The Concept of External Validity

BOBBY J. CALDER
LYNN W. PHILLIPS
ALICE M. TYBOUT*

Many researchers feel that external validity must be emphasized even in theoretical research. The argument for both a sophisticated and a common sense version of this contention is refuted in this paper. It is concluded that the very nature of progress in theoretical research argues against attempting to maximize external validity in the context of any single study.

Concepts of validity, originated by Campbell and Stanley (1966) and elaborated by Cook and Campbell (1976, 1979), have had a substantial impact upon how researchers think about their work. Four types of validity are at issue:

- **Statistical conclusion validity** refers to whether or not statistical inference of covariation between variables is justified.
- **Internal validity** addresses whether or not an observed covariation should be considered a causal relationship.
- **Construct validity** considers whether or not the operational variables used to observe covariation can be interpreted in terms of theoretical constructs.
- **External validity** examines whether or not an observed causal relationship should be generalized to and across different measures, persons, settings, and times.

Although most practicing researchers accept these concepts of validity and are familiar with factors that can threaten valid inferences of each type, several issues remain unresolved. These issues revolve around external validity and are of particular concern in theoretical research. Specifically, there is debate about the priority that should be given to achieving external validity when conducting a theory test. In addition, there is a lack of consensus about interrelationships between external validity and each of the other kinds of validity. Finally, there is a disagreement about the methods appropriate for enhancing external validity in theoretical research.

Two opposing positions regarding these issues can be identified. One view holds that external validity must be a priority in theoretical research. Research that is weak in external validity is not considered to provide an adequate test of theory. This view is grounded in the intuitive belief that valid theories must be shown to account for phenomena as they occur in natural settings. It may be bolstered by the more sophisticated argument that the absence of external validity implies an attendant lack of construct validity. Regardless of the specific form that this argument takes, proponents of this view contend that efforts to incorporate and examine the role of background factors (i.e., variability in persons, settings, and times not explicitly addressed by a theory) will improve the value of theory testing research.

A contrasting view is provided by Cook and Campbell (1979). They argue that when a researcher’s interest is mainly theoretical, the inference of external validity is of little concern (p. 83):

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

In accord with this perspective, external validity is not viewed as necessary for achieving a rigorous theory test and, indeed, may be sacrificed in favor of addressing threats to internal and construct validity. Controlling, rather than varying and examining, background factors is encouraged.

The purpose of this paper is to examine the concept of external validity in theoretical research. We begin with a

*Bobby J. Calder is Professor of Behavioral Science in Management and Professor of Psychology, Northwestern University, Evanston, IL 60201. Lynn W. Phillips is Assistant Professor of Marketing, Stanford University, Stanford, CA 94305. Alice M. Tybout is Associate Professor of Marketing, Northwestern University, Evanston, IL 60201.
critical analysis of the view that theoretical research would be enhanced by efforts to conduct externally valid theory tests. We then argue for the original Cook and Campbell position de-emphasizing external validity in theory testing. It is our contention that external validity is most appropriately addressed through theory development, rather than testing, and that it increases as knowledge accumulates and theories become more complete. (See Calder, Phillips and Tybout (1981) for a more detailed treatment of the relationship between theory development and theory testing.)

THE ARGUMENT FOR EXTERNAL VALIDITY IN THEORETICAL RESEARCH

The assertion that even a study designed to test theory must have some generalizability across measures, persons, settings, and times certainly has much superficial plausibility. It seems self-evident to many researchers, for instance, that having a random sample from some larger population is a better test than employing a convenience (e.g., student) sample. Yet much of this superficial plausibility disappears on close inspection. The reason is that theories are stated at a universal level. As long as a sample is relevant to the universe of the theory, it constitutes a test of that theory.

Kruglanski and Kroy (1976) illustrate this point about sampling by comparing the framing of applied and theoretical research questions. Consider an applied study interested in the incidence of consumers in a particular area Y who own a particular product X. An appropriate way of posing the research question is: "For any random sample of people Y, what is the percentage of ownership of X?" This question clearly calls for random sampling. In contrast, suppose the hypothesis is instead a theoretical one. Consumers receiving a communication causing them to be high on construct X are predicted to be more likely to buy a product than consumers low on construct X. It makes little sense to rephrase the hypothesis as: "For any random sample of people Y, consumers receiving a communication..." The theory is tested as well by a non-random as by a random sample.

Random sampling is not only unnecessary in theoretical research, but it may actually interfere with achieving a severe theory test. This is because it is likely to increase error variance and thereby reduce statistical conclusion validity. Parallel arguments apply to the use of random samples of measures, settings, and times. Thus, it appears that intuitive arguments for external validity in theoretical research can be dismissed. (See Calder et al. (1981) for a more extended discussion of why theoretical research often violates common sense conceptions of external validity.)

There is a more sophisticated argument for external validity that goes beyond common sense conceptions. This view is exemplified by some of the ideas presented in a paper by Lynch (1982). First we will consider the general form of the argument and then turn to the specific ideas of Lynch.

