities, and two other double features were selected to serve as neutral control conditions. A commonly occurring event – solicitation of donations to a nationally known charity with collection boxes set up outside the movie theater lobby – was chosen as the vehicle for a measure of the dependent variable of generosity.

Having located such naturally occurring variants of the laboratory mood-induction operation and altruism measure, the major design problem encountered by the researchers was that of participant self-selection to the alternative movie conditions. Although random assignment of volunteer moviegoers was a logical possibility, the procedures involved in utilizing that strategy would have created many of the elements of artificiality and reactivity that the field setting was selected to avoid. Therefore, the investigators decided to live with the phenomenon of self-selection and to alter the research design to take its effect into consideration. For this purpose, the timing of collection of donations to charity at the various theaters was randomly alternated across different nights so that it would occur either while most people were entering the theater (before seeing the movies) or leaving (after seeing both features). The rate of donations given by arriving moviegoers could then be a check on preexisting differences between the two populations apart from the mood induction. Fortunately, there proved to be no differences in initial donation rates as a function of type of movie, whereas post-movie donations differed significantly in the direction of lowered contribution rates following the sad movies. This pattern of results, then, preserved the logic of random assignment (initial equivalence between experimental conditions) despite the considerable deviation from ideal procedures for participant assignment.

Two points should be emphasized with respect to this illustration of field research. First, the field version of the basic research paradigm was not – and could not be – simply a "transplanted" replication of the laboratory operations. The researchers had considerably less control in the field setting. They could not control the implementation of the stimulus conditions or extraneous sources of variation. On any one night a host of irrelevant events may have occurred during the course of the movies (e.g., a breakdown of projectors or a disturbance in the audience) that could have interfered with the mood manipulation. The researchers were not only helpless to prevent such events but would not have been aware of them if they did take place. In addition, as already mentioned, in the field setting the experimenters were unable to assign participants randomly to conditions and had to rely on luck to establish initial equivalence between groups.

Second, the results of the field experiment as a single isolated study would have been difficult to interpret without the context of conceptually related laboratory experiments. This difficulty is partly due to the ambiguities introduced by the alterations in design and partly to the constraints on measurement inherent in the field situation where manipulation checks, for example, are not possible. The convergence of results in the two settings greatly enhances our confidence in the findings from both sets of operations.

Isolation Versus Construct Validity

Laboratory experiments are inherently artificial in the sense that causal variables are isolated from their normal contextual variation. This isolation and control is the essence of testing causal hypotheses with a high degree of internal validity. It has been pointed out repeatedly in this chapter that isolation does not necessarily jeopardize external validity if the experimental situation has psychological realism, that is, if the causal processes being represented in the lab setting are the same as those that operate in nonlaboratory contexts.

It is this matter of whether the process is the "same" when the context is altered that constitutes the stickiest issue of validity. Greenwood (1982) called this the problem of "the artificiality of alteration," the problem that arises whenever bringing a variable into the laboratory changes its nature. Greenwood argued that this alteration is particularly problematic for social psychology because social psychological phenomena are inherently relational or context-dependent and hence do not retain their identity when isolated from other psychological processes. He takes this as a fatal criticism of laboratory research methods in social psychology, but the truth is that it applies to any context in which a phenomenon is observed as psychological experiences are never exactly the same from one time or place to another.

The issue here is one of the level of abstraction at which constructs or principles are defined. Consider, for example, the construct of "threat to self-esteem." No one would seriously deny that being informed that one had failed a test of creative problem-solving would have more impact on self-esteem of a Harvard undergraduate than it would on a 50-year-old mineworker. Thus, if we were interested in the effects of lowered self-esteem on aggression, we might have to use different techniques to lower self-esteem in the two populations. Threats to self-esteem based on challenges to one's academic self-concept are certainly different in many ways from challenges that threaten one's sense of group belonging or of physical stamina. But if each of these, in their appropriate context, proves to have an impact on anger or aggressiveness, then we have gained confidence in a general principle that threats to areas of self-esteem that are important or central to one's sense of identity increase aggression.

