1975

External Validity and Evaluation Research: A Codification of Problems

Ilene N. Bernstein
Indiana University - Bloomington

George W. Bohrnstedt
Indiana University - Bloomington

Edgar F. Borgatta
CUNY Queens College

This paper delimits and explicates threats to external validity particularly problematic in evaluation research. Five categories of factors are discussed: selection effects, measurement effects, confounded treatment effects, situational effects, and effects due to differential mortality. The paper focuses on pointing up specific ways in which each of the factors threaten generalizability and possible solutions to the methodological problems presented.

EXTERNAL VALIDITY AND EVALUATION RESEARCH

A Codification of Problems

ILENE N. BERNSTEIN
Indiana University
GEORGE W. BOHRNSTEDT
Indiana University
EDGAR F. BORGATTA
Queens College
City University of New York

Evaluative research is the application of scientific methods to the problem of assessing the effectiveness of an activity (or program) in attaining a desired goal. In the last decade there has been an increasing interest in evaluative research as a handmaiden to social policy (Weiss, 1970). It is thought that the results of evaluation research can provide a rational basis for decisions either to modify, terminate, or expand the ever-growing number of social action programs competing for public support.

Despite the commitment to and interest in furthering the development of evaluation research, the applications remain seriously wanting. Bern-

AUTHORS' NOTE: *An earlier version of this paper was presented at 1974 Conference on Evaluation Research Methodology, American Sociological Association Methodology Section, Loyola University, Chicago, Illinois. The authors are grateful for comments on the earlier draft by Dale Blyth, Robert F. Boruch, Peter J. Burke, Martin R. Frankel, and Sheldon Stryker. Author order is according to foot size.*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 4 No. 1, August 1975

stein and Freeman (1975), in a review of the methodological procedures used by federally funded evaluation researchers in fiscal 1970, concluded that the problem was not so much a lack of available methodological techniques as it was the lack of use of those techniques available. In examining the self-reported responses of 236 evaluation researchers, they found, for example: (1) only 25% of the researchers report their study made use of an experimental OR quasi-experimental design with randomization and a control or comparison group; (2) only 59% selected the sample population to be studied on a random basis; (3) only 50% observed a sample representative of the larger population to which they wish to generalize; and (4) only 35% characterized their research as largely quantitative. Clearly, the inference to be drawn is that evaluation research is lacking in the application of appropriate research designs, sampling procedures, and data analytic techniques. More specifically, much evaluation research seems to be plagued by serious problems in research design, making dissemination and utilization of its results problematic if not useless.

Campbell and Stanley (1963) distinguished problems of internal validity from those of external validity for various research designs. Internal validity refers to the degree to which a design allows one to rule out alternative explanations for the way in which a particular independent variable is causally related to the dependent variable of interest. By contrast, external validity refers to the degree of generalizability from one's study to some larger, hypothetical population of interest. While problems of internal validity are as important as problems of external validity (indeed, any threat to internal validity must also logically threaten external validity), this paper concentrates on external validity only, since it is our observation that this topic has received considerably less attention by evaluation research methodologists than have problems of internal validity. As long as evaluation research purports to function as a handmaiden to policy, it is essential that the results of such studies be valid for a variety of intended target populations in varied locales with varied staffs and varied subjects. This is especially true with the advent of national program experiments such as the Negative Income Tax Experiments and the Head Start Program. Resources are just not available to evaluate every action program and project currently under way. If the worth of evaluation research is to be recognized, care needs to be taken to ensure that the results of evaluation studies undertaken provide maximum utility for policy decisions, i.e., they need be as externally valid as possible.

Campbell and Stanley (1963) provided a checklist of factors that might threaten internal and external validity, and Bracht and Glass (1968) provided a more detailed list of factors that might threaten external validity. Delimited below is a list of factors that are specifically germane to evaluation, in that they may threaten the external validity of evaluation research findings. While neither exhaustive nor mutually independent, the factors are grouped into five categories: *selection effects, measurement effects, confounded treatment effects, situational effects,* and *effects due to differential mortality.*

SELECTION EFFECTS

Two major purposes of research are (1) to estimate the effects a set of treatments has on some prespecified set of dependent variables, and (2) to generalize these estimates from the sample studied to some larger target population. Unfortunately, the populations to which evaluation researchers wish to generalize often cannot be easily enumerated, or the expense of doing so would be prohibitive. Therefore, it often is not feasible to draw a probability sample of elements from the target population. Instead, the researcher must sometimes be content with a biased selection of observations, which means that to some extent, external validity will suffer. Four types of such selection biases are listed below, in order of the degree to which they threaten external validity.

SELF-SELECTION OF RESPONDENTS INTO TREATMENT AND CONTROL CONDITIONS

Obviously, if clients themselves determine whether or not to seek treatment, one never knows whether it is the treatment itself which is responsible for observed differences between the experimental and control groups, or whether other variables correlated with the selection of treatment versus control are responsible for the observed effects. While random assignment to experimental and control groups is the optimal procedure to follow, Rossi (1972) points out that many social action programs, once put into effect, are available on a self-selection basis to a larger client population. In this case the target population can be defined as including only those motivated to seek participation in a particular program. Specifically, one might randomly assign subjects to different

treatment conditions and evaluate which is the most effective *for those persons seeking treatment.* For example, persons seeking marital counseling might be randomly assigned to group versus individual counseling; the researcher would then be justified in concluding that for those who sought counseling, one method of therapy is superior. However, generalizing to some larger population of persons motivated to seek counseling would assume that this particular sample of counselees represents the larger target population (i.e., all persons seeking counseling). Without random sampling from this larger population, the degree of external validity is unknown. Quite obviously, if one found that one type of counseling improved marital relations for those seeking counseling, it could be a serious error to generalize that all couples having marital difficulties should seek counseling. Similarly, one could be in serious error in concluding that a campaign should be mounted to get persons with difficulties to volunteer for a program of counseling since this population would not be the same as the original set of volunteers.

In some cases, the use of samples of accessible persons is justifiable even though large-scale experimentation is not feasible, i.e., when one can devise control groups from subjects motivated to participate. For example, if a given Job Corps training program has a limited number of slots available and an excess of applicants, an excellent naturally occurring control group exists, composed of those who sought entrance into the program, but could not be accommodated, assuming entrance is barred by chance alone.

To reiterate, whenever subjects self-select themselves into the treatment conditions, both internal and external validity are jeopardized. However, for programs in which it is anticipated that only volunteers will seek treatment, it is possible to design research from which reasonable generalizations for that target population can be drawn.

SELECTION BY EXCELLENCE

Observational units are chosen because of presumed likelihood of demonstrating the hypothesized effects.

