Theory-Testing, Generalization, and the Problem of External Validity*

JEFFREY W. LUCAS
The University of Akron

External validity refers to the generalization of research findings, either from a sample to a larger population or to settings and populations other than those studied. While definitions vary, discussions generally agree that experiments are lower in external validity than other methodological approaches. Further, external validity is widely treated as an issue to be addressed through methodological procedures. When testing theories, all measures are indirect indicators of theoretical constructs, and no methodological procedures taken alone can produce external validity. External validity can be assessed through determining (1) the extent to which empirical measures accurately reflect theoretical constructs, (2) whether the research setting conforms to the scope of the theory under test, (3) our confidence that findings will repeat under identical conditions, (4) whether findings support the theory being tested, and (5) the confirmatory status of the theory under test. In these ways, external validity is foremost a theoretical issue and can only be addressed by an examination of the interplay between theory and methods.

INTRODUCTION

In the social sciences, a common criterion for evaluating research investigations is the extent to which results can be generalized, with results that are more generalizable viewed as more desirable than results that are less generalizable. However, social scientists are seldom clear in any explicit way about what is meant by generalization. Often, generalization of research results is discussed as an issue of external validity. Treatments and definitions of external validity, however, vary considerably.

This paper represents an effort to identify the conditions under which findings may be generalized, focusing particularly on the concept of external validity. Because external validity is often seen as especially problematic for experimental designs, much of the paper will center around issues of external validity and experimentation. The paper will read in many ways as a defense of experimental methodology against claims of external invalidity. Although this is one of my goals, my overriding objective is to discuss ways in which investigators can produce general knowledge in the social sciences. I conclude that external validity is primarily a theoretical concern and present several criteria by which the external validity of any project designed to test theory can be addressed.

DEFINING EXTERNAL VALIDITY

A primary goal in all sciences, including the social sciences, is the production of general knowledge. General knowledge is knowledge that is not confined to the particulars of time and place. In a simple sense, we might say that we have produced

*Direct correspondence to: Jeffrey W. Lucas, Department of Sociology, University of Akron, Akron, Ohio, 44325-1905. E-mail: jlucas2@uakron.edu. I thank Barry Markovsky, Dave Willer, and John Zipp for valuable comments on an earlier draft.

Sociological Theory 21:3 September 2003

general knowledge when we have confidence that demonstrated relationships will hold beyond the particulars of time, place, and methodology. One criterion for assessing research findings, then, is whether findings will generalize beyond the parameters of the particular study. In the social sciences, the generalization of research findings is often discussed as an issue of external validity.

Although external validity is discussed as the extent to which results can be generalized, its treatments and definitions vary widely beyond this simple notion. Consider definitions of external validity provided in two recently published textbooks on social research methods:

> External validity . . . refers to whether the results of a study can be legitimately generalized to some specified broader population. (McTavish and Loether 2002:133)

> [External validity] concerns the extent to which causal inferences . . . can be generalized to other times, settings, or groups of people. (Monette, Sullivan, and DeJong 2002:236)

Aside from sharing a focus on generalization, these definitions are quite distinct. The first definition treats external validity as generalizing from a sample to a larger population,[1] while the second treats external validity as generalizing to populations or settings *other* than those studied.

This distinction in treatments of external validity is echoed in an examination of the research literature. Some treat external validity as generalizing from a sample to a larger population (see, e.g., Bass and Firestone 1980; Christensen 1977; Devine, Brody, and Wright 1997; Esbensen and Osgood 1999; Hanson and Tuch 1984; Hopkins, Hopkins, and Schon 1988; Ross and Mirowsky 1999; Sellers et al. 1997; Taylor 1980; Thompson, Brownfield, and Sorenson 1996), while others treat external validity as generalizing across settings or populations (see, e.g., Babbie 1989; Easley, Maden, and Dunn 2000; Gergen 1978; Locke 1986; Louviere 1988; Sullivan 2001; Thye 2000; Top 1991; Willems and Howard 1980; Winer 1999). Still others discuss external validity as generalization both to and across settings and populations, or without reference to where generalizations are made (see, for example, Bear 1995; Dipboye and Flanagan 1979; Frankfort-Nachmias and Nachmias 1996; Shadish 2002).

The original statement of external validity was made by Donald Campbell and Julian Stanley (1963:5): "External validity asks the question of generalizability: To what populations, settings, treatment variables, and measurement variables can this effect be generalized?" This definition can incorporate either generalizing to a larger population or generalizing across settings and populations. A later co-publication by one of these original definers of external validity, however, addressed the distinction on the sorts of generalizations to which external validity applies. Thomas Cook and Donald Campbell (1979) stated that external validity involves generalizing (1) to particular target persons, settings, and times and (2) across types of persons, settings, and times.

External validity, then, refers to both generalizing *to* and generalizing *across*. Although this distinction is not often made, particularly by sociologists, it is important. For example, if one treats external validity as generalizing from a sample to a

[1]In many cases, investigators who attempt to draw conclusions about a larger population from a probability sample do not specify a logic for their choice of population. Nevertheless, by "larger population," I refer to the full range of individuals encompassed by the scope conditions of the theory under test. Samples, then, represent a subset of the full class of individuals within which a theoretical proposition is expected to hold.

specified larger population, then the external validity of survey projects might be increased through probability sampling methods. As will be discussed, such sampling techniques, however, would do little to increase external validity if we use it to refer to generalization across populations and settings.

THEORY-TESTING AND GENERALIZATION

Whatever stance one takes on the definition of external validity, it is widely accepted that it refers to the generalizability of research findings. Before determining whether different research methodologies are higher in external validity than others, however, we must determine what we mean by generalization. According to some (e.g., Shavelson and Webb 1991), generalizability refers to how dependable a measure is across contexts. However, in that all measurement is indirect (Blalock 1979), having confidence that measurements will behave similarly across contexts does little to advance the production of general knowledge.

Many concepts in the social sciences are impossible to measure objectively, forcing investigators to rely on subjective assessments (Bollen and Paxton 1998). When testing theoretical principles, our goal is to construct measures that accurately reflect theoretical constructs. However, any empirical investigation is conducted in a concrete setting defined by time and place (Cohen 1980), while our goal in producing general knowledge is to develop general principles unbounded by the particulars of time and place.