This, then, is the ultimate challenge for valid theory-building in social psychology. Our theoretical constructs must be abstract enough to generalize across a range of contexts and specific manifestations, yet precise enough to permit testing at an empirical level. Each empirical demonstration is inevitably limited to a specific context and subject to multiple interpretations. But each time a theoretical proposition is tested in a new setting or with new operations, a contribution is made to the overall picture. Validity is never the achievement of a single research project but the product of cumulative theory-testing and application.

REFERENCES

Aderman, D. (1972). Elation, depression, and helping behavior. *Journal of Personality and Social Psychology, 24,* 91–101.

Aronson, E., & Mills, J. (1959). The effect of severity of initiation on liking for a group. *Journal of Abnormal and Social Psychology, 59,* 177–181.

Aronson, E., Wilson, T., & Brewer, M. B. (1998). Experimentation in social psychology. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 1, pp. 99–142). Boston: McGraw-Hill.

Asch, S. E. (1956). Studies of independence and conformity: A minority of one against a unanimous majority. *Psychological Monographs, 70* (9, Whole No. 416).

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

Berkowitz, L., & Donnerstein, E. (1982). External validity is more than skin deep: Some answers to criticisms of laboratory experiments. *American Psychologist, 37,* 245–257.

Brewer, M. B. (1997). The social psychology of intergroup relations: Can research inform practice? *Journal of Social Issues, 53*(1), 197–211.

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

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

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

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Chicago: Rand-McNally.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand-McNally.

Collingwood, R. G. (1940). *An essay on metaphysics.* Oxford, England: Clarendon.

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

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

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs.* San Francisco: Jossey-Bass.

Gasking, D. (1955). Causation and recipes. *Mind, 64,* 479–487.

Gerard, H. B., & Mathewson, G. C. (1966). The effects of severity of initiation on liking for a group: A replication. *Journal of Experimental Social Psychology, 2,* 278–287.

Greenwald, A. G. (1975). On the inconclusivenes of "crucial" cognitive tests of dissonance versus self-perception theories. *Journal of Experimental Social Psychology, 11,* 490–499.

Greenwood, J. D. (1982). On the relation between laboratory experiments and social behaviour: Causal explanation and generalization. *Journal for the Theory of Social Behaviour, 12,* 225–250.

Mackie, J. L. (1974). *The cement of the universe.* Oxford, England: Oxford University Press.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology, 67,* 371–378.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38,* 379–387.

Orne, M. (1962). On the social psychology of the psychological experiment. *American Psychologist, 17,* 776–783.

Petty, R. E., & Cacioppo, J. T. (1996). Addressing disturbing and disturbed consumer behavior: Is it necessary to change the way we conduct behavioral science? *Journal of Marketing Research, 33,* 1–8.

Rakover, S. S. (1981). Social psychology theory and falsification. *Personality and Social Psychology Bulletin, 7,* 123–130.

Rosenberg, M. J. (1969). The conditions and consequences of evaluation apprehension. In R. Rosenthal & R. Rosnow (Eds.), *Artifact in behavioral research* (pp. 279–349). New York: Academic Press.

Rosenthal, R. (1966). *Experimenter effects in behavioral research.* New York: Appleton-Century-Crofts.

Sears, D. O. (1986). College sophomores in the laboratory: Influence of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology, 51,* 515–530.

Sherif, M. (1935). A study of some social factors in perception. *Archives of Psychology, 27*(187), 1–60.

Tajfel, H. (1970). Experiments in intergroup discrimination. *Scientific American,* 223(2), 96–102.

Underwood, B., Berenson, J., Berenson, R., Cheng, K., Wilson, D., Kulik, J., Moore, B., & Wenzel, G. (1977). Attention, negative affect, and altruism: An ecological validation. *Personality and Social Psychology Bulletin, 3,* 54–58.

Zanna, M., & Cooper, J. (1974). Dissonance and the pill: An attribution approach to studying the arousal properties of dissonance. *Journal of Personality and Social Psychology, 29,* 703–709.