A common practice in evaluation research is the evaluation of programs selected because of their assumed excellence. The logic of this procedure is that such programs are the ones most likely to generate positive results. Therefore, observing them may reveal such factors as: how best to administer the program, what type of staff is conducive to program effectiveness, and what particular program components seem to precipitate

the greatest positive effects. However, the problem with such a selection procedure lies in drawing inferences from the results. If one is comparing excellent programs with nonexcellent programs, the specific program inputs which are causally related to the attainment of program outputs must be delimited in detail such that they can be replicated. Cain and Hollister (1972: 118) refer to this as the replicability criterion:

> It is sometimes argued by administrators that evaluations which are based upon samples drawn from any centers of a program are not legitimate tests of the program concept since they do not adequately take into account the differences in the details of individual projects or of differentiated populations. These attitudes frequently lead the administrators or other champions of the program to select, either *ex ante* or *ex post,* particular "pet" projects for evaluations that "really count." In the extreme, this approach consists of looking at the successful programs (based on observations of ongoing or even completed programs) and then claiming that these are really the ones that should be the basis for the evaluation of the program as a whole. *If* these successful programs have worked with representative participants in representative surroundings and *if* the techniques used–including the quality of the administrative and operational personnel–can be replicated on a nationwide basis, *then* it makes sense to say that the evaluation of the particular program can stand for an evaluation of the overall program. But we can seldom assume these conditional statements. After all, each of the individual programs, a few political plums notwithstanding, was set up because someone thought it was worthwhile.

Quite obviously, then, generalizable conclusions on program effectiveness cannot ordinarily be drawn when units are chosen on the basis of their presumed excellence. The assumption that one can replicate the staff, monitoring procedures, and so on in other locations is dubious at best. There is one exception to this situation. If it is indeed the case that (a) one has randomly chosen units in which the staff has followed the program guidelines to the letter, and (b) the clients served are representative of the target population, and (c) the researcher finds that under these optimal conditions the program has no effect, then it would seem safe to conclude that the program also would have no effect under less optimal conditions. However, this is a highly restrictive set of conditions.

SELECTION BY EXPEDIENCE

Observational units are chosen solely because of availability.

Bernstein and Freeman (1975: ch. 4) cite the case of an evaluator who selected his sample for the evaluation study on the basis of "persons

available and willing to talk." Such a procedure can be titled selection by expedience.

Ideally, in the best of all possible worlds, the evaluator would draw a random sample of persons, centers, or programs, from the population to which he or she wished to generalize. Except for a few large-scale programs, random samples of this sort are not usually feasible. Most action programs are operationalized in only one or a few locations. For example, income maintenance experiments are being carried out in New Jersey, Gary, Seattle, and Denver, and even these experiments vary somewhat from each other in the eligibility requirements for participation. Does the lack of a random sample of the nation's poor mean that inference from these experiments is impossible? Technically speaking, yes; practically speaking, no. If one has no reason to believe that the poor of New Jersey differ significantly from the poor of Los Angeles, then one can make a tentative generalization about the poor of Los Angeles, even though no one from Los Angeles was sampled. Such inferences will be erroneous to the degree that the poor in different areas differ in ways that interact with the treatment. However, simply because one does not have a random sample of the population to which he/she ultimately wishes to generalize does not mean generalizations should not be made. Cornfield and Tukey (1956) argued strongly for this position. They characterized inference as being a bridge with two spans linked by an island: an inference from one's observations to some population (even if it is not specific); and an inference from that population to some larger group of observations. They termed the first span the "statistical span"; the second, the "subject-matter span." For example, an evaluation researcher studying a random sample of welfare recipients participating in the Work Incentive Program (WIN) in 1968 in the Minneapolis-St. Paul metropolitan area, strictly speaking, can only generalize to that area during that time period (the statistical span); but, in fact, he/she and the agency also want to generalize to all welfare recipients in all parts of the country for all years after 1968 (the subject-matter span). Both types of generalizations are important. And in particular, when the subject-matter span is weak, i.e., when there is little knowledge as in the case with most programs, it makes sense to move the island closer to the subject-matter share, even though the statistical span is weaker as a result.

Some caveats are appropriate here. First, it is necessary to stress the tentative nature of any inferences, random sample or not. Thus, generalizations from nonrandom samples must be drawn much more cautiously; second, replication always increases one's confidence in a set of

findings. While the replication of results in a new sample from the *same* population increases one's confidence in the results, replication with samples from *different* populations provides even greater evidence for the original set of findings. For example, if cash transfers do not affect the incentive to work in New Jersey, and this finding is replicated in Gary, Denver, and Seattle, our confidence in this finding would increase greatly even though these cities were not randomly chosen. The fact that the programs have different eligibility criteria would strengthen our confidence in the finding even further, were it replicated in each experiment. Confidence in both positive and negative findings increases with this type of replication.

Problems occur when findings do not replicate from sample to sample. If samples are not random, it is possible that differences are due to systematic differences in the samples, or more precisely there is a sample-treatment interaction effect. If accessibility is thought to be related to variables that interact with the treatment, then problems in external validity seem certain and such samples should be avoided. To illustrate, if one has reason to believe that only agencies, schools, or centers with a history of innovation are volunteering, one should be extremely cautious in making generalizations to other observational units. Obviously, the better the case an evaluator can make (from census data, school records, agency files, and so on) that a nonrandom sample is not systematically different from the target population, the more confidence one can have in the external validity of one's findings.

NONRANDOM POSTTREATMENT MATCHING

Very often the evaluation researcher is called in to evaluate programs after the programs have been in effect for some time. Furthermore, often such programs do not include or provide for control or comparison groups. Thus, the researcher is faced with the seemingly impossible task of trying to draw conclusions about a program in the absence of a basis for comparison. A common technique is to pair each "experimental" subject with a selected "control" subject matched on variables which are assumed to be correlated with the dependent variable(s) of interest, in an attempt to estimate the independent effects of the action program. While the notion that "some control group is better than no control group" may be a truism, matching is a very poor substitute for randomization. First, when randomization is employed, the expected relationship between the potentially confounding variables is zero as is the relationship of each to

the dependent variable(s). But in a nonexperimental research situation it normally is not the case that potentially confounding variables are uncorrelated with each other. Furthermore, if one tries to match subjects on more than three or four variables, often it is very difficult to find matches. Indeed, as the number of variables for matching increases, experience indicates that the control group may be a highly idiosyncratic sample of the population from which it was drawn.

Thus while matching would seem to be a reasonable alternative to randomization, in practice, for most cases, it turns out to be a very poor substitute. As such, it becomes clear that the action and evaluation components must be done contemporaneously in order to avoid the use of designs that hamper the ability to draw valid conclusions about program effects.