Because all measures are indirect, and because all investigations are conducted in concrete settings, it is impossible to produce general knowledge in the absence of theory. In fact, the outcome of all sociological research methods—the finding—has no phenomenological status whatsoever (Willer and Willer 1973). Findings tell us how particular measures behaved in a particular time and place, nothing more (Willer and Webster 1970). Strict sampling procedures might give us confidence that indirect measures will behave similarly at a larger level, but this information will not link our measures to our theoretical constructs. Further, highly realistic research settings might increase the information we can obtain about a particular phenomenon in a particular time and place, but without theory we cannot generalize this knowledge to new settings.

General knowledge is theoretical knowledge. Through their abstractness, theories apply across settings and populations; empirical findings do not (Dobbins, Lane, and Steiner 1988). Thus, the key component of external validity is theory. Further, critiques of investigative techniques as being low in external validity because findings cannot be generalized quite often should be directed at the theory under test, rather than at the methodology employed to test it.

Generalization Critiques and Theory

Methodological discussions of external validity frequently present it as an issue that is particularly problematic for experimental designs. The major drawback to experimental investigations is that, at times, they cannot create or manipulate all theoretically meaningful variables. However, if an experiment does manipulate every theoretically relevant variable and finds an effect, then to say that the effect will not generalize to naturally occurring situations is not a criticism of the experiment as having low external validity; rather, it is a critique of the theory for not taking every factor influencing the phenomenon of interest into account.

Suppose that a theoretical proposition predicts that some concept A will produce some concept B. If an experiment finds no relationship between A and B, but a study in a naturally occurring situations does find an effect of A on B, what does this tell us? One might conclude that the experimental test has low external validity. However, if the experiment is well designed, what the two sets of findings would indicate is that factors vary with A in affecting B that are not specified in the theory under test. Thus, if experiments do not have external validity in testing theories, it indicates, not a problem with the methodology, but more likely a shortcoming of the theory.

If experimental tests can manipulate every theoretically meaningful variable and nothing else, then a problem with the external validity of an experiment is a problem with the utility of the theory. In other words, if a theory is supported in well-designed experiments but the findings do not hold in naturally occurring situations, then there are variables impacting the phenomenon under study for which the theory does not account. In these instances, critiques of external validity should be focused on the theory under test, rather than on the methodology used to test it.

All findings based on any methodology are bounded by particulars of time and place. However, as theories are supported in diverse tests, we begin to have increased confidence in the utility of the theory. One method for gaining increased confidence in theoretical propositions is the replication of research findings. Some have argued that there is a bias against replication research in the social sciences (e.g., Bornstein 1990; Neuliep and Crandall 1993), or that social scientists must carry out more replication research to increase external validity (e.g., McGrath and Brinberg 1983). However, replication does not always equate with testing the same theoretical principles under identical conditions.

The Role of Replication

While all investigations are particular, conducted in concrete settings defined by time and place, theories are abstract, applying to many settings. Following a deductive approach (Nagel 1961), theoretical propositions subsume many empirical instances. Although at first thought it might seem that there is very little replication research in the social sciences, there is, in fact, a great deal. Investigators carry out replication research whenever they test the same theoretical propositions with alternative empirical indicators.

Theoretical replication involves the testing of the same theoretical principles. These tests may or may not employ the same methods and settings as previous tests. For example, institutional theory predicts that organizations will adopt practices that are legitimated in their environments. This proposition has been tested in numerous settings with various measures, from the adoption of human-resources practices (Gooderham, Nordhaug, and Ringdal 1999) to the hiring of female CEOs (Guthrie and Roth 1999) to HIV-prevention practices at treatment centers (D'Aunno, Vaughn, and McElroy 1999). Each of these tests represents a theoretical replication of the prediction that organizations will adopt institutionalized practices.

Replication research, then, is widespread in the social sciences. This research, however, does not involve testing the same principles with the same empirical indicators in the same methodological setting. Instead, theoretical replication often involves testing the same theoretical principles with new measures in new settings. When a theoretical principle is supported in diverse replications, we gain confidence in the theory, and each successive test increases external validity.

EXTERNAL VALIDITY AND EXPERIMENTATION

As noted, external validity is frequently discussed as an issue particularly faced by experimental designs. Experiments involve the creation of artificial conditions in which to test theoretical predictions. In most cases, experimentally created conditions in the social sciences differ in many respects from naturally occurring situations. Further, most social-science experiments do not use probability-sampling techniques; instead, convenience samples, often comprised of undergraduate students, are the norm.

Whichever stance one takes on the definition of external validity, it is widely accepted in the social sciences that experiments are lower in external validity than are other research methods (see, e.g., Aronson, Brewer, and Carlsmith 1985; Babbie 1989; Cheuk et al. 1997; Christensen 1977; Esbensen and Osgood 1999; Frankfort-Nachmias and Nachmias 1996; Gergen 1978; Gordon, Slade, and Schmitt 1986; Hanson and Tuch 1984; Hogarth 1986; McTavish and Loether 2002; Mook 1983; Pettigrew 1988; Shrauger and Schoeneman 1979; Sullivan 2001; Taylor 1980; Top 1991; Wells and Windschitl 1999; Wiggins 1983; Willems and Howard 1980). To some (e.g., Christensen 1977; Esbensen and Osgood 1999; Gordon, Slade, and Schmitt 1986; Hanson and Tuch 1984; Taylor 1980), the external invalidity of experiments is a matter of sampling. Because social-science experiments generally do not employ probability-sampling techniques, the logic goes, they cannot then generalize from the study's sample to a larger population. To others (e.g., Babbie 1989; Gergen 1978; Pettigrew 1988; Sullivan 2001; Top 1991; Willems and Howard 1980), external invalidity stems from the artificiality of experimental conditions. According to this view, experiments create conditions that do not occur in the "real world," and experimental findings thus cannot be generalized from the setting of the study to other, naturally occurring settings.

Both of these schools of thought represent some misconceptions regarding experiments and, more generally, how generalization can happen from any research method. To understand how generalization can happen for any research method, it is useful to address generalizing *to* and generalizing *across* in experimental research.

Generalizing To: External Validity as Generalization to a Larger Population

Experiments are considered by some to be low in external validity because they do not produce findings that can generalize from a sample to a specified larger population. Social-science experiments often include samples comprised of undergraduate students selected from courses at one university. Studies employing surveys, in contrast, usually draw probability samples from larger populations. These differences in sampling techniques lead many to conclude that experiments are low in external validity when external validity is treated as generalization from a study to a larger population. This treatment of external validity is similar to statistical inference validity (Dooley 1990), which questions whether a sample is a good representation of a population.

While sampling techniques may affect the extent to which findings can be generalized from a sample to a larger population, for several reasons, nonprobability samples do not necessarily equate with low external validity.

