

External Validity: A Survey-Experiment Approach
 James S. Jackson and Michael S. Flynn
 University of Michigan

Campbell and Stanley (1966) have suggested that experimental results be evaluated from the perspectives of internal and external validity. The internal validity perspective asks how many rival hypotheses for the results of an experiment can be rejected. While Campbell and Stanley discuss more general threats, such as maturation and statistical regression, which the "true" experimental designs control for, Wiggins (1968) extends this notion to include plausible theoretical alternatives as well. For example, if we attribute the increase in the emission of a desired response by a subject to the automatic effects of a reinforcer, can we discount not only the possible (however implausible) effects of maturation, but can we also rule out the plausible alternative that the reinforcer communicated information about the correctness of the responses to the subject? External validity, on the other hand, concerns the generalizability of results. Would we observe the same relationship between behavior and the delivery of a reinforcer if we used subjects from another population, if we examined different levels and types of reinforcers, and tried to affect the emission rate of other kinds of behaviors?

The concepts of internal and external validity can be used to appraise research results produced by any method. If we examine two frequently used methods of research in social psychology, we observe a trade-off. The experimental method rates higher in internal validity. The specific causal attribution resulting from a well-designed and executed experiment is better grounded than that from a good survey. The well-designed and executed survey, however, gives the researcher greater faith that an observed relationship can be generalized to some larger population.

The observation that experiments and surveys have differing strengths and weaknesses with regard to internal and external validity is not particularly new (e.g., Swanson, 1951; Hovland, 1959; Blalock, 1964; Campbell & Stanley, 1966). Methods and procedures for bridging this gap, however, have not been developed. The results gathered by the two methods are treated as relatively disjoint bodies of knowledge. Hovland (1959) treats this situation in the case of attitude studies, and further illustrates that the "method" used to resolve discrepancies tends to be the "educated guess". One examines the studies, notes the myriad differences in sample, operations, etc., and tries to identify one or more conditions which might account for the different results. But this procedure is so broad and vague that it hinders rather than helps in the integration of the results from the two methods.

What can be done? The external validity of any study is always, in the final analysis, a question of assumption -- are investigators willing to accept the assumption in a particular case or not? Given all the possible threats to external validity in sampling, selection of manipulations, etc., there must be a final reliance on the assumption that results can be generalized. For each study, every plausible threat to external validity cannot be exhaustively tested; but there must be an examination, however tentative and limited, of the reasonableness of this assumption when the opportunity arises.

Perhaps the most frequent attack against the generalization of experimental results, whatever their substantive area, is the charge that their limited sampling base (the proverbial college sophomore) severely restricts their external validity -- sometimes to the point of rendering them trivial. Whatever the intellectual merits of this position for any particular set of experimental results, it remains a pervasive criticism and concern for both experimenters and non-experimenters (e.g., Ross and Smith, 1968; Campbell, 1957). The solution to the sampling problem is relatively straight-forward, however impractical it may be on a routine basis -- simply randomly sample the subjects who are then randomly assigned to the experimental conditions.

Last year we were selected to be co-principal investigators for the Detroit Area Study of the University of Michigan. This gave us the opportunity to participate in a household survey with a total of 574 interviews of approximately 30 minutes in length.

We have used the time available to us in each interview to conduct several experiments with these respondents as experimental subjects, and their homes as "laboratories". Our general purpose has been to examine the assumption of external validity as it relates to

sampling concerns. Consequently we attempted to replicate systematically previous experimental work in which four circumstances obtain. First, a restricted sampling base has been utilized. Second, there is consistency in the experimental results reported. Third, the experimental results have been relied upon to explain the behavior, in other settings, of people with different individual characteristics, such as education, race, SES, etc. might reasonably be expected to interact with the manipulated variables in such a fashion that generalization becomes a more tenuous assumption.

We have not restricted ourselves to the sampling dimension of external validity, but this is the unifying theme across the replications. Where appropriate we have attempted to examine the more idiosyncratic dimensions of external validity, such as selection of manipulations and measurements, and the particular events faced by the subject-respondents. Factors relevant to internal validity have also been built into these experiments -- they are not "mere" replications.