To summarize, we have touched on how selection effects can jeopardize external validity in evaluation research. Under this heading four types of selection types were enumerated. The general conclusions are: (1) selection effects can seriously affect generalizability; (2) even in cases where selection factors are at play it may be possible to devise designs which allow for valid conclusions to be drawn; (3) even though the samples utilized are not random samples of the target population, generalizations may still be warranted when replications are made; and (4) many problems posed by selection biases could be avoided if the evaluation aspects of a given study are simultaneously designed with the design of the action program.

MEASUREMENT EFFECTS

The term "measurement effects" is used very broadly here to include effects due to (a) the unreliability and invalidity of measurement, (b) what Campbell and Stanley (1963) term the "reactivity" of some measures, and (c) interactions between measurement and other variables.

MEASUREMENT ERROR

Measurement error jeopardizes both internal and external validity. As is well known, when one uses unreliable measuring instruments (Lord and Novick, 1968), estimated relationships are biased, usually attenuating the relationships. In effect this means that when measurement error is present, generalization of effects will commonly be conservative. While underesti-

mates are generally more desirable than overestimates, it is clear that in evaluation research settings, underestimates can result either in termination of the action program because of presumed small effects, or an inflated cost-benefit estimate.

The solution to the problem of measurement error is not an easy one since few measures in the social and social psychological domains have high reliabilities. However, since it is known that estimates of effect are biased when measurement error exists, the evaluation researcher should (1) choose or design measures that have demonstrated high reliability, and (2) correct estimates obtained for unreliability in order to get an estimate of the true relationships.

It is sometimes argued that certain phenomena of interest to the policy maker simply are not quantifiable, or that the available measures do not capture the subtleties or complexity of the phenomenon. Certainly, some variables are more difficult to measure than are others, but if program effects are presumed to exist, they must be demonstrable. To suggest that one cannot measure some presumed effect is tantamount to saying that it belongs in a class of extraempirical variables. Such a position is clearly antiscientific since it does not allow for the falsifiability of one's hypotheses.

PRETEST SENSITIZATION

Pretest sensitization refers to the possibility that the administration of a pretest in and of itself might affect experimental results. For example, measuring public attitudes toward ex-convicts prior to an action program aimed at changing the public's view of ex-convicts may sensitize the sample to respond to the program in a way different than a nonpretested sample. Thus generalizations would hold only for pretested populations.

Campbell (1957) and Campbell and Stanley (1963) considered pretest sensitization a sufficiently serious threat to internal and external validity such that they made a strong argument for using a posttest-only design (with randomization) rather than the traditional pretest-posttest design with randomization. However, in a more recent publication, Campbell (1969) withdrew considerably from this position. A series of experiments by Lana and associates (see Lana, 1969) indicated that, across a wide variety of opinions and attitudes, either (1) there was no difference in experimental effects between pretested and posttested groups, or (2) where differences were found, it was shown that smaller changes occurred for the pretested than for the posttested groups—that is, if anything

pretested tended to result in under- rather than overestimates of effects. Based on these findings Campbell concluded that while pretest sensitization may logically jeopardize internal and external validity, the experimental evidence thus far suggests that the actual effects are small. These results suggest that no easy rule of thumb exists to indicate whether the researcher should use a pretest design or not, where generalizations will be to nonpretested populations. The costs of using a posttest-only design are (1) uncertainty whether the experimental and control groups are indeed equivalent after randomization, and (2) loss of pretreatment base-line data.

For certain programs, the population of interest will itself be a pretested one. For example, Anderson (1975) has noted that participants in a national income maintenance program probably would be required to fill out forms and questionnaires prior to receiving cash transfers, in much the same way as is required of persons participating in the current experiments. In this case, or in similar cases, the use of a pretest-posttest design could enhance the external validity of the experiment since the experimental design is isomorphic with the actual program that might be implemented.

POSTTEST SENSITIZATION

Bracht and Glass (1968) pointed out that the administration of the posttest may interact with the treatment, thereby producing results that would not be observed in a population that received the treatment, but was not posttested. They (1968: 464) argued that:

> treatment effects may be latent or incomplete and appear only when formally posttested in the experimental setting. In the natural setting where post-tests are absent, treatment effects may not appear for a want of a sensitizing post-test.

To illustrate, suppose one is interested in the effectiveness of TV spots in changing attitudes toward hiring the handicapped. Time is bought from randomly chosen TV viewing areas, and a posttest is administered both in areas where the spot was shown and where it was not. It may be that the effects of the spot are latent and incomplete until the respondents are actually asked their attitudes about hiring the handicapped. At this time the asking of the questions and the treatment itself combine to affect the answers provided.

Bracht and Glass suggested that in cases in which the experimenter believes that posttest measurement may itself affect the variable of

interest, one should try to employ unobtrusive measures (Webb et al., 1966). However, as Lana (1969) pointed out, the unobtrusive measures available are often unsuitable for many research projects of interest. However, there are some natural experiments in which unobtrusive measures can be meaningfully employed, such as Campbell and Ross's (1968) analysis of the Connecticut crackdown on speeding, and Glass's (1968) analysis of change in Germany's divorce law in 1900. Still, many if not most unobtrusive measures are of unknown reliability and validity.

INTERACTION BETWEEN MEASUREMENT AND INDIVIDUAL LEVEL VARIABLES

Sometimes cognitive and abilities tests measure different variables for certain subgroups of the sample; that is, there is an interaction between measurement and one or more individual level variables. For example, some measures are said not to be "culture free." In the abstract, it is impossible to say whether this is an important threat to external validity, but it is one that the evaluation researcher should consider in choosing measuring instruments. For example, it is clear that questionnaires are inappropriate measuring devices with illiterate or near-illiterate populations, and indeed, any instrument relying on verbal ability including an interview may be problematic. Similarly, the use of the English language with persons not knowing it well leads to equally obvious problems.

The disadvantaged, certain ethnic groups, and very young children do pose special measurement problems for the evaluator. At the same time, some of the arguments raised seem to be political in nature and follow rather than precede the publication of findings, especially negative findings for a popular program that was supposed to have led to some desired change. One of the criticisms leveled at the Westinghouse evaluation of Head Start was the use of instruments not developed for disadvantaged children in measuring cognitive and affective states. But as Williams and Evans (1972) noted, previous studies with these same instruments that showed positive results had rarely been questioned. If nothing more, the potential for such political responses to measures used places emphasis on the need for studies explicitly designed to assess the validity of measures.

