The External Validity of Experiments[1]

GLENN H. BRACHT
GENE V GLASS
University of Colorado

Shortly after publication, Donald T. Campbell and Julian C. Stanley's "Experimental and Quasi-Experimental Designs for Research on Teaching" (1963) gained the status of a classic exposition of experimentation in education. There have been few attempts to extend this pioneering work, as might be expected of a work so comprehensive in conception and so brilliant in execution. Webb *et al.* (1966) produced a work similar in purpose—that being to identify sources of external invalidity which arise from the reactive effect of measurement. We know of no other published work building on Campbell and Stanley's chapter.

We feel that external validity was not treated as comprehensively as internal validity in the Campbell-Stanley chapter. Thus we have endeavored here to refine and elaborate on the sources of external invalidity identified by Campbell and Stanley and to propose and illustrate additional sources of external invalidity which merit attention.

The intent (sometimes explicitly stated, sometimes not) of almost all experimenters is to generalize their findings to some group of subjects and set of conditions that are not included in

[1] We would be remiss if we began without acknowledging a great debt to Donald T. Campbell and Julian C. Stanley for having inaugurated this discussion and, more personally, for having offered suggestions for improving an early draft of this manuscript. We also wish to thank Richard C. Anderson for many helpful suggestions.

the experiment. To the extent and manner in which the results of an experiment can be generalized to different subjects, settings, experimenters, and, possibly, tests, the experiment possesses *external validity*. However, one can identify a number of threats to external validity which cause the effects of a treatment to be specific to some limited population of people or set of conditions. These threats to external validity appear to fall into two broad classes: (1) those dealing with generalizations to populations of persons (What population of subjects can be expected to behave in the same way as did the sample experimental subjects?), and (2) those dealing with the "environment" of the experiment (Under what conditions, i.e., settings, treatments, experimenters, dependent variables, etc., can the same results be expected?). These two broad classes correspond to two types of external validity: *population validity* and *ecological validity*. The remainder of this paper is a detailed examination of the following threats to external validity:

I. Population Validity
   A. Experimentally Accessible Population vs. Target Population: Generalizing from the population of subjects that is available to the experimenter (the accessible population) to the total population of subjects about whom he is interested (the target population) requires a thorough knowledge of the characteristics of both populations. The results of an experiment might apply only for those special sorts of persons from whom the experimental subjects were selected and not for some larger population of persons.
   B. Interaction of Personological Variables and Treatment Effects: If the superiority of one experimental treatment over another would be reversed when subjects at a different level of some variable descriptive of persons are exposed to the treatments, there exists an interaction of treatment effects and personological variable.

II. Ecological Validity
   A. Describing the Independent Variable Explicitly: Generalization and replication of the experimental results presuppose a complete knowledge of all aspects of the treatment and experimental setting.
   B. Multiple-Treatment Interference: When two or more treatments

are administered consecutively to the same persons within the same or different studies, it is difficult and sometimes impossible to ascertain the cause of the experimental results or to generalize the results to settings in which only one treatment is present.

C. Hawthorne Effect: A subject's behavior may be influenced partly by his perception of the experiment and how he should respond to the experimental stimuli. His awareness of participating in an experiment may precipitate behavior which would not occur in a setting which is not perceived as experimental.

D. Novelty and Disruption Effects: The experimental results may be due partly to the enthusiasm or disruption generated by the newness of the treatment. The effect of some new program in a setting where change is common may be quite different from the effect in a setting where very few changes have been experienced.

E. Experimenter Effect: The behavior of the subjects may be unintentionally influenced by certain characteristics or behaviors of the experimenter. The expectations of the experimenter may also bias the administration of the treatment and the observation of the subjects' behavior.

F. Pretest Sensitization: When a pretest has been administered, the experimental results may partly be a result of the sensitization to the content of the treatment. The results of the experiment might not apply to a second group of persons who were not pretested.

G. Post-test Sensitization: Treatment effects may be latent or incomplete and appear only when a post-experimental test is administered.

H. Interaction of History and Treatment Effects: The results may be unique because of "extraneous" events occurring at the time of the experiment.

I. Measurement of the Dependent Variable: Generalization of results depends on the identification of the dependent variables and the selection of instruments to measure these variables.

J. Interaction of Time of Measurement and Treatment Effects: Measurement of the dependent variable at two different times may produce different results. A treatment effect which is observed immediately after the administration of the treatment may not be observed at some later time, and *vice versa*.

I. POPULATION VALIDITY

One of the purposes of a research study is to learn something about a large group of people by making observations on a rel-

atively much smaller group of subjects. The process of generalizing the experimental results from the sample of subjects to a population is known as statistical inference. The identification of this population to which the results are generalizable is treated in the next two sections.

A. Experimentally Accessible Population vs. Target Population

Kempthorne (1961) has distinguished between the *experimentally accessible population* and the *target population.* The former is the population of subjects that is available to the experimenter for his study. The target population is defined as the total group of subjects about whom the experimenter is empirically attempting to learn something. It is the group that he wishes to understand a little better and to whom he wants to apply the conclusions drawn from his findings. For example, an educator has discovered a new approach to teaching fractions to fourth graders. Probably he would like to conclude that his method is better for all fourth-grade students in the United States—the target population. However, he randomly selects his sample from all fourth graders in the local school district—the experimentally accessible population.

The experimenter must make two "jumps" in his generalizations: (1) from the sample to the experimentally accessible population, and (2) from the accessible population to the target population. The first jump, a matter of inferential statistics, usually presents no problem if the experimenter has selected his sample randomly from the accessible population.

In the previous example, the experimenter may have chosen all fourth-grade students in the state as his experimentally accessible population and randomly selected a sample of fourth-grade classrooms. Then the accessible population would probably be more like the target population and inference could be made to the target population with more confidence than in the first example. (However, the experimenter now has a problem in managing the research procedures and maintaining precise control over the treatment because the experiment is being conducted throughout the state.)

Kempthorne (1961) recommended a strict definition of the ex-

perimentally accessible population as an ingredient in the planning of an experiment. He advised that it is better to have reliable knowledge about restricted sets of circumstances (what happens in the school district) and to have the uncertainty of extending this knowledge to the target population (all fourth-grade students in the United States) than to define the experimentally accessible population so broadly as to be uncertain about inferring from the sample to the accessible population.

If the sample has not been *randomly selected* from some experimentally accessible population, the experimenter cannot generalize with probabilistic rigor to some larger group of subjects. In reality, his sample has become his experimentally accessible population. Cornfield and Tukey (1956), however, advocated the application of conclusions to a larger group than the sample. They encouraged generalization from the sample to a population "like those observed."

The second jump, from the experimentally accessible population to the target population, can be made with relatively less confidence and rigor than the first jump. The only basis for this inference is a thorough knowledge of the characteristics of both populations and how these characteristics interact with the experimental treatment. If the mean IQ of fourth graders in the accessible population is 115, can the experimenter generalize to a target population in which the mean IQ is 100? The answer depends, of course, on what finding one wishes to generalize and the relationship between the treatment variable and the characteristics of the target population.

The degree of confidence with which an experimenter can generalize to the target population is never known because the experimenter is never able to sample randomly from the true target population. Kempthorne (1961) pointed out that, even if we could draw a random sample from the target population, by the time the results were analyzed the target population would not be that which had been sampled. "Just how different it will be is a matter of inference about the processes which lead to the target populations. Such an inference is in my opinion impossible to validate in any strict sense" (p. 101). The relevant consideration, however, is not the absolute differences in the target population on two different occasions but how these differences interact with

the treatment variable. Kaplan's (1964, p. 20) comment, perhaps, illustrates the typical approach of the researcher to this type of generalization:

> How we can know that the future will resemble the past, and whether, indeed, some principle of "uniformity of nature" is even presupposed by science—such questions have exercised many philosophers of science. Yet scientists themselves—and surely behavioral scientists—would be quite content to have only as much justification for their predictions as we have for expecting the sun to rise tomorrow.

Three studies illustrate the importance of defining the target population to be like the accessible population. Friedman (1967) found that programed machine instruction was superior to traditional methods for teaching spelling to third graders. With second graders, however, the machine-taught group learned significantly less than the teacher-taught group. Although the difference in grade level appears to have affected the results of the two treatments, the findings may not be internally valid since the two treatment groups were matched on scores from a spelling pretest. In a study of the mediational process in concept learning, Gagné (1966) reported that seven-year-old children are able to switch rapidly from choosing a black card to the opposite (white) card whereas four-year-olds cannot. Incorrect generalizations may have been made by including only one age group in the experiment and defining the target population too broadly. In a study of near and far transposition with children of mental age ranging from 42 to 76 months, Kuenne (1946) found that children in this range of mental age showed no differences on the near transposition test. In far transposition, the three-year-olds scored at a chance level, but the six-year-olds showed practically 100% transposition. If the experimenter had chosen subjects in a more restricted range of mental age, perhaps different conclusions would have resulted. In each of the above studies, an externally invalid result would have been obtained if the experimenter had sought to establish a treatment effect with children of one age and then to generalize the finding across ages.