Sophisticated arguments for external validity accept the universal nature of theoretical research and reject common sense conceptions such as the automatic superiority of random samples. Nevertheless, such arguments contend that external validity inferences can be important for theory testing in the following way. In the context of planning a given study, let us distinguish between two kinds of variables. One is the constructs supplied by the theory under investigation. The other might be called background variables that are not identified by the theory. The contention is that a study lacks external validity when unidentified background factors exist that would, if included in the study, interact with the theoretical variables and thereby modify the effects obtained. The exclusion of such background factors limits the opportunity to falsify theory and therefore compromises the rigor of the theory test.

This argument would have this to say about samples, for example. If type of person is a background factor that interacts with theoretical variables, then it must be included in the research. If, for instance, the effects obtained for working head of households would differ from the effects for students, then both must be included in the study. This would be especially important if the interaction were disordinal. Conducting the study with only a student sample would lack external validity. More critically, it would lack external validity of the sort that weakens the test of theory.

The immediate problem with this argument is that it amounts to a counsel of despair. Even proponents such as Lynch (1982) recognize the extreme difficulty of achieving such external validity. A researcher must literally enumerate and anticipate all of the background factors that could interact with treatments. The researcher must go beyond the predicted causes of the effect (the theory) to include variables that on any intuitive grounds might change the effects obtained. Once included in the research design, analysis of these variables becomes an equally formidable task. The two-stage approach scheme proposed by Lynch only serves to illustrate how ad hoc the analysis would necessarily be.

Beyond the practical difficulty of assessing this sort of external validity is the logical impossibility of doing so. Clearly the set of background factors that could interact with treatments is infinite. Moreover, there is no a priori basis for even the most astute researcher to specify which of these factors will have an impact. Nor is there any logical way of even prioritizing these variables.

The history of science is replete with illustrations of this impossibility. Consider the case of Hertz's famous test of the prediction from electromagnetic theory that radio waves would be produced at the speed of light. What advice would the researcher concerned about external validity have had for Hertz? What background factors should have been included in the study? Different types of meters, different weather conditions, the size of the laboratory, the number of people present, the color of Hertz's lab coat? The single best source of guidance about this—i.e., electromagnetic theory—was by definition silent about all of these. An intuitively astute researcher might have guessed that the size
of the lab could be important (Hertz did not), but the list of possible factors is endless, and in the absence of theory, there is, ipso facto, little reason to choose among possibilities.

The sophisticated argument for external validity has a point to make. The point is that any study can turn out to be wrong. But this does not imply that a researcher can or should try to make a study infallible. As a practical matter, engaging in an obsessive quest for relevant background variables can only dilute the research effort. If it is obvious that a variable moderates the theoretical relationships of interest, then this variable should not be treated as a background factor in an effort to increase external validity—rather, it should be incorporated in the theory being tested. When the impact of background factors is not apparent, these variables are best held constant in any single study so that the researcher can concentrate on other validity inferences. Allowing random heterogeneity on background factors may undermine statistical conclusion validity. The alternative procedure of measuring and blocking on background factors substantially increases costs if any significant number of factors are considered and, as a logical matter, cannot ensure external validity, since omitted and potentially important variables will always exist.

Lynch's (1982) version of the external validity argument focuses specifically on the relationship of external validity and construct validity. While subject to the same practical and logical limitations noted above, Lynch's position is even more problematic. He contends that "if findings 'supporting' one's theory lack external validity, the theory lacks construct validity" (1982, p. 234). The rationale is that the only way to draw inferences about the degree of correspondence between measures and theoretical concepts (i.e., construct validity) is by an examination of the relationships among the empirical measures themselves, preferably across multiple contexts. As an illustration of this approach, a researcher adopting this philosophy would collect observations from two or more contexts within a theory's domain. Associations among the measure would then be examined to determine whether the predictions made by the theory are supported across contexts. If they are, the validity of measures, concepts, and theoretical hypotheses is supported. If they are not, the theory is deemed to lack external validity and, in turn, construct validity.

The problem with this approach is that it attempts to ascertain the validity of a theory's concepts, measures, and hypotheses by focusing solely on empirical measures. Yet associations among empirical measures need not correspond to the true causal or functional relations existing among theoretical concepts. When measurement, specification, or methodological errors are present, observed associations may reflect either underlying theoretical processes, one or more forms of error, or some combination of both. For these reasons, confirmation of a theory through purely empirical assessments provides limited insight. For example, demand character or other methods errors might cause observations consistent with a theory across experimental contexts. In this case, focus on associations among empirical measures alone would lead one to conclude erroneously that construct validity had been achieved. Similarly, rejection of a theory based solely on an examination of empirical association across contexts is not very helpful because one cannot easily determine whether the failure is due to inadequate measures, a misspecified theory, or both. Clearly, attempts to infer construct validity by a focus on empirical associations alone are not only futile but likely to be highly misleading.