OMISSION OF RELEVANT DEPENDENT VARIABLES

One claim sometimes made is that a set of negative findings are "invalid" because the study failed to measure all of the relevant dependent

variables. For example, Williams and Evans (1972) related that some critics of the Westinghouse evaluation of Head Start noted that only cognitive and affective variables were measured, and not measures of health, nutrition, and community objectives as well. If the Head Start programs were designed to affect these latter variables, the criticism would have been valid. However, reviews of the Head Start experience suggest that there was no single Head Start program, but instead a set of very different programs which varied from center to center. As several evaluation methodologists have noted (Freeman and Sherwood, 1965; Suchman, 1967; Hyman and Wright, 1967; Bernstein and Sheldon, 1975), adequate evaluation research depends on agreement of clearly specified program goals. Without this agreement (or acknowledgment of a lack of it) prior to executing the research, every evaluation may be invalid in the sense that someone with a vested interest in the program can later claim that the important goals were not measured.

There are two features characteristic of much of evaluation research that bear on this point. First, because evaluation research is often done in a political setting, different interest groups may hold different goals for the program, some of which may conflict with one another, or at least be unrelated to one another. The evaluation researcher who blindly assumes that the goals formally stated by the program director are definitive may be overlooking what others with political power anticipate for the program. Moreover, he/she may be overlooking, as well, the fact that the target population, the staff, and the agency staff funding the action program may also each have a set of goals in mind and not necessarily a set the same as those enunciated by the program administrator. To illustrate: a former high-ranking administrator of OEO once noted that Head Start was largely born out of political motives: OEO believed it would be popular with congressmen and their constituents. This observation is supported by the fact that Head Start has continued in spite of the negative evaluation it received from the Westinghouse Corporation. This suggests that someone's goals were being met, although they most certainly were not those specified and measured in the formal evaluation.

Second, and related to the above, is that the problem of specifying the goals of a particular program is particularly difficult because, unlike most research which begins with a dependent variable, or a set of dependent variables for which causes are sought, much evaluation research begins with an independent variable and asks what it causes. Typically, a social scientist states an interest in some y and asks which set of x's are causally related to it. In evaluation research, one begins with an x (the program)

and asks what y's it might affect. One can think of a host of variables that a Head Start program, or a negative income tax program, could affect. As such, it is hard to imagine that a priori all of the dependent variables of potential interest will be specified and measured. Thus, the problem results from the fact that programs often begin with the purpose of helping in some general way, but "how" is left largely undefined yet with implication that the how is self-obvious.

The implication of these two features is that valid inferences about the range of effects, or about all of the effects of a social experiment, are much more difficult to make than are inferences from the typical basic-science experiment, other things being equal. Unlike traditional research where the objectives of the research are specified with great care, the evaluation researcher needs to survey carefully those who will be affected by a program, e.g., administrators, staff, and clients, in order to determine the multiplicity of goals that a single program may be expected to serve.

The above discussion is not meant to suggest that findings associated with only some subset of goals for a program will be totally invalid. Rather, our intention is to point out that the structure of evaluation research often requires the researcher to do a prior study in order not to omit variables which reflect important areas of concern for agencies and groups associated with the program in some significant way.

LACK OF CORRESPONDENCE BETWEEN MEASUREMENT AND CAUSAL LAG

One of the difficulties in any research is knowing in advance how much of a time lag exists between application of a treatment and the manifestation of its effect. The effects might be immediate, or they may occur weeks, months, or even years later. Similarly, the effects may be gradual and continuous, or they may occur all at once. Obviously, if the measurement lag (the time between administration of the treatment and its measurement) does not correspond to the causal lag, the evaluation researcher will arrive at incorrect conclusions about the effects of the independent variables (Pelz and Lew, 1970). Hovland et al. (1949) have discussed the well-known "sleeper effect," in which armed forces personnel predisposed to an idea did not show attitude changes immediately after seeing a film, but did show change nine weeks later. Similarly, Borgatta and Evans (1968) showed that antismoking messages did not have an effect until a follow-up a year after the negative findings had been observed.

One evaluation that attempted to assess immediate, intermediate, and long-range effects of a program was the evaluation of the Encampment for Citizenship Program (Hyman et al., 1962). Using an elaboration of what Campbell and Stanley (1963) called a "patched up" design (see Table 1), this evaluation took measures directly preceding the program and immediately after the program to estimate the *immediate* effects ($0_3 - 0_2$, $0_8 - 0_7$, $0_{12} - 0_{11}$, $0_{16} - 0_{15}$ in Figure 1). To estimate the short-term stability of this effect, they compared $0_4 - 0_3$ to $0_3 - 0_2$, using $0_2 - 0_1$ as an estimate for change which would occur as a result of normal maturation. To estimate the *intermediate* effect of the program, they took measures two months after the campers had returned and compared them with the before-after measures: $0_9 - 0_7$ and $0_8 - 0_7$; $0_{13} - 0_{11}$ and $0_{12} - 0_{11}$; and $0_{17} - 0_{15}$ and $0_{16} - 0_{15}$. Last, to measure *long-range effects,* they obtained after-only measures for camp alumni for a period of nine years, and a four-years-after measure was taken on the original New York 1955 group. The use of the alumni follow-up not only extended the range to nine years after exposure, but also allowed the evaluators to chart any discernible patterns in changes as the time lapse since leaving the encampment program increased. Thus by comparing $0_{26} - 0_{11}$ to $0_{12} - 0_{11}$, $0_{25} - 0_{11}$ to $0_{12} - 0_{11} \ldots 0_{18} - 0_{11}$ to $0_{12} - 0_{11}$ they estimated the long-range effects of the program for the period from one to nine years after the camping experience.

In summation, we have discussed six types of measurement problems which can affect the generalizability of results: (1) measurement error per se, (2) pretest sensitization, (3) posttest sensitization, (4) interaction between measurement and individual level characteristics, (5) omission of relevant dependent variables, and (6) the lack of isomorphism between measurement and causal lags.