(1) Experiments often seek to test theoretical relationships, rather than to generalize findings. Some experimentalists (e.g., Martin and Sell 1979; Meeker and Leik 1995; Mook 1983; Webster 2003) have pointed out that the goal of experimental research is

to test theories, not to generalize from a sample to a larger population. Instead of seeking to discover empirical generalizations that apply to larger groups, social-science experiments test theoretical principles. For example, a number of theories (e.g., identity theory, network exchange theory, and power dependence theory) provide explanations of the structural determinants of power in exchange networks. These theoretical explanations are usually tested through experiments in which individuals—often undergraduate students—interact in controlled laboratory experiments. The goal of experimenters in these studies is to test theoretical explanations of the structural constraints on power development, not to produce findings that generalize to some larger population.

When an experiment is designed to test theoretical principles, to ask whether the experiment's sample allows for generalization to a larger population is to ask the wrong question (Berkowitz and Donnerstein 1982). In other words, to say that an experiment is low in external validity because the sample is not representative of a larger population is analogous to critiquing the General Social Survey as a poor device for generating rich description of social phenomena. Both critiques misunderstand the goals of investigations.

(2) Experiments often test theories that do not specify target populations for generalization. A second response to the claim that experiments cannot generalize from a sample to a larger population is that experimental tests of theory, in many cases, do not have a larger population to which investigators aspire to generalize. For example, it seems that when discussing the external validity of experimental research due to sampling procedures, critics often take the United States's adult population as the benchmark. Then, if experiments do not have a probability sample of U.S. residents (which they never do), they are thought to be externally invalid. The theories that experiments test, however, rarely specify a particular population as the parameters within which the propositions of the theories are expected to hold.

Well-specified theories have scope conditions that specify the circumstances under which the relationships expressed in a proposition are expected to hold (Foschi 1997). In a test of a theory, experimental or otherwise, scope conditions specify the class of circumstances to which a theory is predicted to apply. In many cases, experiments test theories that do not restrict the population of individuals to which a theory applies. Status characteristics theory (Berger, Cohen, and Zelditch 1966, 1972), for example, has clearly specified scope conditions that are not limited by particular population parameters, and the theory has been supported in a large number of experimental tests. Criticizing an experimental test of status characteristics theory that employs undergraduate students as having low external validity because results cannot be generalized to a larger population misses the point. The theory makes propositions on human behavior unbounded by the particulars of population parameters; there is no larger population to which generalizations are intended to be made.

Experimental tests of theories often have the goal of uncovering laws of human behavior (Henshel 1980). These sorts of investigations are also not guided by particular population parameters. In seeking laws of human behavior, one cannot draw a representative sample of the population, because there is no larger population from which to draw a sample. To criticize these experiments as low in external validity because of unrepresentative samples follows from applying the logic of survey research to the goals of experimental inquiry (Freese 1980). Experimental tests that seek to uncover laws of human behavior are not intended to produce results that generalize from a sample to some larger population.

(3) Generalizing to larger populations is a matter of procedure, not a matter of methodology. A third response to the claim that experiments are low in external validity because of unrepresentative samples is that the criticism applies to experimental tests, not to experimental methodology. In that surveys are more likely to use probability-sampling techniques than are experiments, it may be more likely that a given survey will contain a sample more representative of a larger population than will a given experiment. However, a general claim that experiments are lower in external validity than other methodological approaches would be inappropriate. The difference in sampling between many surveys and many experiments is a matter of procedure, an empirical matter.

There is nothing inherent about surveys as compared to experiments that makes for more representative samples. Differences in representativeness between survey and experimental studies are a matter of design, not a matter of method. Thus, to claim that experiments are lower in external validity than are surveys because experimental samples do not represent larger populations is a prejudicial generalization. As social scientists, we would never accept statements of this sort when describing social phenomena. For example, European-Americans, on average, score higher than African-Americans on standardized tests of achievement. There is no evidence for differences in mental ability between European-Americans and African-Americans, and test-score differences are at least largely due to social processes. Suppose that, based on differences in test scores, a sociologist claimed a difference in mental ability between European-Americans and African-Americans. This statement would be roundly and justifiably criticized for ignoring factors that vary with race and that affect scores on tests of achievement.

A statement that experiments are lower in external validity than are surveys because of unrepresentative samples follows the same logic. Particular experiments may be less likely to employ probability samples than particular surveys, but nothing about experiments requires that researchers do not draw probability samples. Further, for all surveys, several factors decrease the extent to which samples are representative of larger populations (e.g., most surveys use geographically restricted samples and response rates over 75 percent are rare in survey research). Thus, while experimental projects, on average, may have less representative samples than do survey projects, it is inappropriate to critique experimental *methodology* as allowing for less representative samples than does survey methodology.

(4) Generalizations of any sort can only happen through theory. The most important response to the claim that experiments lack external validity because of unrepresentative samples is that generalizations of any sort can only occur through theory. Suppose that two researchers have completed independent studies. One researcher has completed a survey with a probability sample drawn from a larger population. The other researcher has completed an experiment with a sample drawn from classes at one university. Even though generalization to a larger population is rarely, if ever, the goal of experimenters, further suppose that both the survey researcher and experimenter seek to generalize their findings to a larger population. In this case, we might conclude that the survey has higher external validity, in that its findings can more easily be generalized to the larger population.

To make this sort of conclusion, we must determine what we mean by generalization. If by generalization we mean that measurements taken from a sample will behave similarly under the same conditions at the population level, then it may be appropriate to conclude that the survey would be more generalizable. Generalizations of

this sort, however, do not do us much good. Measurements in any theory-testing study represent empirical indicators of theoretical constructs. These empirical indicators are always constrained by particulars of time, place, and measurement. Thus, we might have confidence that findings from a sample will hold in a larger population, but that tells us very little—only that empirical indicators will vary similarly in a larger group.

Data are always limited to the special case of what happened when the measurements were made (Ahl and Allen 1996), while theory subsumes our data. Theory links our empirical indicators to the constructs we intend them to measure; in other words, a construct can never be observed, only an instantiation of a construct (Chow 1996). Without theory, the most we can know is how our measurements will behave among different groups of individuals. To produce general knowledge, however, we must leave behind our empirical indicators and be able to draw conclusions about the constructs our indicators measure. As will be elaborated in a later section, this can only occur through theory.

Critiques of experiments as having low external validity because of unrepresentative samples fail to recognize that representative samples alone do not allow for generalization from a sample to a larger population. Even if we have confidence that relationships demonstrated in a sample will hold at the population level, we only produce general knowledge when our empirical indicators are linked to our constructs through theory.