One of the sources of external invalidity that arises from generalizing from the experimentally accessible population to the target population is the "selection by treatment" interaction. This

occurs when seemingly "similar" studies give different results. A closer investigation of the studies may reveal that the accessible population of one study was accustomed to curricular innovations and experimentation, whereas the other was not. An illustration of the "selection by treatment" interaction is Brownell's (1966) comparison of two instructional programs in England and Scotland. Conflicting results from the two countries were resolved in the experimenter's mind when he discovered that the teachers and pupils in England were accustomed to innovation, but the teachers and students in Scotland were experiencing a new program for the first time. Thus the two experimentally accessible populations were not representative of the same target population, and the difference was relevant to the outcome of the study.

Maturational level is another obvious way in which accessible populations may differ. Kendler and Kendler (1959) found that fast learners responded faster to reversal shifts and that slow learners responded faster to nonreversal shifts. Thus the mediational S-R theory was a better predictor for fast learners, but the single-unit S-R theory was a better predictor for slow learners. These results probably apply only to this age group (kindergarten) because these children are in a transitional stage of development, i.e., some children function on a single-unit S-R basis while others make relevant mediated responses. Brownell and Moser (1949) compared two instructional methods (meaningful vs. rote) and two subtraction procedures (equal addition vs. decomposition) for teaching borrowing in third-grade subtraction. On the average, the meaningful method was better in all schools, but the rote method was relatively effective with the equal additions procedure in schools whose students had a better background of meaningful arithmetic experiences. Thus differences in arithmetic background in the schools included in this study did affect the pattern of results in the different types of schools.

The finding (Barker and Gump, 1964) of a relationship between size of high school and student participation and satisfaction certainly has implications for the external validity of research with high school students. Although the findings showed a strong relationship between school size and student behavior, there are problems in attributing a causal relationship to size

of school. One of the confounding effects in this study was the type of community in which the schools were located—small schools in rural communities and large schools in larger towns and cities.

See Campbell and Stanley (1963 pp. 189-190) for an excellent discussion and additional examples of the "selection by treatment" interaction as a source of external invalidity.

B. Interaction of Personological Variables and Treatment

Generalization is the ability to make general statements about the effect of some treatment. Interactions between the treatment variable and characteristics of the subjects, however, may limit the generality of the inference, depending on the type of interaction. Lubin (1961) has distinguished between ordinal and disordinal interactions for the purpose of determining whether one treatment can be prescribed for all subjects in the target population or whether different treatments should be prescribed for subjects who possess different measures of some personological variable. A statistically significant interaction, such as the one reported in Table 1, is ordinal when the lines which represent the effect of the various treatment levels across the levels of the personological variable do not cross (cf. Figure 1). Although such interactions lend to the meaningfulness of interpreting the data, they do not limit generalizability, i.e., one treatment can be prescribed for all levels of the personological variable.

When the interaction is statistically significant (cf. Table 2) and the lines cross (cf. Figure 2), further analysis is necessary before it is known if this interaction is ordinal or disordinal, i.e., the crossing of the lines is not sufficient evidence for the

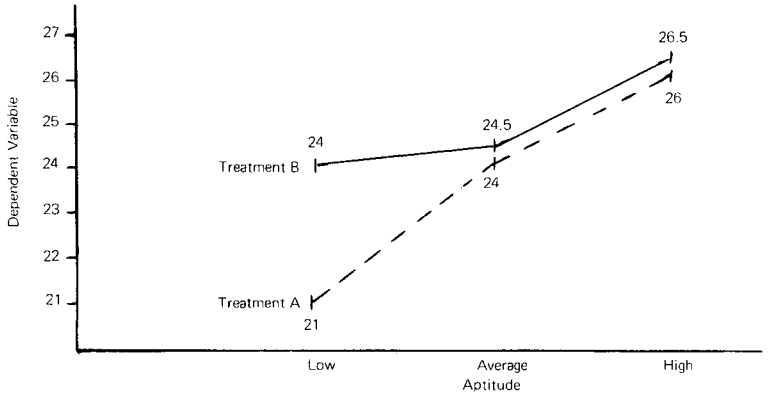
TABLE 1

Analysis of Variance for a Fixed Model Two-Way Classification (Two Treatment Groups by Three Levels of Aptitude)

SV	*df*	*MS*	*F*	*P*
Treatment	1	605	2.09	N.S.
Aptitude	2	920	3.17	$<.05$
Interaction	2	890	3.07	$<.05$
Within	144	290		

FIG. 1

Interaction of Treatments A and B with Three Levels of Ability.



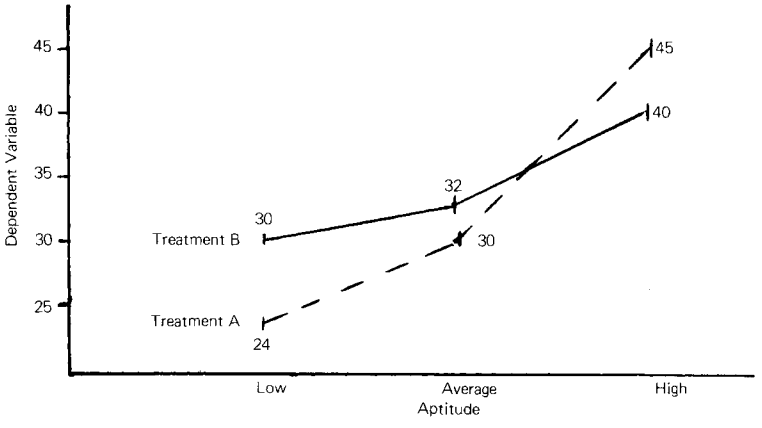
existence of a disordinal interaction. There exists an inferential statistical problem for detecting disordinal interactions which Millman (1967) also has discussed. In a two-way fixed model ANOVA design, one tests for the presence of *interaction,* not for *ordinal* interaction *as against disordinal* interaction. The typical *F*-test rejects the null hypothesis of no interaction with high power for either ordinal or disordinal interactions—at the stage of the *F*-test the two types of interaction are not distinguished. Curiously, researchers have divested themselves of their "inferential" scruples and designated a particular significant interaction *disordinal* merely if the lines of the graph cross at any point.

TABLE 2

Analysis of Variance for a Fixed Model Two-Way Classification (Two Treatment Groups by Three Levels of Aptitude)

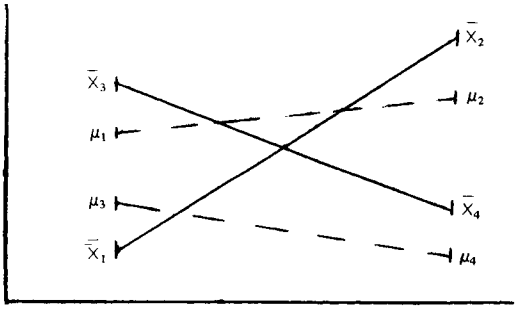
SV	*df*	*MS*	*F*	*P*
Treatment	1	425	1.23	N.S.
Aptitude	2	1215	3.52	$<.05$
Interaction	2	1325	3.84	$<.025$
Within	144	345		

FIG. 2
Interaction of Treatments A and B with Three Levels of Ability.



The objection we wish to raise against this procedure is apparent in Figure 3. *A statistically significant observed disordinal interaction may be only a chance deviation from an ordinal interaction in the population means.*

FIG. 3
Illustration of an ordinal *interaction in the population means giving rise to a* disordinal *interaction in the sample means.*



What is needed, of course, is a significance test which distinguishes ordinal and disordinal interactions in the population parameters. Such a test is not easily devised, and we can only suggest an imperfect possibility for the 2 x 2 design. Suppose that four population means in a 2 x 2 design are numbered μ_1, . . . , μ_4 as in Figure 3. The hypothesis to be tested is that $\mu_1 - \mu_3$ and $\mu_2 - \mu_4$ are both non-zero and differ in algebraic sign. Having observed non-zero differences $\overline{X}_1 - \overline{X}_3$ and $\overline{X}_2 - \overline{X}_4$ which differ in sign (suppose $\overline{X}_1 < \overline{X}_3$ and $\overline{X}_2 > \overline{X}_4$), one could perform individual *t*-tests of the null hypothesis for the two pairs of means against the alternatives $\mu_1 < \mu_3$ and $\mu_2 > \mu_4$. Assuming these two tests to be independent (which would be only approximately true unless the mean square within was partitioned into two separate parts, one for each test), each directional *t*-test could be run at a level of significance α_1 which produces the desired level of significance, α, for the pair of tests when substituted into the formula $\alpha = 1 - (1-\alpha_1)^2$.