CONFOUNDED TREATMENT EFFECTS

In most experiments observational units are randomized into one of the treatment control conditions, and there is no ambiguity about which observational units have received which treatment. Unfortunately, with social experimentation in natural settings it is often difficult to determine what the treatment really is. Some subjects may be participating in several programs, and hence there is no uniform treatment (or set of treatments) for all subjects in a given treatment condition. In still other cases, subjects whose eligibility characteristics have changed (eligibility determines which

TABLE 1

Evaluation of Encampment Programs—An Example of an Elaboration Design

	Before Measures			After Measures			Years After								
Groups	6 weeks Before Program	First Day Of Arrival	Program	Last Day Of Program	6 Wks. After	2 Mo. After	1	2	3	4	5	6	7	8	9
N.Y. 1955	O_1^*	O_2	X_1	O_3	O_4					O_5					
N.Y. 1957	O_6	O_7	X_2	O_8		O_9									
N.Y. 1958	O_{10}	O_{11}	X_3	O_{12}		O_{13}									
Cal. 1958**	O_{14}	O_{15}	X_4	O_{16}		O_{17}									
Alumni 1946			X_5												O_{18}
Alumni 1947			X_6											O_{19}	
Alumni 1948			X_7										O_{20}		
Alumni 1949			X_8									O_{21}			
Alumni 1950			X_9								O_{22}				
Alumni 1951			X_{10}							O_{23}					
Alumni 1952			X_{11}						O_{24}						
Alumni 1953			X_{12}					O_{25}							
Alumni 1954			X_{13}				O_{26}								

*Only one-third of the campers were sent a questionnaire six weeks before arriving at the Encampment. This was done in order to obtain a measure of normal maturation over a six week period. Normal maturation then was the difference between $O_2 - O_1$.

**The California 1958 Encampment was added to vary the ecological setting. Whereas in New York (Riverdale) the program was set up like a total community, in California it was less self-contained.

treatment condition they are in) may "wander" from treatment condition to treatment condition throughout the life of the experiment. Finally, there are cases in which treatment effects are apparently specific only to subjects with certain characteristics. All of these effects can be grouped together and termed "confounded treatment effects."

LACK OF STANDARD TREATMENTS ACROSS SAMPLING UNITS

In our earlier discussion of selection of sample units by excellence, we noted that evaluation studies often attempt to assess the effectiveness of multiple programs by observing a sample of them. For example, the Head Start evaluation by Westinghouse examined a sample of 104 Head Start centers in order to assess the effectiveness of the Head Start Program in precipitating intellectual and social-personal development among first-, second-, and third-grade children. However, taking a probability sample of local programs provides no assurance that the programs were alike in terms of implementation. As McDill et al. (1972: 149) discussing Title I programs asserted,

> Most governmental poverty programs have earmarked federal funds to be used in an area of recognized need but have left the determination of the means and goals to be pursued almost entirely to those at the local level. Thus the individual programs "purchased" with these monies and the rationales behind them are diffuse, a point which must always be kept in mind when a national program is evaluated.

These authors noted further the wide variation in emphasis and services provided by the various Head Start projects.

One can conclude from the above that, when national programs are being evaluated, a random sample of local versions of that program provides no basis for generalizability about the effectiveness of a single program. Rather what seems appropriate is a design that carefully specifies in detailed ways program inputs and program outputs, and a list of local projects that share these specific definitions of inputs and outputs. Once that list of projects has been compiled, process measure evaluations should be examined for each, i.e., measures should be taken to assess whether or not the program has been implemented according to stated guidelines. A new list can then be formulated of projects that have indeed implemented the same set of program inputs for the purpose of attaining the same set of desired outputs. From that list, a probability sample can be drawn to

assess the effectiveness of programs with components ABC to effect changes XYZ. And those results can be generalizable to other programs with the same inputs. However, this assumes that project centers are randomly assigned to treatments, and further, that differential adherence to program guidelines is random. Selective factors related to the effects of interest may also determine the degree to which the guidelines are followed.

Alternatively, if one knows ex post facto how local projects differ from one another, and if one assumes that these differences are randomly distributed among the projects, then these differences can be treated as a set of treatments and the data analyzed using traditional analysis-of-variance techniques. The problem, of course, is that it probably is incorrect to make the assumption that the various project centers are randomly assigned in fact to the various treatments.

MULTIPLE TREATMENT EFFECTS

Campbell and Stanley (1963: 6) posited that multiple treatment effects are likely to occur when subjects have participated in multiple programs aimed at effecting change. Similarly, Bracht and Glass (1968: 438) asserted that multiple treatment interference precludes generalizations to populations which have not been subjected to the same sequential set of multiple treatments. In evaluation research, this problem is particularly severe since the research is so often aimed at assessing the effectiveness of programs for the socially and economically disadvantaged, a population often comprising persons participating in several social action programs.

In terms of external validity, the problem is generalizing to populations of persons who have or have not participated in multiple programs, and/or separating out the effects of participation in one program from those of having participated in more than one program. For example, persons receiving monies from the Negative Income Tax Experiment may also be participating in Model Cities Programs, Job Training Programs, and the like. The researcher measuring change on any particular variable can never be certain which program should be credited with precipitating the desired change. Also, it may be that participation in a number of programs may involve subtle selective processes on the part of clients disposed to change, or simply the participation in many programs may establish a "culture of change" for the clients.

It remains unclear whether simultaneous participation in multiple programs has selective, additive, or interaction effects. Even if changes are

noted, the population for which results are generalizable remains unclear. In order to help assess whether multiple treatments are occurring, clients could be (a) asked to respond to a set of items designed to assess multiple program participation, and (b) asked to report their history of participation in the program being evaluated. Unless one can make the assumption that subjects are randomly distributed among the various multiple treatment effects, there really is no good solution to the problem of generalizability. If, however, the assumption of random assignment seems reasonable, one can apply the analysis of covariance, although this assumes no interaction between the various multiple treatment effects and the experimental program variable(s). Equivalently, one can construct a dummy variable for each presumed multiple treatment effect and test for its significance within a multiple regression framework (Cohen, 1968; Kerlinger and Pedhazur, 1973).

Another variety of multiple treatment effects also can be noted. This one applies to programs that define the treatment condition to which one is assigned according to a set of eligibility criteria. For example, the New Jersey Income Maintenance Program was limited to married, male-headed families in five metropolitan areas in New Jersey and Pennsylvania. However, during the course of the three-year experiment, some experimental families became eligible for AFDC because of the loss of a husband due to death, desertion, divorce, or separation. Such families were required to report all AFDC income to the evaluators, but cash transfers to these families from the experiment were not reduced. Similarly, the Gary income program used criteria based on age, family composition, and income to define eligibility for participation in the program, but changes in these variables meant that the treatment for certain subjects changed with changes in the variables defining eligibility.

Again, if such wandering can be thought of as occurring randomly across treatment conditions, one might code "wanders" as a dummy variable and estimate the seriousness of the bias. The severity of the problem depends, of course, on the prevalence of the condition. If only a small percentage of the subjects change in these ways, the estimates of treatment effects will be only minimally affected. On the other hand, if this "movement" is considerable, one must question whether a set of discrete treatments existed at all.