Claims that experiments cannot generalize from a sample to a larger population stem from the fact that experiments do not use representative samples. Missing from sociological discussions of this issue are the advantages of using nonprobability samples when testing theory.

The Advantages of Nonprobability Samples

As discussed, the purpose of experiments is to test theoretical propositions. Theoretical propositions, of course, can never be proven to be correct. The most support we can hope to achieve for a theoretical proposition in the social sciences is that it has escaped refutation in a number of diverse tests (Calder, Phillips, and Tybout 1982). Thus, the goal of any experimental test is to construct a situation in which a proposition may or may not be refuted. This occurs through designing a setting that conforms to the scope conditions of the theory under test.

Following the falsification approach (discussed in a later section) that dominates in the social sciences, experiments create situations under which researchers can determine whether the null hypothesis has been falsified. When this occurs, we gain increased confidence that the alternative hypothesis may be correct. With this in mind, the most stringent test of a theory is one that allows the highest likelihood for falsifying the null hypothesis if it is indeed false for the particular group under study.

Nonprobability samples (e.g., undergraduate students from the same university) will, in most cases, be more homogenous than probability samples. Thus, one outcome of using probability samples in experimental research will very likely be increased variability in measurements (Lynch 1982). This variability will increase the likelihood of false rejection of the null hypothesis. Thus, when constructing a setting in which to falsify a theory, homogenous samples have the advantage over representative samples that heterogeneity of participants reduces the likelihood of identifying violations of a theory when it is false (Lynch 1983). This is a significant benefit to using nonprobability sampling techniques in experimental research.

In addition to being critiqued as not allowing for generalization from a sample to a larger population, experiments are criticized for not allowing generalization *across* populations and settings.

Generalizing Across: External Validity as Generalizing to New Populations or Settings

External validity is defined both as generalizing from a sample to a larger population and as generalizing to populations and settings other than those studied. Generalizing from a sample to a larger population is usually viewed as a sampling issue, with more representative samples seen as allowing for greater generalization than do less representative samples. Stringent sampling techniques, however, can do nothing to increase generalization *across* settings and populations.

When considering the ability of experimental designs to generalize across settings and populations, there are two primary issues to consider—sampling procedures and artificiality. Both are thought to limit the generalization of experimental findings.

Sampling Procedures and Generalizing Across Populations

Social-science experiments are thought to be low in external validity because they do not employ probability-sampling techniques (e.g., Ferber 1977). The previous section addressed why this critique is misguided when considering generalization from a sample to a larger population. Further, when researchers aspire to generalize beyond the populations studied, no sampling techniques—however stringent—can produce external validity.

Suppose that two investigators seek to test a theoretical proposition that higher status will be accompanied by higher self-efficacy. Further suppose that one investigator chooses to test this prediction through an experiment and the other through a survey. The experimenter obtains a sample of female freshman from a course at a university. She then randomly assigns each participant to either a high- or low-status condition by manipulating the characteristics of fictitious partners. She finds higher self-efficacy in the high-status condition. The survey researcher obtains a probability sample of working adult Americans. Her survey asks respondents to list their occupations and to complete a self-efficacy scale. She finds higher self-efficacy for higher-status occupations.

Conventional wisdom tells us that the survey test of the status/self-efficacy hypothesis has higher external validity because the researcher used probability-sampling techniques. The survey test included a sample of adult Americans and thus can be generalized to working adults in the United States. The experiment, on the other hand, included only female freshman and will not allow for generalizations beyond female freshman. Thus, most (see, e.g., Esbensen and Osgood 1999; Hanson and Tuch 1984) would determine that the survey test is higher in external validity than is the experimental test.

Notice, first, the criteria that are applied above when examining the external validity of the survey and experimental designs. When discussing the external validity of survey research, social scientists often focus on whether findings from a sample can be generalized to the population from which the sample was drawn (e.g., Devine, Brody, and Wright 1997; Hopkins, Hopkins, and Schon 1988). When examining the external validity of experimental designs, social scientists generally attempt to determine whether findings can be generalized to new populations. Due, in part, to restricted samples, the answer is that they cannot. This leads many to conclude that experiments are low in external validity. Thus, in traditional views, surveys are

considered externally valid if findings from a sample can be generalized to the population from which the sample was drawn. Laboratory experiments are considered externally invalid because findings from the population studied cannot be generalized to populations other than the one from which the sample was drawn.

When considering generalization across populations, however, probability samples allow for no more generalization than do convenience samples. Samples in research, by definition, are drawn from a larger population, and findings from a sample, taken alone, can tell us nothing about *other* populations. We would have no more reason to think that the survey with a sample of adult Americans could generalize across populations than to think that the experiment comprised of female freshman could generalize across populations. In and of themselves, no methodological procedures can allow for generalization across settings and populations. This can only occur through theory.

Artificiality and Generalizing Across Settings

Generalizing experimental findings across *settings* is often discussed as an issue of artificiality. Some (e.g., Gergen 1978; Pettigrew 1988) contend that the artificial settings of experiments do not mirror the "real world" and that these artificial designs allow for low generalization to other settings. However, just as no sampling procedures can allow for generalization across populations, no methodological procedures in and of themselves will allow for generalization across settings.

Consider another test of the status/self-efficacy hypothesis. Suppose that a researcher conducts a project to test this prediction, relying primarily on participant observation. The researcher observes groups of high-school math students interacting in a class to solve a task. He finds that members with the highest status (i.e., those who talked the most, had the most influence, and so on) scored higher, on average, on a standard self-efficacy scale than did students with lower status.

When considering generalization across settings, conventional wisdom would again lead us to conclude that the experimental test is lower in external validity, this time less than the qualitative test.[2] The qualitative test involved "real people" interacting on a real task in the real world, while the experimental test involved a highly artificial setting. It seems reasonable to conclude that a test in a more natural environment will allow for more generalization than will a test in a more artificial environment.

The qualitative findings, however, are no more generalizable to *other* types of groups than are the experimental findings. The qualitative researcher studies high-school students interacting in the classroom. The findings of this research may generalize to other groups of high-school math students,[3] but this sort of generalization is almost certainly not the researcher's goal. Assuming that the qualitative researcher seeks to generalize his findings beyond groups of high-school math students, will his findings be generalizable to executives interacting in a boardroom or to a family interacting at a dinner table? There is no reason to think so. In the same way, the findings of the experimental study might generalize to other female freshman

[2]I present an example in which qualitative analysis is used to test a theoretical prediction. In most cases, qualitative approaches are used to generate theory, rather than to test it. However, there is no logical reason to believe that qualitative analyses cannot be used to test deductive theory.