To what extent do disordinal interactions (treatments with personological variables) occur in educational settings? Both Cronbach (1966) and Kagan (1966) expressed the belief that the discovery method has more value for some students than for others; some students will perform better with inductive teaching, and some will respond better to didactic teaching. Cronbach (1966, p. 77) contended that generalizations will have to be stated with several qualifications in the form: "With subject matter of this nature, inductive experience of this type, in this amount, produces this pattern of responses, in pupils at this level of development." He also expects discovery to interact more with personality variables than with ability.

Stolurow (1965) also hypothesized the interaction of personological variables with learning strategies and suggested that learning can be optimized by searching out these interactions. He has cited several studies as evidence that such interactions do exist. However, we have reviewed five of these studies (Eigen, 1962; Little, 1934; McNeil, 1962; Reed and Hayman, 1962; and Spence and Taylor, 1951) and found only one (Reed and Hayman, 1962) to contain a statistically significant interaction of treatment with a personological variable. Reed and Hayman found that a high school programed text covering English gram-

mar, punctuation, and capitalization was more effective for high-ability students, but that classroom instruction was better for low-ability students.

Unfortunately, Stolurow has interpreted interactions on the basis of a difference in means in the treatment levels and the difference in correlation of the personological variable with the dependent variable in the different treatment groups. This procedure is not legitimate for discovering differentially effective treatments and can lead to a misrepresentation of the data. Treatment groups A and B may differ on dependent variable Y, and the correlation coefficients between Y and a personological variable X may be quite different, though not differing in sign, in groups A and B without there being a disordinal interaction between X and groups A and B.

In his APA presidential address, Cronbach (1957) expressed optimism for discovering interactions between aptitude and treatment. He charged psychologists, both correlational and experimental, to invent constructs and form a network of laws which permits prediction. Interactions between organismic and treatment variables were hypothesized to form a part of this network of laws. Others, including Eckstrand (1962) and Tiedeman and Cogan (1958) feel that research has been devoted primarily to discovering general statements about the teaching-learning process and has failed to account for individual differences in learning. Edwards and Cronbach (1952) have recommended that the most promising organismic variables should be built into the experimental design so that gains can be assessed separately for each variable.

Stern *et al.* (1956) have suggested an interaction between performance in college and character. It seems likely to them that marked differences between the stereopath and other types of persons would be found in connection with academic performance. They predict that the stereopath will encounter particular difficulties in such areas as the humanities and social sciences where considerable emphasis is placed on abstract analysis, relativity of values and judgment rather than fixed standards, and an intraceptive rather than an impersonal orientation. They also predict that the stereopath is more likely to be found among those entering careers such as law, medicine, business, engineering,

etc., and less likely to be among those preparing for academic work or for the expressive arts.

The empirical evidence for disordinal interactions is presently far less convincing than the arguments stated above. Snow *et al.* (1965) reported that attitude toward instructional films, ascendancy, responsibility, numerical aptitude, verbal aptitude, past experience with entertainment films, and past use of college library instructional films interacted with instructional treatments (film vs. live demonstrations) in a college physics course. Six of these interactions were on an immediate-recall criterion, and the other two were on the delayed recall test. Although the lines crossed for all of these interactions, the authors of this paper have concluded from a further analysis of the data given in the original report (Snow, 1963) that only two of these interactions were disordinal. Our use of one-sided *post hoc* multiple *t*-tests at the 5% level of significance to test for differences between the treatment groups at certain levels of the personological variables must be interpreted as very liberal. Thus the probability of the occurrence of these two disordinal interactions is somewhat greater than the obtained level of significance (.05 and .025).

Kress and Gropper (1966) reported a significant disordinal interaction on a retention test (26-27 days after the treatment) of fixed tempo in program instruction and a student's characteristic work rate. When the fixed pace was slow, characteristically slow workers performed better, but when the tempo was fast, the fast workers scored better. The characteristically fast and slow workers were matched on intelligence. It should be noted that the illustration on page 277 of Kress and Gropper's article does not reflect the significant interaction which is shown in their Table 1. The illustration in their Figure 1 is based on two tempos, not four, and on two characteristic work rates, not fourteen. Although a statistical test for the data in Figure 1 was not performed, the authors of this paper feel that, even with a very liberal estimate, a disordinal interaction does not exist.

Hovland *et al.* (1949) used radio transcriptions to indoctrinate members of the Armed Forces during World War II. One experimental group heard both sides of the argument for some opinion (Program II) while the other experimental group heard only the favorable side of the same opinion (Program I). Program I was

more effective for the men who initially favored the opinion, but Program II was more effective for changing the opinion of the men who initially were opposed. Using educational background as a personological variable, they found that Program I was more effective for changing the opinions of the lower educational group (non-high school graduates), but Program II was more effective for the higher educational group (high school graduates). When both initial opinion and educational background were employed as variables in the analysis, the investigators concluded:

> Giving the strong points for the "other side" can make a presentation more effective at getting across its message, at least for the better educated men and for those who are already opposed to the stand taken. This difference in effectiveness, however, may be reversed for the less educated men and, in the extreme case, the material giving both sides may have a negative effect on poorly educated men already convinced of the major position taken by a program. From these results it would be expected that the total effect of either kind of program on the group as a whole would depend on the group's educational composition and on the initial division of opinion in the group. Thus, ascertaining this information about the composition of an audience might be of considerable value in choosing the most effective type of presentation (p. 215).

Since complete data were not presented, it is not possible to ascertain if these interactions are significantly disordinal.

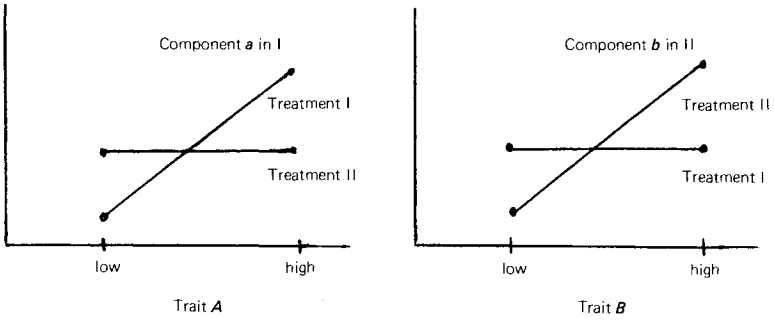
Cronbach and Gleser (1965) have extended the analysis of the data from Osburn and Melton (1963) to illustrate a disordinal aptitude-treatment interaction. Assuming linear regression and a normal distribution of the personological variable, they estimated the treatment means for various levels of aptitude. The results showed that a modern high school algebra course was superior for students in the upper three-fourths of the distribution on *DAT Abstract*, but the traditional course was better for the students in the lower fourth of the distribution. Although a test for the significance of this interaction was not reported, Cronbach and Gleser did conclude that the gain from differential placement would be too small to have much practical value.

Undoubtedly there are other experiments in which disordinal interactions of personological variables with treatment effects have been found, but the authors have no knowledge of them. It is

assumed that the frequency of such studies is small. The authors feel that the *molarity* (as opposed to *molecularity*) of both the personological variables and the treatments incorporated into many experiments *may* tend to obscure disordinal interactions which *might* be observable when both the variables and the treatments are more narrowly defined. Suppose that treatment I contains a component *a* which facilitates the performance of persons who are high on trait *A* and interferes with the performance of those low on *A*. And suppose that treatment II contains a component *b* which facilitates the performance of persons high on trait *B* and interferes with the performance of those low on *B*.

FIG. 4.

Two Opposing Disordinal Interactions of Traits and Treatment Components Present in the "Same" Personological Variable and Treatments.



If, in an experimental comparison of treatments I and II, factorially complex measures of the personological variable (containing traits *A* and *B*) and the dependent variable are used, the two disordinal interactions in Figure 4 may counterbalance each other and produce no disordinal interaction between the personological variable and treatments I and II. If this supposition is true, one would expect to find personological variable-treatment disordinal interactions more frequently with narrowly defined treatments and variables than in experiments employing broadly defined (complex) variables and treatments, e.g., intelligence and two curricula. Though this observation may suggest where to look for disordinal

interactions of personological variables and treatments, it may also suggest that searching for such interactions with treatments as necessarily complex as instructional curricula may be fruitless.