INTERACTION OF INDIVIDUAL LEVEL VARIABLES WITH THE EXPERIMENTAL TREATMENT EFFECTS

When there are interactions between the treatment and the individual level variables, generalizability of inferences may be severely limited. Lubin (1961) noted two classes of interaction effects. In one case a treatment has opposite effects for different classes of persons (called disordinal interaction); in the other, the effects are in the same direction (ordinal interaction). Knowing that ordinal interaction exists is of importance since it suggests that the cost-benefit ratio differs for different subgroups of the sample. However, the problems for inference when disordinal interaction exists are patently more serious. Few published examples of disordinal interaction can be located, suggesting that it may be a relatively rare problem. Another possibility is that few researchers check for it. One exception is the work of Hunt and Hardt (1969) who, in a study of disadvantaged high school students, found that over a 21-month period the GPAs of the black Upward Bound students and their controls both decreased while the white students in both the experimental and control groups increased, although the increase was not significant.

While the number of individual level variables that might interact with a treatment is very large, it seems imperative that the evaluator at least consider as possibilities some standard sociological variables such as age, sex, race, education, and income. The last three virtually define what we mean by "disadvantaged." Since many social action programs are directed at the disadvantaged, one would want to be certain that at least these variables are not interacting with the treatment in a way which suggests that the treatment is effective only for the more "advantaged" in the sample.

In summation, we have reviewed three types of confounded treatment effects which might affect generalizability: (1) the lack of a single treatment across sampling units, (2) multiple treatment effects, and (3) the interaction of individual level variables with the experimental treatment. In the final section which follows, we treat a broad class of threats to external validity termed situational effects.

SITUATIONAL EFFECTS[1]

Threats to external validity may be posed by the fact that the experiment, be it natural or contrived, occurs in a particular context not

representative of all contexts to which one wishes to generalize. Hyman and Wright (1967), discussing evaluation research methods, suggested that one take care to consider the effects of the ecology, setting, and staff of social program experiments since it is not unlikely that these factors will interact in some significant ways with the treatment effects. The following represent factors which seem most likely to be problematic for the external validity of evaluation research studies.

STAFF EFFECTS

Evaluation researchers have long since recognized that observed effects may be dependent on the administrator, director, group leader, or therapist directly involved with program administration. To illustrate, the effectiveness of a therapy program for soon-to-be-released convicts may depend heavily on the charismatic abilities of the therapist(s), rather than on a particular therapeutic approach. Unless the personal qualities of the staff are treated as part of the specific program input, and findings are noted to be generalizable only to programs administered by staff with equivalent personal qualities, the results cannot be generalizable.

It is this "situational factor" that may account for the difficulty often noted in implementing a successful widespread effort based on an earlier successful demonstration program. That is, uniqueness may be associated with the effort of a staff that believes in the effectiveness of the program. However, the same ability, ideological commitment, and enthusiasm may not be present among those involved in a later effort for widespread implementation of the program. Cain and Hollister (1972: 117) made a similar point in stating that to focus on the characteristics of the staff may be to focus on the nonreplicable aspects of the program: "it has sometimes been stated that the success of a compensatory education program depended upon the warmth and enthusiasm of the teachers. In a context of a nationwide program, [however] no administrator has control over the level of warmth and enthusiasm of teachers."

Apart from the action program staff, the evaluation staff may pose a problem, although it is less likely since presumably they have no stake in either positive or negative results regarding program effectiveness. However, in some instances, it may be that the presence of the evaluator may create a particular environment, e.g., the staff may feel threatened by him/her, or the clients may be aware of the evaluators which may then affect the clients' behavior. This latter point is treated more fully in the next section.

Staff effects such as the above may be seen as analogous to the so-called "experimenter effects" that are treated in great detail in the social psychological literature of the past several years, e.g., McGuire (1969) and Rosenthal (1969).

HAWTHORNE EFFECTS

The Hawthorne effect refers to changes in the attitudes or behaviors of subjects in an experiment precipitated by the awareness of being observed. Most often the Hawthorne effect results in an increased effort on the part of the subject to try to do that which he/she perceives as desirous or socially acceptable, although the client can just as well try to perform badly if so motivated. The first case in which this effect was noted involved a study of factory workers in the Western Electric Company. The workers responded favorably to almost all experimental changes in the work environment, such as lighting and music. Roethlisberger and Dickson (1939) eventually concluded that their observation of an increased work effort could not be attributed to changes in light intensity or to the addition of music, since responses to these innovations were being confounded by the subjects' awareness of the experiment.

In evaluation research settings, knowledge of being a participant in an evaluation of an experimental program can produce similarly biased effects. This can be especially problematic if program participants recognize that the results of the evaluation may be used in the decision process which determines the fate of the experimental program. Moreover, it is sometimes reported that staff with a vested interest in the continuance of the program may play on this known reaction to experimental evaluation. That is, staff may use as the basis for "pep talks" and encouragement for increased effort such statements as, "Our group is being observed as an important experimental group and therefore, we should do our best to be the very best ever . . . such that in the future everyone can have. . . ." The problem is that the findings must be generalizable to other settings in which the experimental nature of the program may no longer be a factor. The question one is left with, then, is whether the same effects would occur in a nonexperimental setting where the participants are not being evaluated. However, one must not overestimate these effects either, since in the absence of an experiment the staff might still give such "pep talks" for equally important reasons, e.g., to get ahead themselves, because they do want to help the clients, or even to keep the program alive. In fact, one might argue that such factors

should be incorporated into every program if indeed they are found to function to persuade the client to perform in a way that results in desired behavioral or status changes.

Working on the identical problem, social psychologists have attempted various procedures aimed at ruling out the biases posed by "demand characteristics," demand characteristics being the effects of the subject's perception of his or her role in the experiment and/or his or her perception of the hypothesis being tested. Orne (1969: 144-145) stated:

> Because subjects are active, sentient beings, they do not respond to the specific experimental stimuli with which they are confronted as isolated events but rather they perceive these in the total context of the experimental situation. . . . The subject's recognition that he is not merely responding to a set of stimuli but is doing so in order to produce data may exert an influence upon his performance. Inevitably he will wish to produce "good" data, that is, data characteristic of a "good" subject.

Unlike experimenter effects, which can be studied and somewhat controlled for by varying experimenters, demand characteristics are not so easily uncovered, since demand characteristic effects depend partly on the subject's perception. Orne (1969) suggested three methods for dealing with demand characteristics; only the first is applicable for evaluation research. The first method is what he terms "postexperimental inquiry," i.e., inquiring into the subject's perception of the experiment's demand characteristics after it is completed. While this seems useful, importantly it does not eliminate the bias, but rather serves to document whether or not it is a problem worthy of concern. It is useful for evaluation research if one discovers that the perception of certain demand characteristics increases the likelihood of change in the desired direction. Given the purpose is always to discover some effective method of inducing desired change, the explication of demand characteristics can become a deliberate part of program inputs for future implementations.