[3]While the qualitative findings on status and self-efficacy may generalize to other high-school math students, designs in natural settings face methodological issues that may limit generalization—as do experiments. These include changes in behavior due to the presence of an intrusive investigator (Berg 2001) and the fact that naturalistic reports will always be filtered through researchers' perspectives (Esterberg 2002).

interacting with fictitious partners, but we would have no reason to think that the findings would generalize to executives in a boardroom or to a family at a dinner table.

Although conventional wisdom tells us that the qualitative findings would be more generalizable to other settings than would the experimental findings, neither test in and of itself can allow for generalizations beyond the particular types of groups studied. Generalizations of this sort can only occur through theory. Further, not only is a critique of experimental artificiality as producing low external validity not appropriate, but it also ignores the major advantage of experimental investigations.

The Advantages of Artificiality in Experimental Research

Because the goal of experiments is to test theoretical principles, experimentalists seek to create conditions in which theoretical constructs can be observed and measured. However, theoretical constructs can never be directly observed, only instantiations of constructs (Xiao and Vicente 2000). This is true for any investigation grounded in any methodology—when testing theories, we can never directly observe our theoretical constructs.

The objective of experiments is to create conditions that satisfy the requirements of the theory under test. To this end, well-designed experiments simplify naturally occurring situations, incorporating only theoretically relevant elements (Webster 1994). This is the advantage of artificiality. Artificial experimental conditions allow researchers to examine only elements of the situation that are relevant to theory under test; other elements that may mask or vary with predicted effects are eliminated (Lovaglia et al. 1998).

To produce findings that are generalizable, we must be able to transcend the data we collect. Because we can never directly measure theoretical constructs, findings from any single field setting are no more generalizable than are results from a single laboratory experiment (Lynch 1999). In this way, artificiality does not reduce external validity. Instead, the key to evaluating any research design is whether the situation adequately represents the concepts or relations specified in the theoretical hypotheses under test (Kruglanski and Kroy 1976).

One might argue that tests of theory in natural settings collect more detailed information on the topic under study and thus produce more generalizable findings. However, the more particular information we gather about a specific phenomenon, the less that information will generalize to *other* phenomena (Markovsky 1994). Further, theories do not apply to unique, nonrecurring situations (Webster and Kervin 1971). Since generalization, by definition, applies knowledge outside the sample of cases studied to any given point, if we are to generalize findings, our knowledge must transcend information given by the data that have been observed (Willer and Heckathorn 1980). The advantage of artificiality is that it simplifies naturally occurring situations to only theoretically meaningful aspects, eliminating variables that are not relevant and making generalization more likely (Greenwood 1982).

What Is the "Real World?"

Critiques of experimental artificiality (e.g., Gergen 1978; Top 1991) frequently note that experiments do not accurately represent the real world. According to this view, the artificiality of experiments creates less "real" situations than do studies in natural

settings. Because of this artificiality, the logic goes, individuals behave differently than they otherwise would. However, both experiments and tests in natural settings involve the indirect observation of theoretical constructs. Experiments do manipulate features of situations to eliminate aspects that are not theoretically relevant, but the experiences of participants are as real as in any other situation.

As an academic, I am bothered by statements that professors are detached from the real world and make proclamations from ivory towers. I want to ask those who make these statements how my world is less "real" than their world. I go to work every day, have relationships with co-workers, shop at the grocery store, interact with neighbors, watch television, and so on. I cannot identify any ways in which my world is less real than that of others.

To critique experiments as not faithfully representing the real world is analogous to critiquing academics as living in ivory towers. Participants in experiments must respond to manipulated features of the situation within which they find themselves. The features of experiments might be unique, but this does not make the experiences of participants less real. The advantage of well-designed experiments is that participants respond only to factors that are theoretically relevant—but only the factors to which participants respond are artificially created, not their responses.

Falsification and Theory-Testing

I have pointed out that sociologists generally ascribe to Popper's (1959) falsification approach to theory-testing. According to this view, a theory is never confirmed; it merely escapes falsification. Because theories are general but apply to specific instances, a theory can be falsified if its predictions can be shown to be inaccurate for any subset of instances to which it is intended to apply. The goal of any test of theory in the social sciences should be to create a situation in which the propositions of a theory are intended to apply and then to seek to falsify the predictions.

The objective of experiments, then, is to create situations in which a theory can be falsified. If the theory under test is falsified, we have evidence that the theory requires modification. If the theory escapes falsification, we gain confidence in the utility of the theory. As a theory escapes falsification in multiple tests, we begin to have confidence that the theory will hold in diverse situations. It is in this way that we produce general knowledge. No single study, however, can produce general knowledge in the absence of theory. This is because no study, taken alone, generalizes to new settings and populations. Further, no study, however well designed, can directly measure theoretical constructs, and thus no study can be generalized from a sample to a larger population, except in terms of how indirect measurements of constructs might behave at a larger level.

The falsification approach to theory-testing has implications for how the generalization of findings might occur for studies based on any research methodology.

ASSESSING EXTERNAL VALIDITY

The major task in any science, including the social sciences, is the development of theory (Schmidt 1992). In that general knowledge is theoretical knowledge, no findings based on any research methodology can have external validity in the absence of theory. For this reason, it is meaningless to discuss different methodological approaches as though they have higher or lower external validity than others. The assessment of external validity relies primarily on characteristics of the theory

under test. For any test based on any methodology that supports the propositions of a theory, we can use the following five criteria to assess the external validity of the test.[4]

(1) Construct validity. Construct validity concerns the extent to which measures accurately reflect the theoretical concepts they are intended to measure (Carmines and Zeller 1979). Although construct validity and external validity are usually discussed as separate issues, construct validity is a necessary prerequisite to external validity. Because theoretical concepts are never measured directly, and because generalization can only occur through applying findings to theoretical concepts, measures in any study must relate to each other consistently with theoretically derived hypotheses for the study to have external validity. If measures do not accurately reflect theoretically predicted relationships, then findings cannot link to the theory and generalization of findings is not possible.

(2) Relevance. Relevance refers to the degree to which the situation designed to test the theory adheres to the scope conditions of the theory under test. Scope conditions are statements that define the class of circumstances in which a theoretical proposition is applicable (Cohen 1989). Scope conditions allow findings to have external validity by specifying the universe of situations within which the predictions of the theory should be expected to hold.