II. ECOLOGICAL VALIDITY

In addition to generalizing the results to a population of persons, the experimenter wants to say that the same effect will be obtained under other environmental conditions. Such a generalization assumes that the experimental effect is independent of the experimental *environment* (hence, the choice of the word "ecological").

Our separation of this paper into two parts should not be interpreted to mean that population validity and ecological validity are independent considerations for designing experiments and interpreting experimental results. It will be observed that threats to population validity may be the result of some source of ecological invalidity. There are many reasons why generalization is often restricted to a smaller population than may be desirable, and sources of ecological invalidity account for many of those reasons. Thus experiments which were cited in the first section of this paper may appear again in this section.

The following questions are illustrative of those which arise when considering the generalization of results to other ecological settings:

1. Are the treatment effects dependent to some extent on the use of certain audio-visual aids?
2. Is the length of treatment, both daily and over-all, a factor contributing to the effects?
3. Is the physical setting, e.g., size and shape of room, temperature, barometric pressure, etc., a factor in the treatment effects?
4. Are the treatment effects independent of the time of day?

Internal validity and population validity are not sufficient considerations in designing an experiment. Brunswik (1956, p. 39) suggested that "proper sampling of situations and problems may in the end be more important than proper sampling of subjects, considering the fact that individuals are probably on the whole much more alike than are situations among one another." The importance of the representativeness of educative conditions, even at

the sacrifice of some rigid control of the experimental treatment, has also been emphasized by Page (1958a) and others.

Exact replication of all experimental conditions is not desirable in educational research because then the theory-relevant aspects of the treatment cannot be separated from the extraneous variables. Campbell (1957) has suggested a "transition" experiment to discover the aspects of the treatment that really make the difference. The aspects of the treatment which are hypothesized to be independent of the theory would be varied in the multiple treatment groups, and the theory-relevant aspects of the treatment would be an "exact" replication in all groups. Thus there would be a replication of the experiment in varying ecological settings to determine what aspects of the treatment are causing the effect.

Millman (1966) emphasized that the experiment should be representative of a variety of conditions to which it may be desired to generalize the results. He also pointed out that by sampling across conditions one is more likely to detect any meaningful interactions which might exist. Concerning concepts related to learning theories, Hastings (1966) stressed the need to conduct research "across various content areas and with various age levels so that broad principles or generalizations can be made about new materials without having to carry on additional investigations."

Brunswik (1955, 1956) argued for the study of subjects in natural situations. This "representative design," as he called it, does not permit any artificial covarying or separating of variables either statistically or by experimental control. The total stimulus situation is studied only in its natural ecological setting and the data are then analyzed with correlational techniques. Cattell (1966) believes that the progress of psychology as a science must depend increasingly on non-manipulative designs. He maintained that "the great drawback of manipulation—apart from its being unusable with most important human learning—is that it risks disturbing, by 'side effects', the very process to be observed" (p. 8). May's (1953, p. 36) comments on research in psychotherapy also stressed the need for studying the subject in his natural setting:

> On the basis of the analysis in this paper, the crucial prerequisite for a new method is that the *irreducible unit for study be taken as the individual human being in a real-life situation.* The term "real" here means a situation in which the given human being is confronted

with some decision (in the particular sense this term is used previously in this paper) which involves in greater or lesser degree his own happiness and welfare and for which, therefore, he has some inescapable responsibility. . . . Confronting an individual with a conflict situation for experimental purposes in a laboratory does not produce a "real" situation. The person's "real" situation is not that of a human being facing the situation of, let us say, having to turn on the light at one signal and off at another, and then being confronted with the conflict situation of both signals at once, but rather that of a person cooperating with a friend or teacher in an experiment. His actually "real" state of mind may well not be the conflict from which, he knows, he will be entirely free when he leaves the laboratory in an hour—but rather that of curiosity or boredom or mild frustration and resentment that he is subjected to the experiment. . . . The cogency and value of these experiments will depend on how clearly the experimenter discerns the way in which the particular segment being isolated for the experiment fits into the total situation of the human beings involved.

Barker (1965) illustrated his distinction between psychologists as transducers (observing behavior without intervening or manipulating) and as operators (observing manipulated behavior) by reporting the results of two different studies of frustration in children. The results of the earlier experiments, where frustration was contrived for the subjects in a laboratory setting, were not replicated by Fawl (1963), who investigated frustration as it occurs in the natural habitat of children. Fawl observed that frustration occurs rarely in children, and when it does occur it does not have the behavioral consequences observed in the laboratory. The laboratory experiments did not simulate frustration as life prescribes it for children. The inescapable conclusion (Barker, 1965, p. 5) is "that psychologists as operators (O) and as transducers (T) are not analogous, and that the data they produce have fundamentally different uses within the science." He continued:

So far as behavior structure is concerned, O systems are, indeed, great simplifiers; the question is: Are they also great destroyers of essential attributes of psychological phenomena? One wonders, for example, how the properties which behavior units possess when they are lined up Indian file by an operator are modified when they occur in overlapping formation, as they so often do in the phenomena reported by T data (p. 7).

The reader of this section might conclude that we are attacking laboratory-based research and manipulative studies in the "field;" we are not. Both laboratory-based and field-manipulative research serve a vital function in contributing to our knowledge. In fact, for most educational research the non-manipulative designs do not appear to be as useful as the manipulative factorial design. However, these studies are not the basis for generalization to a variety of situations in which the human being normally interacts with his environment. Generalization to those situations which are not similar to the experimental setting is fraught with indeterminate risks.

A. Describing the Independent Variable Explicitly

Kempthorne (1961) stressed that the set of operations in an experiment must be replicable *to some degree.* This requires that the description of the set of operations must be sufficient to permit another experimenter to reproduce the set of operations to a reasonable extent. Such a detailed and complete description of the experiment is also necessary for the reader who must estimate to what extent the results can be generalized to other situations. If the description is not sufficient, the scientific value of the experiment is diminished.

Wittrock (1966) remarked that many empirical studies of the learning-by-discovery method have not clearly defined the independent variable. Experimenters sometimes report that learning by discovery is more effective than the traditional method, but they fail, unfortunately, to report in detail the procedures and activities of the experiment. Since the discovery method is interpreted differently by various people, it is not known what is being manipulated. Therefore, it is unknown to what situations the results can be generalized. Although not all research studies relate directly to some theory, the recommendations of Holzkamp (see Brandt, 1967) do suggest several considerations in designing and reporting the procedures of a study: (1) the length of the treatment should correspond to the hypotheses generated by the theory; (2) the experimental stimuli should cover the range embodied in the theory; and (3) the subjects should be involved in the experimental events to the extent called for by the theory.

The results of a study by Duncan (1964) clearly illustrate that

the length of the treatment can be an important aspect of the experimental situation. He found a negative transfer from day 1 to day 2 in learning paired-associate lists, but there was a gain in learning from day 2 to day 5. In addition, there was an ordinal interaction of length of treatment and conditions of learning. The difference between the paired-associate method and the response-discovery method was greater on day 1 than on day 5. Thus there was greater improvement over days in the response-discovery condition than in the paired-associate method.

B. Multiple-Treatment Interference

In some experiments two or more treatments are administered consecutively to the same subjects. In such cases it is possible to measure the effect of the initial treatment, but special problems arise in estimating the effect of subsequent treatments. It is not known to what extent responses to later treatments depend on the earlier treatments (Cox, 1958; Lana and Lubin, 1963). Cox pointed out that the effect of one treatment on the subsequent treatment observations usually is not represented by anything as simple as the addition of single constants (see pp. 20-21). In such cases either a special hypothesis must be set up appropriate to the problem or the treatments must be taken as a whole sequence of stimuli. (See Cox, chapter 13, for a discussion of cross-over designs which may be appropriate when each subject receives several treatments.)

Multiple-treatment interference may also result when the same subjects participate in more than one experiment. Weitz (1967) noted how the previous participation of subjects (college psychology students) in a study of guilt caused them to be overly suspicious (to the point where their responses had to be disregarded) of the innocent actions of a subsequent experimenter conducting a separate study on cognitive dissonance. Weitz asked, "Are certain psychological theories so dependent on the particular type of naïveté a person possesses that the theory has very limited generality?" The question is particularly relevant when asked of studies which employ deception of the subject by various means, e.g., experiments on conformity, persuasion, and aspects of cognitive dissonance.