To overcome this problem, it is necessary to employ a research methodology that (1) estimates the degree of correspondence between measurements and the concepts they represent; and (2) identifies and corrects for errors in measurement when testing nonobservational propositions. Researchers are currently seeking to develop procedures for these tasks in the context of both experimental and nonexperimental research (Bagozzi 1980; Bagozzi and Phillips 1982; Jöreskog and Sörbom 1979, 1982). Several recent applications of these procedures have produced findings that refute the proposition that the lack of external validity implies a lack of construct validity (Bagozzi 1981; Bentler and Speckart 1980; Phillips 1982; Phillips, Chang, and Buzzell 1982). These studies are characterized by the adoption of a research approach in which (1) the degree of correspondence between measures and concepts is explicitly represented—i.e., construct validity is explicitly modeled; (2) predictions concerning relationships between theoretical concepts are tested across multiple groups or situations—i.e., the external validity of the hypothesized causal relationships is explicitly tested across contexts; and (3) hypotheses are tested such that lawlike generalizations rather than raw empirical associations are the focus of inquiry. Common to each of these studies is the finding that a high degree of correspondence between measures and concepts is achieved across all contexts examined (i.e., construct validity is achieved in all contexts), while the predicted relations between theoretical concepts emerge in some contexts, but not in others (i.e., external validity is not achieved across all contexts). In some cases, the different pattern of results across contexts was anticipated by competing theories (e.g., Bagozzi 1981; Phillips et al. 1982). In other cases, the pattern was unexpected and theory development is required to account for the findings (e.g., Bentler and Speckart 1981; Phillips 1982). Regardless of whether or not the different results across contexts were predicted by theory, these studies clearly show that one can achieve construct validity but not external validity in the context of a single study. This contradicts Lynch's (1982) proposition that "if data supporting one's theory lack external validity, the theory lacks construct validity."

In sum, external validity and construct validity are unique. The question of whether measurement operations faithfully represent theoretical concepts (construct validity) is distinct from the question of whether relationships between theoretical concepts are generalizable across contexts (external validity). Construct validity is a necessary but not sufficient condition for external validity (Calder et al. 1981), but the reverse is not true. Further, neither construct
nor external validity may be ascertained by reliance on research philosophies that deal strictly with empirical observations and their associations. Unless these points are recognized, it will be impossible for researchers to distinguish between fundamental generalizations based on underlying theoretical processes and accidental generalizations arising from concomitant or spurious associations.

THE ARGUMENT AGAINST EXTERNAL VALIDITY IN THEORETICAL RESEARCH

The sophisticated argument for the importance of external validity in theoretical research has been shown to be practically infeasible and logically impossible. It is also possible, however, to argue against the importance of external validity on more constructive grounds. The essential point here is that research progress does not require external validity for any single research study.

There are many detailed views of the research process. However, most philosophers of science would agree in rejecting an inductive view in which research progress is attained by the accumulation of study on top of study. It is accepted that progress results more through the logic of refutation than through confirmation. We learn most by extending and developing theories when predictions have been falsified. This is Popper's (1959) falsificationist perspective; the reader is referred to Calder et al. (1981) for a more general description. It is important to note that it does not matter if individual researchers do not view their own activities in a falsificationist light. The issue is how progress occurs, not the intentions or insights of individual researchers.

Consider an example that is typical of research progress. Early work found that the attitudes of people who were highly involved with an issue were more difficult to change than those of less involved people. A more recent study by Petty and Cacioppo (1979) found that low involvement produced more attitude change only when message arguments were weak. Lynch (1982) gives this example to illustrate how the background variable, strength-of-message-arguments, should have been included in the early research. Because it was not, the earlier research lacked external validity. The test of theory was therefore inadequate in the original studies.

There is no doubt that the above pattern is a common one. But it is wrong to suppose that such examples show that external validity is critical to theoretical research: they indicate just the opposite. Despite the fact that no one study can hope to attain the sort of external validity at issue, the research process is one in which new studies contradict previous results and lead to new theory. This may be due to new hypotheses or to serendipitous observations. In either case, progress hinges on this very process of refutation and conjecture—new theories arising from rejected theories. No one study proves a theory, nor is a theory ever proven. The process simply confers scientific acceptance on a theory that has not yet, but may well be, disproven.

To imagine that this process, and consequent research progress, can be telescoped down into one study in the name of external validity is simply wrong. Background variables that interact with theoretical constructs will be found as they were found in the case of Hertz's radio waves. But they will only be found through the process of research in which they are incorporated into expanded and new theories. They will not be found by the researcher attempting to ensure external validity at a point in time.

CONCLUSION

Although we reject even the sophisticated argument for emphasizing external validity in theoretical research, this should not be taken to mean that we consider the issues discussed here to be closed. The subject of how advances in research come about, of how theoretical progress is made, is by no means well understood at either a philosophical or operational level. The fact that the practicing researcher cannot emphasize external validity inferences at a point in time should not be interpreted to mean that researchers should ignore the issue of how theoretical progress is made. Nor should it be concluded that improvements are impossible. We are, in fact, in real need of methodological guidance in this respect.

[Received August 1982.]

REFERENCES