It is only appropriate to generalize findings beyond the population or setting studied when the findings are based on a design that meets the scope conditions of the theory under test (Walker and Cohen 1985). Unfortunately, many theories do not explicitly state their scope conditions, but they are necessary if findings are to be generalized. If a theory has been supported in diverse tests, then we can generalize findings from one situation to another when both situations are described by the same abstract properties and satisfy the same conditions (Zelditch 1969).

(3) Reproducibility. No study, taken alone, can produce general knowledge. For this reason, the replication of research findings is an important step toward external validity. We gain increased confidence in a theory with each successful replication (Raman 1994). Reproducible findings are findings that are repeated under identical conditions, and this is what is usually meant by replication research. Will the study, repeated again in the same conditions, produce the same findings? If so, then the external validity of the findings is increased.

(4) Consistency. Consistency refers to the extent to which the observations in a study are consistent with each other and with the theory under test. In simple terms, consistency asks whether the findings of the study support the theoretical proposition(s) under test. The distinguishing feature of scientific research is that it is theoretical (Willer 1987), and findings increase external validity when they support the theory that the study is designed to test.

(5) Confirmatory status. Because findings from any study can only be generalized through theory, characteristics of the theory under test are the most important determinants of external validity. The confirmatory status of a theoretical proposition refers to the extent to which the proposition has been supported in numerous tests in diverse settings.

[4]Three of these five criteria (relevance, reproducibility, and consistency) draw heavily from Wagner's (1994) work on how empirical evidence bears on theoretical claims.

If a theory has been widely supported, then the external validity of any single test supporting a theory will be greatly enhanced. In the same way, if a test supporting a theory flies in the face of a number of tests that have refuted the theory, then we must say that the study has low external validity. Scientific knowledge is cumulative, and findings gain increased external validity with each successful theoretical replication.

These five criteria, then, can be used to assess the external validity of any test designed to test theoretical principles. Do the measures of the study accurately reflect theoretical constructs, does the test meet the scope conditions of the theory, are we confident that the findings would hold if the study were repeated, do findings support the theory under test, and has the theory escaped falsification in diverse settings? If these questions are answered affirmatively, then a study is likely to produce findings that can be generalized across settings and populations. In these ways, general knowledge can be produced.

EXTERNAL VALIDITY AS A THEORETICAL ISSUE

In current conceptions, external validity is treated primarily as a methodological concern. For instance, if covered in graduate social-science programs, external validity is usually taught in methods courses. Further, many discussions focus on how certain methodological approaches are higher in external validity than other approaches. However, any findings can only be linked to theoretical constructs at the level of the larger population in the context of theory. Further, the link between any methodology and new settings or populations can only occur through theory. Thus, in these terms, external validity is foremost a problem of theory.

By treating external validity as a primarily theoretical issue, I do not imply that the theory-testing function of most experiments makes them more externally valid than other types of investigative techniques. All investigations of any type are guided by decisions regarding how best to indirectly measure theoretical constructs. Many nonexperimental investigations are guided by well-specified theories, and some experimental investigations are guided by vague conjectures. However, whatever the methodological approach employed to test a theory, external validity must be addressed primarily as a theoretical issue.

There is only one instance in which it is appropriate to generalize findings beyond the setting or population studied. This instance is when findings are based on a design that meets the scope conditions of a theory (Walker and Cohen 1985). If a test meets the scope conditions of a theory, if the theory has been supported in diverse tests, and if measures accurately reflect theoretical constructs, then findings will be generalizable to other situations that also meet the scope of the theory. This includes settings and populations other than those studied.

If a study is designed to test a theory, and if it fits the conditions of the theory, it can have external validity and can tell us something about other populations and settings that also meet the conditions of the theory. If a study is driven by an overly vague or somehow inadequate theory, then the findings will not be generalizable and the test will lack external validity.

CONCLUSION

While treatments and definitions of the concept vary widely, external validity always refers to the production of general knowledge. Findings alone and without theory, however, cannot contribute to general knowledge. Thus, it is inappropriate to discuss

certain methodologies as being higher in external validity than others. Any study based on any methodology in the absence of theory is completely without external validity.

I propose that external validity is primarily a problem of theory and that it can be assessed only by an examination of the interplay between theory and methods. If a design meets the scope of the theory under test, if measures accurately reflect theoretical constructs, if findings support the propositions of the theory under test, and if the theory has been supported in diverse tests, then findings will have external validity. Other than creating situations that accurately reflect the conditions and concepts of the theory under test, no methodological procedures—including strict sampling techniques or highly realistic settings—can lead to an increase in external validity.

REFERENCES

Ahl, V., and T. F. H. Allen. 1996. *Hierarchy Theory: A Vision, Vocabulary, and Epistemology*. New York: Columbia University Press.

Aronson, E., M. Brewer, and J. M. Carlsmith. 1985. "Experimentation in Social Psychology." Pp. 441–86 in *Handbook of Social Psychology*, Vol. 1, *Theory and Method*, ed. G. Lindsey and E. Aronson. New York: Random House.

Babbie, E. 1989. *The Practice of Social Research*. 5th ed. Belmont, CA: Wadsworth.

Bass, A. R., and I. J. Firestone. 1980. "Implications of Representativeness for Generalizability of Field and Laboratory Research Findings." *American Psychologist* 35:463–64.

Bear, G. 1995. "Computationally Intensive Methods Warrant Reconsideration of Pedagogy in Statistics." *Behavioral Research Methods, Instruments, and Computers* 27:144–47.

Berg, B. L. 2001. *Qualitative Research Methods for the Social Sciences*. 4th ed. Needham Heights, MA: Allyn and Bacon.

Berger, J., B. P. Cohen, and M. Zelditch, Jr. 1966. "Status Characteristics and Expectation States." Pp. 29–46 in *Sociological Theories in Progress*, Vol. 1, ed. J. Berger, M. Zelditch, Jr., and B. Anderson. Boston: Houghton Mifflin.

Berger, J., B. P. Cohen, and M. Zelditch, Jr. 1972. "Status Characteristics and Social Interaction." *American Sociological Review* 37:241–55.

Berkowitz, L., and E. Donnerstein. 1982. "External Validity Is More than Skin Deep: Some Answers to Criticisms of Laboratory Experiments." *American Psychologist* 37:245–57.

Blalock, H. M., Jr. 1979. "The Presidential Address: Measurement and Conceptualization Problems: The Major Obstacle to Integrating Theory and Research." *American Sociological Review* 44:881–94.