C. *Hawthorne Effect*

A subject's knowledge that he is participating in an experiment may alter his response to the treatment. In such cases the experimental results cannot be accounted for entirely by the treatment effect. An additional factor, the perceived "demand characteristics of the experimental situation," must be hypothesized to account for the subjects' behavior to some extent. Orne (1962) has defined the "demand characteristics of the experimental situation" to include all the cues which convey an experimental hypothesis to the subject and so become significant determiners of the subject's behavior. The extent to which the demand characteristics affect the responses of the subject is a function of the extent to which the purpose of the experiment is clear to him. If the purpose is ambiguous, the effect will not be consistent and clear-cut.

Associated with the subject's perception of the experimental situation is the anxiety generated by participation in an experiment. Rosenberg (1965) reported that subjects who experience comparatively high levels of "evaluation apprehension" are more likely to confound the effects of the treatment.

There are several reasons why subjects may respond differently when they know they are participating in an experiment. Orne (1962) concluded that some subjects are motivated by a high regard for the aims of science and experimentation. They tend to hope that the study in which they are participating will contribute to a body of scientific knowledge and perhaps ultimately to human welfare. Thus the subject believes that the experimental task, whatever it is, is important, and his effort and discomfort are justified. Orne reported several experiments in which subjects who were aware of being in an experimental setting continued to perform boring, unrewarding, and nonsensical tasks. Not only did the subjects display remarkable compliance in carrying out the task, but they did so with a surprisingly high degree of diligence.

Social desirability is another motivating force for experimental behavior. A subject wants to do the "right thing" and be well evaluated, especially if he volunteered for the experiment. The perceived roles of subject and experimenter (Orne, 1962, p. 777) also influence a subject's impression of what is socially desirable.

A particularly striking aspect of the typical experimenter-subject relationship is the extent to which the subject will play his role and place himself under the control of the experimenter. Once a subject has agreed to participate in a psychological experiment, he implicitly agrees to perform a very wide range of actions on request without inquiring as to their purpose. . . .

The placebo effect, which is produced by the subject's faith in the efficacy of the experimental treatment, occurs with great regularity in medical studies (Rosenthal and Frank, 1956) and certainly is relevant to the external validity of research in psychotherapy and other areas of the behavioral sciences. Where the placebo effect does operate, evidence for the efficacy of the experimental treatment can be shown only if the improvement is greater than or qualitatively different from the placebo effect. An appropriate design for experiments of this type should include another form of treatment in which the subjects have equal faith, so that the placebo effect operates equally in both treatment levels.

The study by Page (1958) is an excellent illustration of controlling the reactive effect. He reported that the subjects were totally naive about his study of teacher comments and student performance: "In none of the classes were students reported to seem aware or suspicious that they were experimental subjects" (p. 175). Page achieved this by working through the seventy-four participating teachers, who administered the treatment and collected the data as part of the classroom routine, so that the students had no reason to know he existed or that they were participating in an experiment.

Cook (1967) studied the effect of direct and indirect cues on student achievement in elementary-school SMSG and conventional mathematics programs. The direct cue consisted of telling the students that they were participating in an experiment. The indirect cues included the introduction of a new curriculum (SMSG mathematics) and a different teacher for the math period. No significant differences were obtained between the presence and absence of either direct or indirect cues on student achievement at the end of one and two years of the study. On the basis of this study and an extensive review of the literature, Cook (1967) concluded that the Hawthorne effect probably does not contaminate experimental results in measures of academic achievement to the

extent that some researchers have claimed. However, Cook did call for additional research of the Hawthorne effect in academic and especially in the measurement of personality variables.

D. Novelty and Disruption Effects

A new and unusual experimental treatment, e.g., a curricular innovation, may be superior to a traditional treatment primarily because it is novel. Under conditions of diminished novelty, the superiority of the experimental treatment may disappear. Earlier in this paper, Brownell's (1966) study was cited as an illustration of potential population invalidity. The experimental samples in England and Scotland were not representative of the same population because the teachers were enthusiastic about different programs in the two countries. In Scotland the new program (A) was enthusiastically inaugurated into the schools, and the teachers became quite skillful in using it. In England, however, a new program (B) had previously been inaugurated, and the teachers had mastered its techniques. Thus Brownell's new program (A) was greeted with relatively less enthusiasm in England. Results such as this have led Cronbach (1963) to despair over comparative studies of competing curricula because it is never certain whether the advantage of one program is attributable to the innovative effect or the superiority of the curriculum. Scriven (1967), on the other hand, has recommended the matching of enthusiasm for the comparative evaluation of competing curricula and has suggested procedures for achieving this.

The antithesis of the novelty effect is the *disruption* effect which sometimes occurs with a new and unfamiliar treatment which is sufficiently different to the experimenter to render it somewhat less than effective during the initial try-out. After the experimenter has attained facility with the treatment, the results may be equal or superior to a traditional treatment. There is also the possibility that the novelty and disruption effects counterbalance each other in the same experiment.

An estimate of the novelty and disruption effects can be obtained by extending the experimental treatment over time. Even then, problems arise if the novelty or disruption of the treatment has led to the development of relatively permanent skills and traits in the experimenters and/or subjects.

E. *Experimenter Effect*

In the behavioral sciences, experimenters may unintentionally and to some indeterminate extent affect the behavior of their subjects. Rosenthal (1966), who has identified approximately eighteen different sources of experimenter effects which are relevant to ecological validity, has made a distinction between active and passive experimenter effects. Active effects are associated with unintended differences in the experimenter's *behavior*, e.g., encouragement, verbal reinforcement, and annoying mannerisms, that influence the subject's behavior. Passive effects are ascribed to the *appearance*, e.g., sex, age, and race, and are not associated with his behavior.

The experimenter effect may also reveal itself in the observation and recording of behavior. Boring (1962) observed that when an experimenter is making subjective observations, he may fail to see and report certain significant findings and fail to reach appropriate conclusions because of his theoretically based expectations. Kaplan (1964) also stressed the need for independence of experimental effect and experimenter when he asserted that intersubjectivity, i.e., a scientific observation that could be made by any other observer in the same situation, is the important methodological requirement for scientific acceptability.

Kintz *et al.* (1966, p. 224) have reviewed empirical studies of the experimenter effect and suggested that experimenters should form an independent variable in the experimental design.

> It is the present authors' contention that wherever an experimenter-subject relationship exists, the possibility also exists for E to contaminate his data by one or more of a multitude of conveyances. It appears that experimental psychology has too long neglected the experimenter as an independent variable. By relating some of the findings of clinical and social psychologists, as well as the few experimental studies to date, it is hoped that experimental psychologists will no longer accept on faith that the experimenter is necessary but harmless.

F. *Pretest Sensitization*

In experiments where a pretest has been administered, there is the possibility that the experimental effect is really a confounding of the treatment effect and the sensitization to the treatment.

Only if the design of the study permits him to conclude that there was no pretest effect can the experimenter generalize his findings to situations where a pretest will not be administered.

The bulk of the evidence for the pretest effect deals with attitudes and opinions. Campbell (1957) reported that Paul Lazarsfeld followed up on the United Nations Information Campaign in Cincinnati (Star and Hughes, 1950) and found that the group which was interviewed before the campaign showed significant attitude changes, a high degree of awareness of the campaign, and important increases in information. The pre-interview sensitized the persons to the topic of the U.N. and made the information campaign effective only for that group. Two consequences of interviewing, mental stimulation and a clarification of view (Crespi, 1948), support the evidence that the interview sensitizes subjects to a subsequent treatment.

Nosanchuk and Hare (1966) found that subjects who completed a pretest questionnaire were sensitized more to topics of current interest than were subjects who read descriptive statements about the topics. The measure of sensitization was the number of times a subject recognized words and phrases which were related to the concepts included in the pretest questionnaire.

The effect of the pretest on changes in attitude for both salient and non-salient topics has been investigated by Nosanchuk and Marchak (undated). Experimental subjects who completed a pretest questionnaire showed less change in attitude over a six-day period on a semantic differential than did both control groups. One control group read descriptive paragraphs about the same issues presented to the experimental group, and the other control group completed a pretest questionnaire and read descriptive paragraphs about issues unrelated to the topics in the other two groups. The control groups did not differ from each other, providing support for the hypothesis that a pretest format has a greater sensitizing effect than the reading of a descriptive paragraph. However, Lana (1959a, 1959b) found that the pretest did not affect attitude change either when the topic was of relatively little interest (vivisection) or of great concern (ethnic prejudice). Since there were twelve days between the pretest and the treatment in both studies, it is possible that the passage of time decreased the pretest effect. In several later experiments, Lana

(1966) found that the pretest groups showed significantly less change in opinion when presented with pro and con arguments than did the no-pretest and disguised pretest groups. Cognitive dissonance has been suggested (Nosanchuk and Marchak, undated) as an explanation for these results, i.e., subjects strive for consistency in attitudes where they have made a prior commitment.