Method

Each of our respondents participated in two different experimental situations. The first was drawn from the general area of equity, and involved the respondent dividing a hypothetical sum of money among four participants in a research project. The task the participants supposedly engaged in (Physical, Social and Cognitive) as well as their race and sex were varied between respondents. The performance level of the participants and the amount of time they spent on the task were varied within subjects.

The second experiment was a verbal operant conditioning task. Respondents were asked to attach emotional labels to 54 photographs. The less frequently used emotional category on the first nine photographs were reinforced on a continuous basis for the remaining 45 photographs. Respondents were exposed to a satiation-deprivation manipulation, an achievement-no-achievement manipulation, and a dependency manipulation (three levels) on a between respondent basis.

Embedded within the reward distribution experiment were items asking the respondent to rate the ability, luck, effort and task difficulty of each supposed participant in the research. These questions were earlier asked about hypothetical successes or failures in "real life". These items provide information on the generalizability of a number of attribution experiments.

Finally, standard survey materials and questions were used to gather information relevant to the respondent's past history of reinforcement.

Approximately half of the interviews were conducted by students associated with the Detroit Area Study Research Practicum. These interviewers were randomly assigned to groups of 4-6 dwelling units (half-blocks). The remainder of the interviews were completed by professional interviewers on the staff of the Field Section of the Institute for Social Research. These interviewers were assigned by pairs to geographic areas, and then randomly assigned to half-blocks within their areas. This randomization of interviewers, of course, ensures that our large number of interviewer-experimenters does not increase and confound experimenter effects (McGuigan, 1963). After initial refusals, normal survey practice was followed in the assignment of follow-up interviewers. Approximately 70% of the interviews were completed by randomly assigned interviewers.

Discussion

What data will be brought to bear upon the assumption of external validity by these experimental replications? For a number of substantively important results, a broad range of plausible alternative hypotheses and threats to generalizability based upon individual characteristics of the subject sample can be examined. While the assumption of external validity for my particular of results will always be one that some will accept and other reject, this study should provide an assessment of the general reasonableness of the assumption. If, for example, in all the experiments there is a consistent difference in results by race, then there will be evidence that race is sufficiently a threat to external validity that investigators must include it in research designs, or accept appropriate limits to the generalizability of their results gathered with race ignored or held constant. On the other hand, if race consistently fails to have an effect, then those who object to cross-race generalization

will be on notice that research in a particular area will have to be produced to support their objections. In either case, the question should become less one of "scholarly disagreement" and more one of serious theoretical analysis and empirical research.

Of course, this assumes the generalizability of this particular sample, these particular experiments, and the results observed. This study, like any other, at some point must assume external validity; its merit is that through somewhat unusual methods and procedures it pushes the point at which the assumption must be made further back.

The overall pattern of results obtained for these experiments will be presented, focusing especially upon the effects produced by respondent characteristics. At the same time, these data can be brought to bear upon a number of questions of internal validity, specific to each experiment. Additionally, of course, this study will shed light upon the general feasibility of including experimental design and procedures within the context of household surveys, and hopefully encourage exploration and utilization of these techniques by other researchers.

References

- Blalock, H. M. Causal inferences in non-experimental research. Chapel Hill: The University of North Carolina Press, 1964.
- Campbell, D. T., and Stanley, J. Experimental and quasi-experimental designs for research. Chicago: Rand McNally and Company, 1963.
- Hovland, C. I. Reconciling conflicting results derived from experimental and survey studies of attitude change. American Psychologist, 1959, 4, 8-17.
- McGuigan, F. The experimenter: A neglected stimulus object. Psychological Bulletin, 1963, 60, 421-428.
- Ross, J., and Smith, P. Orthodox experimental designs. In H. M. and A. B. Blalock (Eds.) Methodology in social research. McGraw-Hill, 1968.
- Swanson, G. E. Some problems of laboratory experiments with small populations. American Sociological Review, 1951, 348-358.
- Wiggins, J. A. Hypothesis validity and experimental laboratory methods. In H. M. and A. B. Blalock (Eds.) Methodology in social research. McGraw-Hill, 1968.