Bollen, K. A., and P. Paxton. 1998. "Detection and Determinants of Bias in Subjective Measures." *American Sociological Review* 63:465–78.

Bornstein, R. F. 1990. "Publication Politics, Experimenter Bias, and the Replication Process in Social Science Research." *Journal of Social Behavior and Personality* 5:71–81.

Calder, B. J., L. W. Phillips, and A. M. Tybout. 1982. "The Concept of External Validity." *Journal of Consumer Research* 9:240–44.

Campbell, D. T., and J. C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research*. Chicago, IL: Rand McNally and Company.

Carmines, E. G., and R. A. Zeller. 1979. *Reliability and Validity Assessment*. Beverly Hills, CA: Sage Publications.

Cheuk, W. H., K. S. Wong, B. Swearse, and S. Rosen. 1997. "Stress Preparation, Coping Style, and Nurses' Experience of Being Spurned by Patients." *Journal of Social Behavior and Personality* 12:1055–64.

Chow, S. I. 1996. *Statistical Significance: Rationale, Validity, and Utility*. London: Sage.

Christensen, L. B. 1977. *Experimental Methodology*. Boston: Allyn and Bacon.

Cohen, B. P. 1980. "The Conditional Nature of Scientific Knowledge." Pp. 71–110 in *Theoretical Methods in Sociology: Seven Essays*, ed. L. Freese. Pittsburgh, PA: University of Pittsburgh Press.

———. 1989. *Developing Sociological Knowledge: Theory and Methods*. 2nd ed. Chicago, IL: Nelson-Hall.

Cook, T. D., and D. T. Campbell. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston, MA: Houghton Mifflin.

D'Aunno, T., T. E. Vaughn, and P. McElroy. 1999. "An Institutional Analysis of HIV-Prevention Efforts by the Nation's Outpatient Drug Abuse Treatment Units." *Journal of Health and Social Behavior* 40:175–92.

Devine, J. A., C. J. Brody, and J. D. Wright. 1997. "Evaluating an Alcohol and Drug Treatment Program for the Homeless: An Econometric Approach." *Evaluation and Program Planning* 20:205–15.

Dipboye, R. L., and M. F. Flanagan. 1979. "Research Settings in Industrial and Organizational Psychology: Are Findings in the Field More Generalizable than in the Laboratory?" *American Psychologist* 32:141–50.

Dobbins, G. H., I. M. Lane, and D. D. Steiner. 1988. "A Note on the Role of Laboratory Methodologies in Applied Behavioural Research: Don't Throw Out the Baby with the Bath Water." *Journal of Organizational Behavior* 9:281–86.

Dooley, D. 1990. *Social Research Methods*. 2nd ed. Englewood Cliffs, NJ: Prentice-Hall.

Easley, R. W., C. S. Maden, and M. G. Dunn. 2000. "Conducting Marketing Science: The Role of Replication in the Research Process." *Journal of Business Research* 48:83–92.

Esbensen, F., and W. D. Osgood. 1999. "Gang Resistance Education and Training (GREAT): Results from the National Evaluation." *Journal of Research in Crime and Delinquency* 36:194–225.

Esterberg, K. G. 2002. *Qualitative Methods in Social Research*. New York: McGraw-Hill.

Ferber, R. 1977. "Editorial: Research by Convenience." *Journal of Consumer Research* 4:57–58.

Foschi, M. 1997. "On Scope Conditions." *Small Group Research* 28:535–55.

Frankfort-Nachmias, C., and D. Nachmias. 1996. *Research Methods in the Social Sciences*. 5th ed. New York; St. Martin's Press.

Freese, L. 1980. "The Problem of Cumulative Knowledge." Pp. 13–69 in *Theoretical Methods in Sociology: Seven Essays*, ed. L. Freese. Pittsburgh, PA: University of Pittsburgh Press.

Gergen, K. J. 1978. "Experimentation in Social Psychology: A Reappraisal." *European Journal of Social Psychology* 8:507–27.

Gooderham, P. N., O. Nordhaug, and K. Ringdal. 1999. "Institutional and Rational Determinants of Organizational Practices: Human Resource Management in European Firms." *Administrative Science Quarterly* 44:507–31.

Gordon, M. E., L. A. Slade, and N. Schmitt. 1986. "The 'Science of the Sophomore' Revisited: From Conjecture to Empiricism." *Academy of Management Review* 11:191–207.

Greenwood, J. D. 1982. "The Relation between Laboratory Experiments and Social Behaviour: Causal Explanation and Generalization." *Journal for the Theory of Social Behaviour* 12:225–50.

Guthrie, D., and L. M. Roth. 1999. "The State, Courts, and Equal Opportunities for Female CEOs in US Organizations: Specifying Institutional Mechanisms." *Social Forces* 78:511–42.

Hanson, S. L., and S. A. Tuch. 1984. "The Determinants of Marital Instability: Some Methodological Issues." *Journal of Marriage and the Family* 46:631–42.

Henshel, R. L. 1980. "Seeking Inoperative Laws: Toward the Deliberate Use of Unnatural Experimentation." Pp. 175–99 in *Theoretical Methods in Sociology: Seven Essays*, ed. L. Freese. Pittsburgh, PA: University of Pittsburgh Press.

Hogarth, R. M. 1986. "Generalizing in Decision Research: The Role of Formal Models." *IEEE Transactions on Systems, Man, and Cybernetics* 16:439–49.

Hopkins, K. D., B. R. Hopkins, and I. Schon. 1988. "Mail Surveys of Professional Populations: The Effects of Monetary Gratuities on Return Rates." *Journal of Experimental Education* 56(4): 173–75.

Kruglanski, A. W., and M. Kroy. 1976. "Outcome Validity in Experimental Research: A Reconceptualization." *Representative Research in Social Psychology* 7:166–78.

Locke, E. A. 1986. *Generalizing from Laboratory to Field Settings*. Lexington: MA: D. C. Heath and Company.

Louviere, J. J. 1988. "Conjoint Analysis Modeling of Stated Preferences: A Review of Theory, Methods, Recent Developments, and External Validity." *Journal of Transport Economics and Policy* 10:93–119.

Lovaglia, M. J., J. W. Lucas, J. A. Houser, S. R. Thye, and B. Markovsky. 1998. "Status Processes and Mental Ability Test Scores." *American Journal of Sociology* 104:195–228.

Lynch, J. G. 1982. "On the External Validity of Experiments in Consumer Research." *Journal of Consumer Research* 9:225–39.

———. 1983. "The Role of External Validity in Theoretical Research." *Journal of Consumer Research* 10:109–11.