Windle (1954) reviewed forty-one studies in which there was a test-retest with personality inventories and suggested that the interval of time between pretest and post-test may be related to the change in response. The tendency for better adjustment on the retest appeared mostly in studies where the interval between tests was less than two months.

Three studies are relevant to pretest sensitization when the dependent variable is a measure of academic achievement. Lana and King (1960) administered the pretest to male college students as a learning task, i.e., the pretest group read a summary of a film and then immediately recalled the summary. The no-pretest groups read the same summary but were not asked to recall it. Twelve days later the film was shown and all groups were asked to immediately recall the story as completely as possible. The results showed that the pretest had an effect on the post-test scores beyond the treatment effect.

Solomon (1949) found that a spelling pretest had a depressive effect on training in spelling. However, it appears from the report of this experiment that the disruption of the classroom routine by sending the pretest group out of the room during the pretest, thus creating a somewhat unnatural situation, may be a source of external invalidity for this finding. In addition, the groups were roughly matched in spelling ability by means of teacher judgments, a possible source of internal invalidity.

Entwisle (1961a) reported that there was no difference between pretest and no-pretest groups who were matched on IQ. She used fourth graders who were taught the state locations of large U.S. cities by a fixed tempo projection of slides in three twenty-minute training sessions which were scheduled for two, five, and seven weeks after the pretest. The post-test was administered one week after the third training session. A later re-analysis of these data and a modified replication of the experiment (Entwisle,

1961b) showed a three-way interaction of the pretest, sex, and IQ factors. However, the matching of subjects in the groups, the large number of statistical tests which maximize the likelihood of a "chance" significant interaction, and the nesting of IQ within teachers which confounds the interaction effect are sources of internal invalidity for these results.

The design of an experiment by Daw and Gage (1967) is conspicuously excellent for its control of pretest sensitization. They found that teachers' ratings of their ideal and actual principals did not affect a second rating of their actual principal either six or twelve weeks later.

The results of empirical investigations of pretest sensitization indicate that the effect is most likely to occur when the dependent variable is a self-report measure of some aspect of personality, attitude, or opinion. The pretest effect on academic achievement is apparently less prevalent, but the results are inconclusive since the studies which have been conducted are not representative of experimental situations where it usually is necessary to use a pretest.

*G. Post-Test Sensitization**

The possibility exists that a treatment effect will arise only if a post-test is administered. For example, an elementary school science curriculum may attempt to teach some subtle mechanical concept. Conceivably, the act of administering a mastery test at the conclusion of instruction to compare the experimental and control groups could provide a crucial opportunity for the student to acquire the concept (perhaps because of a fortuitous wording of the test questions or the illustrations employed). The same instructional program might not result in a mastery of the concept in the normal school curriculum in which the post-test of the experiment is not administered.

As a second example we can alter Campbell and Stanley's illustration of pretest sensitization of an anti-Semitism questionnaire administered in a study of the effects of the movie *Gentlemen's Agreement* on anti-Jewish prejudice. Suppose that a group of *Ss* is randomly divided into two halves, one of which saw

* We are indebted to Mr. Scott Harrington for having pointed out this possible source of external invalidity.

Gentlemen's Agreement and the other of which did not. A post-test of anti-Semitism is to be made by means of a self-report attitude inventory. (Because of random assignment to experimental and control groups, no pretesting is necessary.) When confronted with the post-test, the members of the experimental group, who saw the film, have sufficient time to be affected by the film's message while responding to the inventory. The strength of the effect would depend somewhat on their ability to reconstruct details of the film in their minds; it may be that portions of the film were subtle and escaped the attention of the unsensitized experimental subjects.

The point being made here is that treatment effects may be latent or incomplete and appear only when formally post-tested in the experimental setting. In the natural setting where post-tests are absent, treatment effects may not appear for want of a sensitizing post-test. For example, the anti-prejudice message of *Gentlemen's Agreement* may never be communicated to the casual movie-goer who receives neither pretest nor post-test. In experiments where post-test sensitization may effect the measurement of the treatment effect, the experimenter should try to employ valid *unobtrusive measures* (Webb *et al.*, 1966).

H. Interaction of History and Treatment Effects

Historical conditions at the time of an experiment may affect the results of the treatment in such a way that the effect would not be found on other occasions. Emotion-packed activities of a relatively brief duration, e.g., the firing of a top official in the local government, or a state basketball tournament, which occur during or immediately prior to an experiment might produce behavior which would not be typical at other times. However, the experimenter can usually arrive at some conclusion as to whether or not such activities have invalidated the results to some extent. On the other hand, historical conditions of a relatively longer duration, e.g., wartime, or exceptionally high student morale, may have an effect on the treatment which is not immediately obvious or estimable. Evidence for the existence of an interaction between history and the experimental treatment can be obtained by replicating the treatment across time.

I. Measurement of the Dependent Variable

Both the conceptualization of the dependent variable and the operational definition of the dependent variable are relevant to the generalization of experimental results. Since the conceptualization of the dependent variable is directly related to the hypotheses formulated by the experimenter, it seems logical that more than one dependent variable should be defined for most experimental studies (see Cronbach, 1957). Wittrock (1966) has emphasized the need for a variety of dependent variables in learning-by-discovery studies, and Cronbach (1966) has listed twelve outcomes which have been claimed by various spokesmen of the discovery method. Scriven (1959) pointed out that psychoanalysts seem to have little agreement about the goals of therapy, and, thus, the contributions which can be made by research on psychotherapy are related to the measurement of multiple outcomes.

> The fact of a cure is thus impossible to pin down using only one indicator; but by employing multiple criteria we shall at least be able to locate a number of clear-cut cases, and our study can rely on them and not insist on decisions where substantial conflicts of criteria exist (pp. 245-246).

The operational definition of the dependent variable refers to the selection of a measuring instrument which is assumed to measure both reliably and validly the underlying construct. For some dependent variables there are several instruments from which the experimenter can choose. This raises the question of the comparability of tests which supposedly measure the same thing. It seems that an important contribution to empirical knowledge would be a sampling of similar tests in an experiment.

Webb *et al.* (1966) and Campbell (1967) have advocated the use of *unobtrusive measures* in experimental settings. In addition to being relatively free of response sets and many sources of unreliability, they appear to be valid measures for some dependent variables.

J. Interaction of Time of Measurement and Treatment Effects

A treatment effect which is observed immediately after the treatment period may not be maintained at some later time, e.g., a

month or six months after the treatment period. Most experimenters fail to take the time element into account and thus risk invalid generalization of treatment effects to other points in time. In Krumboltz and Weisman's (1962) study of different modes of response (overt, covert, etc.) with programed instruction, the three experimental groups did not differ on an immediate retention test, but significant differences were observed on an alternate form of the test administered two weeks later. Although this particular finding has never been duplicated (Anderson, 1967), it does illustrate the source of external invalidity in question. Other studies (Goldbeck and Campbell, 1962; Krumboltz and Kiesler, 1965) have also shown that differential results may be observed with immediate and delayed retention tests of programed learning with various response modes.

Differential effects over time have also occurred in experiments of opinion change. Hovland *et al.* (1949), using films to indoctrinate members of the Armed Forces during World War II, observed differences between the immediate and long-term effects (nine weeks later) of the films. The nature of the differences prompted them to suggest

> . . . that "sleeper" effects are obtained among individuals already *predisposed* to accept an opinion but who have not yet accepted it. According to this hypothesis, a person soon "forgets" the ideas he has learned which are not consonant with his predispositions, but that he retains without loss or even with an increment those ideas consonant with his predispositions (pp. 192-193).

Hovland and Weiss (1951) found a difference in the immediate effects of presentations by trustworthy and untrustworthy communicators. With the passage of time (four weeks), the initial differences disappeared.

An experimental design which includes the measurement of the dependent variables at several points in time will increase the ecological validity of the results. Daw and Gage's (1967) experiment of the effect on principals of knowing their teachers' ratings of "actual" and "ideal" principals is an excellent illustration of such a design.

CONCLUSION

The purpose of this paper is to identify and illustrate sources of external invalidity of experiments. We hope that by illustrating sources of external *invalidity* that we have not engendered in the reader a pessimistic view of the possibilities for externally valid experimental designs. Our intent is to encourage care in designing experiments and interpreting experimental results. There are many ways that an experiment can be designed to produce valid results; there are far fewer ways that an experiment can produce invalid results. Thus an identification of the sources of external *invalidity* is intended to serve as a useful mnemonic for evaluating the generalizability of experimental results. Undoubtedly, the list we have proposed could be profitably lengthened.