NOVELTY EFFECTS

Bracht and Glass (1968) defined "novelty effects" as that which occurs when some new treatment is first introduced. In evaluation settings, the newness of the program may precipitate a response to the treatment that would not occur under more ordinary conditions. The possibility of this novelty effect further demonstrates the need for measures of the stability of effects over time. Previously, we referenced the patched-up design as

used by Hyman et al. (1962). It would seem that a design that provides comparative estimates of effects on replicated instances of the same program would be useful as a mechanism for reducing errors of inference posed by the bias of novelty effects. For example, if the effects attributed to the implementation of the encampment program were far greater in the first year (New York 1955) than in later years (New York 1957), then one could conclude that novelty effects, the effects of the newness of the program, were interacting with the treatment outcomes. Replication thus seems to be one useful way of controlling for novelty effects.

HISTORY

Just as the experimental setting may affect the outcome of the treatment program, so the historical time in which an experiment takes place may also affect the outcome of the treatment program. Many of the experimental social action programs recently evaluated were conceived and implemented during the late sixties, a historical period that clearly could not be characterized as typical.[2] It is not hard to imagine that the emotion-packed, liberation-reform tempo of that period probably affected the attitudes and behaviors of the socially and economically disadvantaged. And this might well have affected the way in which the disadvantaged population responded to program treatments aimed at social reform and social rehabilitation. The problem now is for decision makers to draw inferences from these evaluations and to generalize from them for the purpose of policy decisions about programs that will be implemented and executed in times not characterized by such political and social upheaval in, say, the quite different recession period of the 1970s.

One way of dealing with the threat posed to external validity by the interaction of history with treatment effects is to replicate the experimental treatment at various points in time. Since replication also helps to eliminate other threats to external validity we underscore its value. Bernstein and Freeman (1975: ch. 7), in setting forth a set of policy recommendations for evaluation research, suggested that experimental programs be implemented and evaluated long before relevant policy decisions need to be made. Such long-range preplanning would allow for replications over time of the experimental program in varied settings with varied staff and varied program inputs. Done this way, the culminating experimental program could be a program based on modifications made in accordance with systematic, ongoing evaluation. Furthermore, such a procedure would increase the likelihood of program effectiveness and

would allow for some control over sampling fluctuations, setting, history, and/or staff interactions with the treatment effects. However, this requires a willingness to commit funds for experimental program ressearch that, with only a few exceptions, is not characteristic of government policy thus far. The more common policy has been to try a program only after the need for it has become acute and thus long after the time for repeated experimentation has passed.

GEOGRAPHIC SETTING

> In the attempt to conceptualize a program, the evaluator may be led astray by the very term itself. He may think of the treatment and forget the context in which it is imbedded. . . . [Importantly] the staff and the program are contained within a site, and the ecology of sites often contributes to the effectiveness of programs and should [thus] be conceptualized by an evaluator [Hyman and Wright, 1967: 196].

The geographic site in which an experimental program takes place might well interact with the treatment to produce particular outcomes that might not occur were the program set in alternate sites. For example, in some programs subjects are removed from their natural environment and taken for treatment to an environment better suited to the execution of the treatment. The Daytop Village Drug Rehabilitation Program is a case in point. Hard core drug addicts from the New York metropolitan area were taken to a country estate overlooking Long Island Sound, where they lived for several years in a commune-type arrangement and participated in a self-examination, self-help therapy program. The program directors asserted that the core of the program was the self-examination, self-help therapy sessions. However, when a similar program was instituted in a tenement in New York's lower east side, program administrators soon realized that the country setting, because of its aesthetic pleasantness and/or lack of access to drugs, was an important component of the rehabilitation program. Apparently, they had made an error in generalizing the experience of the country setting site to similar programs set in different sites.

A similar finding was noted by Hyman et al. (1962) in their evaluation of the encampment program. By comparing the observations of participants in the New York program (located in an isolated residential site) with observations of participants in the California program (located in a large major city), they detected what appeared to be interaction effects between the ecology of the setting and the encampment program. Campers

in the New York program showed more marked changes on the desired outcome variables than did campers in the comparable California program. In both the Daytop experiment and the encampment experiment, geographic setting was suggested post hoc as the crucial variable accounting for the observed differences. However, in fact, differences may have been due to staff, procedural, or other factors. Once again, replicating the experimental program over varied geographical settings would help to sort out the effects of experimental sites.

EFFECTS DUE TO DIFFERENTIAL MORTALITY

Differential mortality effects stem from differential subject and program losses from social experiments. If the loss of observational units is different for the treatment and control groups, and the differences cannot be attributed to chance, then external validity is threatened. While estimation models exist for missing data when the attrition is assumed to be random across treatment and control conditions, they do not appear to be models applicable when the sample loss is assumed to be nonrandom. Unfortunately, it is the latter case, where attrition cannot be assumed to be nonrandom, that may be typical of much evaluation research.

DIFFERENTIAL ATTRITION OF SUBJECTS

In experiments involving desired services such as income maintenance, nutrition supplements, and the like, it is realistic to assume that differential attrition will occur between the experimental and the control conditions since the motivation to continue in the study is significantly lower for the control subjects. For example, Kershaw (1972) indicated that in the income maintenance experiments attrition indeed was higher in the control group than in the experimental groups, even though counterattrition methods were employed.

Minimally, it seems that evaluators need to determine what the correlates of differential attrition are in order to further the development of new counterattrition methods. Presently, the personnel of the Gary Income Maintenance Experiment are considering a simulation experiment to determine the effects of differential attrition on the estimates of interest. This appears to be a reasonable way to proceed; hopefully, their efforts will provide useful suggestions.

DIFFERENTIAL ATTRITION OF PROGRAMS

Generally, when one thinks of threats to validity posed by differential mortality, the reference is to attrition of subjects in an experimental situation. Because of the political nature of evaluation research, the attrition problem is often more complex insofar as the attrition may not only be of subjects, but of whole programs as well. For example, if a national evaluation of "X" poverty program is being conducted and a random sample is taken of local X poverty programs for the evaluation study, when local programs drop out of the sample because they have been canceled (for whatever reason) the external validity of the findings can be called into question. This is especially problematic if it is the case that canceled programs were characterized by a blatant lack of effectiveness, corruption, or the like. Thus, unless one can assume that sample mortality is randomly distributed, generalizability is limited.