———. 1999. "Theory and External Validity." *Journal of the Academy of Marketing Science* 27:367–76.

Markovsky, B. 1994. "The Structure of Theories." Pp. 3–24 in *Group Processes: Sociological Analyses*, ed. M. Foschi and E. J. Lawler. Chicago, IL: Nelson-Hall.

Martin, M. W., and J. Sell. 1979. "The Role of the Experiment in the Social Sciences." *Sociological Quarterly* 20:581–90.

McGrath, J. E., and D. Brinberg. 1983. "External Validity and the Research Process: A Comment on the Calder/Lynch Dialogue." *Journal of Consumer Research* 10:115–24.
McTavish, D. G., and H. J. Loether. 2002. *Social Research: An Evolving Process*. 2nd ed. Boston, MA: Allyn and Bacon.
Meeker, B. Foley, and R. K. Leik. 1995. "Experimentation in Sociological Social Psychology." Pp. 629–49 in *Sociological Perspectives on Social Psychology*, ed. Karen S. Cook, G. A. Fine, and J. S. House. Needham Heights, MA: Allyn and Bacon.
Monette, D. R., T. J. Sullivan, and C. R. DeJong. 2002. *Applied Social Research: Tool for the Human Services*. 5th ed. Fort Worth, TX: Harcourt College Publishers.
Mook, D. G. 1983. "In Defense of External Invalidity." *American Psychologist* 38:379–87.
Nagel, E. 1961. *The Structure of Science*. New York: Harcourt Brace.
Neuliep, J. W., and R. Crandall. 1993. "Reviewer Bias against Replication Research." *Journal of Social Behavior and Personality* 8:21–29.
Pettigrew, T. F. 1988. "Influencing Policy with Social Psychology." *Journal of Social Issues* 44:205–19.
Popper, K. 1959. *Objective Knowledge: An Evolutionary Approach*. Oxford, UK: Clarendon Press.
Raman, K. 1994. "Inductive Inference and Replications: A Bayesian Perspective." *Journal of Consumer Research* 20:633–43.
Ross, C., and J. Mirowsky. 1999. "Disorder and Decay: The Concept and Measurement of Perceived Neighborhood Disorder." *Urban Affairs Review* 34:412–32.
Schmidt, F. L. 1992. "What Do Data Really Mean? Research Findings, Meta-Analysis, and Cumulative Knowledge in Psychology." *American Psychologist* 47:1173–81.
Sellers, R. M., S. A. J. Rowley, T. M. Chavous, N. J. Shelton, and M. A. Smith. 1997. "Multidimensional Inventory of Black Identity: A Preliminary Investigation of Reliability and Construct Validity." *Journal of Personality and Social Psychology* 73:805–15.
Shadish, W. R. 2002. "Revisiting Field Experimentation: Field Notes for the Future." *Psychological Methods* 7:3–18.
Shavelson, R. J., and N. M. Webb. 1991. *Generalizability Theory: A Primer*. Newbury Park, CA: Sage Publications.
Shrauger, J. S., and T. J. Schoeneman. 1979. "Symbolic Interactionist View of Self-Concept: Through the Looking Glass Darkly." *Psychological Bulletin* 86:549–73.
Sullivan, T. J. 2001. *Methods of Social Research*. Fort Worth, TX: Harcourt College Publishers.
Taylor, M. W. 1980. "The Noninfluence of 'Family Background' on Intellectual Attainment." *American Sociological Review* 45:855–58.
Thompson, K. M., D. Brownfield, and A. M. Sorenson. 1996. "Specialization Patterns of Gang and Nongang Offending: A Latent Structural Analysis." *Journal of Gang Research* 3:25–35.
Thye, S. R. 2000. "Reliability in Experimental Sociology." *Social Forces* 78:1277–309.
Top, T. J. 1991. "Sex Bias in the Evaluation of Performance in the Scientific, Artistic, and Literary Professions: A Review." *Sex Roles* 24:73–106.
Wagner, D. 1994. "The Growth of Theories." Pp. 25–42 in *Group Processes: Sociological Analyses*, ed. M. Foschi and E. J. Lawler. Chicago, IL: Nelson-Hall.
Walker, H. A., and B. P. Cohen. 1985. "Scope Statements: Imperatives for Evaluating Theory." *American Sociological Review* 50:288–301.
Webster, M., Jr. 1994. "Experimental Methods." Pp. 43–69 in *Group Processes: Sociological Analyses*, ed. M. Foschi and E. J. Lawler. Chicago, IL: Nelson-Hall.
———. Forthcoming. "Laboratory Experiments." In *Encyclopedia of Measurement*, chief ed. K. Kempf-Leonard. New York: Academic Press.
Webster, M., Jr., and J. B. Kervin. 1971. "Artificiality in Experimental Sociology." *Canadian Review of Sociology and Anthropology* 8:263–72.
Wells, G. L. and P. D. Windschitl. 1999. "Stimuli Sampling and Social Psychological Experimentation." *Personality and Social Psychology Bulletin* 25:1115–25.
Wiggins, J. A. 1983. "Family Violence as a Case of Interpersonal Aggression: A Situational Analysis." *Social Forces* 62:102–23.
Willems, E. P., and G. S. Howard. 1980. "The External Validity of Papers on External Validity." *American Psychologist* 35:387–88.
Willer, D. 1987. *Theory and the Experimental Investigation of Social Structures*. New York: Gordon and Breach Science Publishers.
Willer, D. E., and D. Heckathorn. 1980. "Cumulation, Explanation, and Prediction." Pp. 11–141 in *Theoretical Methods in Sociology: Seven Essays*, ed. L. Freese. Pittsburgh, PA: University of Pittsburgh Press.

Willer, D., and J. Willer. 1973. *Systematic Empiricism: Critique of a Pseudoscience*. Englewood Cliffs, NJ: Prentice-Hall.

Willer, D, and M. Webster, Jr. 1970. "Theoretical Constructs and Observables." *American Sociological Review* 35:748–57.

Winer, R. S. 1999. "Experimentation in the 21st Century: The Importance of External Validity." *Journal of the Academy of Marketing Science* 27:349–58.

Xiao, Y., and K. J. Vicente. 2000. "A Framework for Epistemological Analysis in Empirical (Laboratory and Field) Studies." *Human Factors* 42:87–101.

Zelditch, M., Jr. 1969. "Can You Really Study an Army in a Laboratory?" Pp. 528–39 in *A Sociological Reader in Complex Organizations*, ed. A. Etzioni. New York: Holt, Rinehart, and Winston.