The references for each source of external invalidity were included to help us to illustrate the threats to external validity. Although the references are not exhaustive, we did search extensively through the literature to find illustrations of sources of external invalidity. Thus we hope that this paper does not give the impression that most experiments in the past have produced invalid results.

Each source of external invalidity suggests an experimental design which effectively controls it. For example, one design which guards against the interaction of time and measurement and treatment effects is the following:

(Experimental)	R	X	O_1	O_2	$\cdots$	O_m
(Control)	R		O_1	O_2	$\cdots$	O_m

where R denotes random assignment to groups, X denotes the experimental treatment, and $O_{1 \ldots m}$ denotes measurements on the dependent variable. Revising experimental designs to control for other sources of external invalidity would be an extension of the present work.

REFERENCES

ANDERSON, RICHARD C. "Educational Psychology." In Paul R. Farnsworth *et al.* (Eds.), *Annual Review of Psychology*, Vol. *18*. Palo Alto, California: Annual Reviews, Inc., 1967. pp. 129-164.

BARKER, ROGER G. "Explorations in Ecological Psychology," *American Psychologist* 20:1-14; January 1965.

BARKER, ROGER G., and GUMP, PAUL V. *Big School, Small School.* Stanford: Stanford University Press, 1964. 250 pp.

BORING, EDWIN G. "Newton and the Spectral Lines," *Science* 136: 600-601; May 1962.

BRANDT, LEWIS W. "Wanted: A Representative Experiment," *Contemporary Psychology* 12:192-193; April 1967.

BROWNELL, WILLIAM A. "The Evaluation of Learning Under Dissimilar Systems of Instruction." In Jeremy D. Finn, (Ed.), *Introduction to Research and Evaluation.* Buffalo, New York: State University of New York at Buffalo, 1966. pp. 142-150.

BROWNELL, WILLIAM A., and MOSER, HAROLD E. *Meaningful vs. Mechanical Learning: A Study in Grade III Subtraction.* Duke University Research Studies in Education, No. 8. Durham: Duke University Press, 1949. 207 pp.

BRUNSWIK, EGON. "Representative Design and Probabilistic Theory in a Functional Psychology," *Psychological Review* 62:193-217; May 1955.

BRUNSWIK, EGON. *Perception and the Representative Design of Psychological Experiments.* Berkeley, California: University of California Press, 1956. 154 pp.

CAMPBELL, DONALD T. "Factors Relevant to the Validity of Experiments in Social Settings," *Psychological Bulletin* 54:297-312; July 1957.

CAMPBELL, DONALD T. "Administrative Experimentation, Institutional Records, and Nonreactive Measures." In Julian C. Stanley (Ed.), *Improving Experimental Design and Statistical Analysis.* Chicago: Rand McNally & Company, 1967. pp. 257-291.

CAMPBELL, DONALD T. and STANLEY, JULIAN C. "Experimental and Quasi-Experimental Designs for Research on Teaching." In N.L. Gage (Ed.), *Handbook of Research on Teaching.* Chicago: Rand McNally & Company, 1963. pp. 171-246. (Also published as a separate, *Experimental and Quasi-Experimental Designs for Research,* by Rand McNally & Company, 1966).

CATTELL, RAYMOND B. "Guest Editorial: Multivariate Behavioral Research and the Integrative Challenge," *Multivariate Behavioral Research* 1:4-23; January 1966.

COOK, DESMOND L. *The Impact of the Hawthorne Effect in Experimental Designs in Educational Research.* Cooperative Research Project No. 1757, U. S. Office of Education, June 1967. 160 pp.

CORNFIELD, JEROME, and TUKEY, JOHN W. "Average Values of Mean Squares in Factorials," *The Annals of Mathematical Statistics* 27:907-949; December 1956.

COX, D. R. *Planning of Experiments.* New York: John Wiley & Sons, Inc., 1958. 308 pp.

CRESPI, LEO P. "The Interview Effect in Polling," *The Public Opinion Quarterly* 12:99-111; Spring 1948.

CRONBACH, LEE J. "The Two Disciplines of Scientific Psychology," *The American Psychologist* 12:671-684; November 1957.

CRONBACH, LEE J. "Course Improvement Through Education," *Teachers College Record* 64:672-683; May 1963.

CRONBACH, LEE J. "The Logic of Experiments on Discovery." In Lee S. Shulman and Evan R. Keislar (Eds.), *Learning by Discovery: A Critical Appraisal.* Chicago: Rand McNally & Company, 1966. pp .77-92.

CRONBACH, LEE J., and GLESER, GOLDINE C. *Psychological Tests and Personnel Decisions.* Urbana: University of Illinois Press, 1965. 347 pp.

DAW, R. W. and GAGE, N. L. "Effect of Feedback from Teachers to Principals," *Journal of Educational Psychology* 58:181-188; June 1967.

DUNCAN, CARL P. "Learning to Learn in Response-Discovery and in Paired-Associate Lists," *The American Journal of Psychology* 77: 367-379; September 1964.

ECKSTRAND, GORDON A. "Individuality in the Learning Process: Some Issues and Implications," *The Psychological Record* 12:405-416; October 1962.

EDWARDS, ALLEN L. and CRONBACH, LEE J. "Experimental Design for Research in Psychotherapy," *Journal of Clinical Psychology* 8:51-59; January 1952.

EIGEN, LEWIS D. "A Comparison of Three Modes of Presenting a Programed Instruction Sequence," *The Journal of Educational Research* 55:453-460; June-July 1962.

ENTWISLE, DORIS R. "Attensity: Factors of Specific Set in School Learning," *Harvard Educational Review* 31:84-101; Winter 1961(a).

ENTWISLE, DORIS R. "Interactive Effects of Pretesting," *Educational and Psychological Measurement* 21:607-620; Autumn 1961 (b).

FAWL, CLIFFORD L. "Disturbances Experienced by Children in Their

Natural Habitats." In Roger G. Barker (Ed.), *The Stream of Behavior.* New York: Appleton-Century-Crofts, 1963. pp. 99-126.

FRIEDMAN, MYLES I. "The Effectiveness of Machine Instruction in the Teaching of Second and Third Grade Spelling," *The Journal of Educational Research* 60:366-369; April 1967.

GAGNÉ, ROBERT M. "Varieties of Learning and the Concept of Discovery." In Lee S. Shulman and Evan R. Keislar (Eds.), *Learning by Discovery: A Critical Appraisal.* Chicago: Rand McNally & Company, 1966. pp. 135-150.

GOLDBECK, ROBERT A., and CAMPBELL, VINCENT N. "The Effects of Response Mode and Response Difficulty on Programed Learning," *Journal of Educational Psychology* 53:110-118; June 1962.

GOODMAN, LEO A. "Ecological Regressions and Behavior of Individuals," *American Sociological Review* 18:663-664; December 1953.

HASTINGS, J. THOMAS. "Curriculum Evaluation: The Why of the Outcomes," *Journal of Educational Measurement* 3:27-32; Spring 1966.

HOVLAND, CARL I., JANIS, IRVING L., and KELLEY, HAROLD H. *Communication and Persuasion.* New Haven: Yale University Press, 1953. 315 pp.

HOVLAND, CARL I., LUMSDAINE, ARTHUR A., and SHEFFIELD, FRED D. *Experiments on Mass Communication.* Princeton: Princeton University Press, 1949. 345 pp.

HOVLAND, CARL I., and WEISS, WALTER. "The Influence of Source Credibility on Communication Effectiveness," *The Public Opinion Quarterly* 15:635-650; Winter 1951.

KAGAN, JEROME. "Learning, Attention and the Issue of Discovery." In Lee S. Shulman and Evan R. Keislar (Eds.), *Learning by Discovery: A Critical Appraisal.* Chicago: Rand McNally & Company, 1966. pp. 151-161.

KAPLAN, ABRAHAM. *The Conduct of Inquiry: Methodology for Behavioral Science.* San Francisco: Chandler Publishing Company, 1964. 428 pp.

KEMPTHORNE, OSCAR. "The Design and Analysis of Experiments with Some Reference to Educational Research." In Raymond O. Collier, Jr., and Stanley M. Elam, (Eds.), *Research Design and Analysis: Second Annual Phi Delta Kappa Symposium on Educational Research.* Bloomington, Indiana: Phi Delta Kappa, 1961, pp. 97-126.

KENDLER, TRACY S. and KENDLER, HOWARD H. "Reversal and Nonreversal Shifts in Kindergarten Children," *Journal of Experimental Psychology* 58:56-60; July 1959.

KINTZ, B. L., *et al.* "The Experimenter Effect," *Psychological Bulletin* 63:223-232; April 1965.

KRESS, GERARD C., Jr., and GROPPER, GEORGE L. "A comparison of Two Strategies for Individualizing Fixed-Paced Programed Instruction," *American Educational Research Journal* 3:273-280; November 1966.