CONCLUSION

Our purpose here has been to provide a codification and discussion of factors that preclude maximum utility of evaluation research findings by virtue of the limits they place on generalizability. As threats to external validity, these factors potentially stand in the way of policy makers, who need to draw inferences from individual studies so as to be able to decide rationally whether programs should be modified, terminated, or expanded. This list should be most useful to evaluation researchers as a checklist against which precautions should be taken in designing and executing evaluation studies.

Suggestions have been put forth to rule out some of the biases noted. However, not all of the suggestions will always be feasible nor have suggestions been provided for all of the potential threats. Our hope is that this illumination of factors will aid future efforts in experimentation with methods and models that control for threats to external validity. And that in so doing, the state of evaluation research methodology will be advanced.

NOTES

1. Bracht and Glass (1968) include many of the same factors as falling under the rubric of "ecological validity."
2. Of course, one can easily argue that no period is "typical."

REFERENCES

ANDERSON, A. (1975) Principal Investigator: Evaluation of Gary Income Maintenance Experiment. Personal conversations with author.

BERNSTEIN, I. and H. E. FREEMAN (1975) Academic and Entrepreneurial Research: The Consequences of Diversity in Federal Evaluation Studies. New York: Russell Sage.

BERNSTEIN, I. and E. B. SHELDON (1975) "Method of evaluative research," in R. Smith (ed.) Social Science Methods. New York: Free Press.

BORGATTA, E. F. and R. R. EVANS (1968) Smoking and Health. Chicago: Aldine.

BRACHT, G. H. and G. V. GLASS (1968) "The external validity of experiments." Amer. Educ. Research J. 5: 437-474.

CAIN, G. C. and R. G. HOLLISTER (1972) "The methodology of evaluating social action programs," pp. 109-137 in P. Rossi and W. Williams (eds.) Evaluating Social Programs. New York: Seminar Press.

CAMPBELL, D. T. (1969) "Perspective: artifact and control," pp. 351-382 in R. Rosenthal and R. L. Rosnow (eds.) Artifact in Behavioral Research. New York: Academic Press.

--- (1957) "Factors relevant to the validity of experiments in social settings." Psych. Bull. 54: 297-312.

--- and H. L. ROSS (1968) "The Connecticut crackdown in speeding: time-series data in quasi experimental analysis." Law and Society Rev. 3, 1: 33-53.

CAMPBELL, D. and J. C. STANLEY (1963) Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally.

COHEN, J. (1968) "Multiple regression as a general data analytic system." Psych. Bull. 70: 426-443.

CORNFIELD, J. and J. W. TUKEY (1956) "Average values of mean squares in factorials." Annals of Mathematical Statistics 27: 907-949.

FREEMAN, H. E. and C. C. SHERWOOD (1965) "Research in large-scale intervention programs," pp. 262-276 in F. G. Caro (ed.) Readings in Evaluation Research. New York: Russell Sage.

GLASS, G. V. (1968) "Analysis of data on the Connecticut speeding crackdown as a time series quasi-experiment." Law and Society Rev. 3, 1: 55-76.

HOVLAND, C. I., A. A. LUMSDAINE, and F. D. SHEFFIELD (1949) Experiments on Mass Communication. Princeton: Princeton Univ. Press.

HUNT, D. E. and R. H. HARDT (1969) "The effect of Upward Bound programs on the attitudes, motivation, and academic achievement of Negro students." J. of Social Times 25: 117-129.

HYMAN, H. and C. R. WRIGHT (1967) "Evaluating social action programs," pp. 185-220 in F. G. Caro (ed.) Readings in Evaluation Research. New York: Russell Sage.

--- and T. HOPKINS (1962) Applications of Methods of Evaluation: Four Studies of the Encampment for Citizenship. Berkeley: Univ. of California Press.

KERLINGER, F. N. and E. J. PEDHAZUR (1973) Multiple Regression in Behavioral Research. New York: Holt, Rinehart & Winston.

KERSHAW, D. (1972) "Issues in income maintenance experimentation," in P. H. Rossi and W. Williams (eds.) Evaluating Social Programs. New York: Seminar Press.

LANA, R. E. (1969) "Pretest sensitization," pp. 119-141 in R. Rosenthal and R. L. Rosnow (eds.) Artifact in Behavioral Research. New York: Academic Press.

LORD, F. N. and M. R. NOVICK (1968) Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley.

LUBIN, A. (1961) "The interpretation of significant interaction." Educ. and Psych. Measurement 21: 807-817.

McDILL, E. L., M. S. McDILL, and J. T. SPREHE (1972) "Evaluation in practice: contemporary education," pp. 141-185 in P. Rossi and W. Williams (eds.) Evaluating Social Programs. New York: Seminar Press.

McGUIRE, W. J. (1969) "Suspiciousness of experimenter 'intent', " pp. 13-57 in R. Rosenthal and R. L. Rosnow (eds.) Artifact in Behavioral Research. New York: Academic Press.

ORNE, M. T. (1969) "Demand characteristics and the concept of quasi controls," pp. 143-179 in R. Rosenthal and R. L. Rosnow (eds.) Artifact in Behavioral Research. New York: Academic Press.

PELZ, D. C. and R. A. LEW (1970) "Heise's causal model applied," pp. 28-37 in E. F. Borgatta and G. W. Bohrnstedt (eds.) Sociological Methodology. San Francisco: Jossey-Bass.

ROETHLISBERGER, K. and W. DICKSON (1939) Management and the Worker. Cambridge: Harvard Univ. Press.

ROSENBERG, M. J. (1969) "The conditions and consequences of evaluation apprehension," pp. 279-349 in R. Rosenthal and R. L. Rosnow (eds.) Artifact in Behavioral Research. New York: Academic Press.

ROSENTHAL, R. (1969) "Interpersonal expectations: effects of the experimenter 'o hypothesis," pp. 181-277 in R. Rosenthal and R. L. Rosnow (eds.) Artifact in Behavioral Research. New York: Academic Press.

ROSSI, P. (1972) "Testing for success and failure in social action," pp. 11-49 in P. Rossi and W. Williams (eds.) Evaluating Social Programs. New York: Seminar Press.

——— and W. WILLIAMS [eds.] (1972) Evaluating Social Programs. New York: Seminar Press.

SUCHMAN, E. (1967) Evaluative Research. New York: Russell Sage.

WEBB, E. J. et al. (1966) Unobtrusive Measures: Nonreactive Research in the Social Sciences. Chicago: Rand McNally.

WEISS, C. (1970) "The politicization of evaluation research." J. of Social Issues 26, 4: 57-67.

WILLIAMS, W. and J. W. EVANS (1972) "The politics of evaluation: the case of Head Start," pp. 247-264 in P. Rossi and W. Williams (eds.) Evaluating Social Programs. New York: Seminar Press.