KRUMBOLTZ, JOHN D. and KEISLAR, CHARLES A. "The Partial Reinforcement Paradigm and Programed Instruction," *The Journal of Programed Instruction* 3:9-14; 1965.

KRUMBOLTZ, JOHN D. and WEISMAN, RONALD G. "The Effect of Overt versus Covert Responding to Programed Instruction on Immediate and Delayed Retention," *Journal of Educational Psychology* 53:89-92; April 1962.

KUENNE, MARGARET R. "Experimental Investigation of the Relation of Language to Transposition Behavior in Young Children," *Journal of Experimental Psychology* 36:471-490; December 1946.

LANA, ROBERT E. "Pretest-Treatment Interaction Effects in Attitudinal Studies," *Psychological Bulletin* 56:293-300; July 1959(a).

LANA, ROBERT E. "A Further Investigation of the Pretest-Treatment Interaction Effect," *Journal of Applied Psychology* 43:421-422, December 1959(b).

LANA, ROBERT E. "Inhibitory Effects of a Pretest on Opinion Change," *Educational and Psychological Measurement* 26:139-150; Spring 1966.

LANA, ROBERT E., and KING, DAVID J. "Learning Factors as Determiners of Pretest Sensitization," *Journal of Applied Psychology* 44:189-191; June 1960.

LANA, ROBERT E., and LUBIN, ARDIE. "The Effect of Correlation on the Repeated Measures Design," *Educational and Psychological Measurement* 23:729-739; Winter 1963.

LITTLE, JAMES KENNETH. "Results of Use of Machines for Testing and for Drill upon Learning in Educational Psychology," *The Journal of Experimental Education* 3:45-49; September 1934.

LUBIN, ARDIE. "The Interpretation of Significant Interaction," *Educational and Psychological Measurement* 21:807-817; Winter 1961.

LUCHINS, ABRAHAM S. "Mechanization in Problem Solving: The Effect of *Einstellung*," *Psychological Monographs*, 54, No. 6, 1942.

MAY, ROLLO. "Historical and Philosophical Presuppositions for Understanding Therapy." In O. Hobart Mowrer (Ed.), *Psychotherapy: Theory and Research*. New York: The Ronald Press Company, 1953. Pp. 9-43.

McNEIL, JOHN D. "Programed Instruction as a Research Tool in Read-

ing: An Annotated Case," *The Journal of Programed Instruction* 1:37-42; 1962.

MILLMAN, JASON. "In the Service of Generalization," *Psychology in the Schools* 3:333-339; October 1966.

MILLMAN, JASON. "Should Differential Treatments be Used: A Critique of Present Procedures of Analysis and Some Suggested Options." A Paper Presented at the 1967 Annual Meeting of the Educational Research Association of New York State, 1967. 10 pp.

NOSANCHUK, T. A. and MARCHAK, M. P. "Pretest Sensitization and Attitude Change." Duplicated, University of British Columbia, Undated. 14 pp.

NOSANCHUK, T. A. and HARE, ROBERT D. "Word-Recognition Threshold as a Function of Pretest Sensitization," *Psychonomic Science* 6:51-52; September 1966.

ORNE, MARTIN T. "On the Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characteristics and Their Implications," *American Psychologist* 17:776-783; November 1962.

OSBURN, H. G. and MELTON, R. S. "Prediction of Proficiency in a Modern and Traditional Course in Beginning Algebra," *Educational and Psychological Measurement* 23:277-287; Summer 1963.

PAGE, ELLIS BATTEN "Educational Research: Replicable or Generalizable?" *Phi Delta Kappan* 39:302-304; March 1958(a).

PAGE, ELLIS BATTEN. "Teacher Comments and Student Performance: A Seventy-Four Classroom Experiment in School Motivation," *The Journal of Educational Psychology* 49:173-181; August 1958(b).

REED, JERRY E. and HAYMAN, JOHN L., JR. "An Experiment Involving Use of English 2600, An Automated Instruction Text," *The Journal of Educational Research* 55:476-484; June-July 1962.

ROBINSON, W. S. "Ecological Correlations and the Behavior of Individuals," *American Sociological Review* 15:351-357; June 1950.

ROSENBERG, MILTON J. "When Dissonance Fails: On Eliminating Evaluation Apprehension from Attitude Measurement," *Journal of Personality and Social Psychology* 1:28-42; January 1965.

ROSENTHAL, DAVID and FRANK, JEROME D. "Psychotherapy and the Placebo Effect," *Psychological Bulletin* 53:294-302; July 1956.

ROSENTHAL, ROBERT. "On the Social Psychology of the Psychological Experiment: The Experimenter's Hypothesis as Unintended Determinant of Experimental Results," *American Scientist* 51:268-283; June 1963.

ROSENTHAL, ROBERT. *Experimenter Effects in Behavioral Research.* New York: Appleton-Century-Crofts, 1966. 464 pp.

SCRIVEN, MICHAEL. "The Experimental Investigation of Psychoanalysis." In Sidney Hook (Ed.), *Psychoanalysis, Scientific Method, and Philosophy.* Washington Square, New York: New York University Press, 1959. Pp. 226-251.

SCRIVEN, MICHAEL. "The Methodology of Evaluation." In AERA Monograph Series on Curriculum Evaluation, No. 1, *Perspectives of Curriculum Evaluation.* Chicago: Rand McNally & Company, 1967. Pp. 39-83.

SNOW, RICHARD E. *The Importance of Selected Audience and Film Characteristics as Determiners of the Effectiveness of Instructional Films.* Report to the United States Office of Education. Lafayette: Audio Visual Center, Purdue University, 1963. 263 pp.

SNOW, RICHARD E., TIFFIN, JOSEPH, and SEIBERT, WARREN F. "Individual Differences and Instructional Film Effects," *Journal of Educational Psychology* 56:315-326; December 1965.

SOLOMON, RICHARD L. "An Extension of Control Group Design," *Psychological Bulletin* 46:137-150; March 1949.

SPENCE, K. W. and TAYLOR, JANET. "Anxiety and Strength of the UCS as Determiners of the Amount of Eyelid Conditioning," *Journal of Experimental Psychology* 42:183-188; September 1951.

STAR, SHIRLEY A. and HUGHES, HELEN MacGILL. "Report on an Educational Campaign: The Cincinnati Plan for the United Nations," *The American Journal of Sociology* 55:389-400; January 1950.

STERN, GEORGE G., STEIN, MORRIS I., and BLOOM, BENJAMIN S. *Methods in Personality Assessment.* Glencoe, Illinois: The Free Press, 1956. 271 pp.

STOLUROW, LAWRENCE M. "Model the Master Teacher or Master the Teaching Model." In John D. Krumboltz (Ed.), *Learning and the Educational Process.* Chicago: Rand McNally & Company, 1965. Pp. 223-247.

TIEDEMAN, DAVID V., and COGAN, MORRIS. "New Horizons in Educational Research," *Phi Delta Kappan* 39:286-291; March 1958.

VOLSKY, THEODORE, JR., *et al. The Outcomes of Counseling and Psychotherapy: Theory and Research.* Minneapolis: University of Minnesota Press, 1965. 209 pp.

WEBB, EUGENE J., *et al. Unobtrusive Measures: Nonreactive Research in the Social Sciences.* Chicago: Rand McNally & Company, 1966. 225 pp.

WEITZ, JOSEPH. "Tiny Theories," *American Psychologist* 22:157; February 1967.

WINDLE, CHARLES. "Test-Retest Effect on Personality Question-

naires," *Educational and Psychological Measurement* 14:617-633; Winter 1954.

WITTROCK, M. C. "The Learning by Discovery Hypothesis." In Lee S. Shulman and Evan R. Keisler (Eds.), *Learning by Discovery: A Critical Appraisal.* Chicago: Rand McNally & Company, 1966. pp. 33-76.

(Received March, 1968)
(Revised May, 1968)

AUTHORS

GLASS, GENE V. *Address:* Laboratory of Educational Research, University of Colorado, Boulder, Colorado 80302 *Title:* Associate Professor of Education *Age:* 28 *Degrees:* B.A., Univ. of Nebraska; M.S., Ph.D., Univ. of Wisconsin *Specialization:* Statistics, experimental design, psychometrics, evaluation.

BRACHT, GLENN H. *Address:* Laboratory of Educational Research, University of Colorado, Boulder, Colorado 80302 *Title:* NDEA Fellow *Age:* 29 *Degrees:* B.S., Concordia, Seward, Nebr.; M.A., Univ. of Colorado *Specialization:* Educational research, evaluation, and measurement.