PSYCHOLOGICAL BULLETIN
VOL. 54, NO. 4, 1957

FACTORS RELEVANT TO THE VALIDITY OF EXPERIMENTS IN SOCIAL SETTINGS[1]

DONALD T. CAMPBELL
Northwestern University

What do we seek to control in experimental designs? What extraneous variables which would otherwise confound our interpretation of the experiment do we wish to rule out? The present paper attempts a specification of the major categories of such extraneous variables and employs these categories in evaluating the validity of standard designs for experimentation in the social sciences.

Validity will be evaluated in terms of two major criteria. First, and as a basic minimum, is what can be called *internal validity:* did in fact the experimental stimulus make some significant difference in this specific instance? The second criterion is that of *external validity, representativeness,* or *generalizability:* to what populations, settings, and variables can this effect be generalized? Both criteria are obviously important although it turns out that they are to some extent incompatible, in that the controls required for internal validity often tend to jeopardize representativeness.

The extraneous variables affecting internal validity will be introduced in the process of analyzing three pre-experimental designs. In the subsequent evaluation of the applicability of three true experimental designs, factors leading to external invalidity will be introduced. The effects of these extraneous variables will be considered at two levels: as simple or main effects, they occur independently of or in addition to the effects of the experimental variable; as interactions, the effects appear in conjunction with the experimental variable. The main effects typically turn out to be relevant to internal validity, the interaction effects to external validity or representativeness.

The following designation for experimental designs will be used: *X* will represent the exposure of a group to the experimental variable or event, the effects of which are to be measured; *O* will refer to the process of observation or measurement, which can include watching what people do, listening, recording, interviewing, administering tests, counting lever depressions, etc. The *X*s and *O*s in a given row are applied to the same specific persons. The left to right dimension indicates temporal order. Parallel rows represent equivalent samples of persons unless otherwise specified. The designs will be numbered and named for cross-reference purposes.

THREE PRE-EXPERIMENTAL DESIGNS AND THEIR CONFOUNDED EXTRANEOUS VARIABLES

The One-Shot Case Study. As Stouffer (**32**) has pointed out, much social science research still uses De-

[1] A dittoed version of this paper was privately distributed in 1953 under the title "Designs for Social Science Experiments." The author has had the opportunity to benefit from the careful reading and suggestions of L. S. Burwen, J. W. Cotton, C. P. Duncan, D. W. Fiske, C. I. Hovland, L. V. Jones, E. S. Marks, D. C. Pelz, and B. J. Underwood, among others, and wishes to express his appreciation. They have not had the opportunity of seeing the paper in its present form, and bear no responsibility for it. The author also wishes to thank S. A. Stouffer (**33**) and B. J. Underwood (**36**) for their public encouragement.

sign 1, in which a single individual or group is studied in detail only once, and in which the observations are attributed to exposure to some prior situation.

$X \quad O$ 1. One-Shot Case Study

This design does not merit the title of experiment, and is introduced only to provide a reference point. The very minimum of useful scientific information involves at least one formal comparison and therefore at least two careful observations (**2**).

The One-Group Pretest-Posttest Design. This design does provide for one formal comparison of two observations, and is still widely used.

$O_1 \; X \; O_2$ 2. One-Group Pretest-Posttest Design

However, in it there are four or five categories of extraneous variables left uncontrolled which thus become rival explanations of any difference between O_1 and O_2, confounded with the possible effect of X.

The first of these is the main effect of *history*. During the time span between O_1 and O_2 many events have occurred in addition to X, and the results might be attributed to these. Thus in Collier's (**8**) experiment, while his respondents[2] were reading Nazi propaganda materials, France fell, and the obtained attitude changes seemed more likely a result of this event than of the propaganda.[3] By history is meant the specific event series other than X, i.e., the extra-experimental uncontrolled stimuli. Relevant to this variable is the concept of experimental isolation, the employment of experimental settings in which all extraneous stimuli are eliminated. The approximation of such control in much physical and biological research has permitted the satisfactory employment of Design 2. But in social psychology and the other social sciences, if history is confounded with X the results are generally uninterpretable.

The second class of variables confounded with X in Design 2 is here designated as *maturation*. This covers those effects which are systematic with the passage of time, and not, like history, a function of the specific events involved. Thus between O_1 and O_2 the respondents may have grown older, hungrier, tireder, etc., and these may have produced the difference between O_1 and O_2, independently of X. While in the typical brief experiment in the psychology laboratory, maturation is unlikely to be a source of change, it has been a problem in research in child development and can be so in extended experiments in social psychology and education. In the form of "spontaneous remission" and the general processes of healing it becomes an important variable to control in medical research, psychotherapy, and social remediation.

There is a third source of variance that could explain the difference between O_1 and O_2 without a recourse to the effect of X. This is the effect of *testing* itself. It is often true that persons taking a test for the second time make scores systematically different from those taking the test for the first time. This is indeed the case for intelligence tests, where a second mean may be expected to run as much as five IQ points higher than the first one. This possibility makes important a distinction between *reactive* measures and *nonreactive* measures. A reactive measure is one

[2] In line with the central focus on social psychology and the social sciences, the term *respondent* is employed in place of the terms *subject*, *patient*, or *client*.

[3] Collier actually used a more adequate design than this, an approximation to Design 4.

which modifies the phenomenon under study, which changes the very thing that one is trying to measure. In general, any measurement procedure which makes the subject self-conscious or aware of the fact of the experiment can be suspected of being a reactive measurement. Whenever the measurement process is *not* a part of the normal environment it is probably reactive. Whenever measurement exercises the process under study, it is almost certainly reactive. Measurement of a person's height is relatively nonreactive. However, measurement of weight, introduced into an experimental design involving adult American women, would turn out to be reactive in that the process of measuring would stimulate weight reduction. A photograph of a crowd taken in secret from a second story window would be nonreactive, but a news photograph of the same scene might very well be reactive, in that the presence of the photographer would modify the behavior of people seeing themselves being photographed. In a factory, production records introduced for the purpose of an experiment would be reactive, but if such records were a regular part of the operating environment they would be nonreactive. An English anthropologist may be nonreactive as a participant-observer at an English wedding, but might be a highly reactive measuring instrument at a Dobu nuptials. Some measures are so extremely reactive that their use in a pretest-posttest design is not usually considered. In this class would be tests involving surprise, deception, rapid adaptation, or stress. Evidence is amply present that tests of learning and memory are highly reactive (**35**, **36**). In the field of opinion and attitude research our well-developed interview and attitude test techniques must be rated as reactive, as shown, for example, by Crespi's (**9**) evidence.

Even within the personality and attitude test domain, it may be found that tests differ in the degree to which they are reactive. For some purposes, tests involving voluntary self-description may turn out to be more reactive (especially at the interaction level to be discussed below) than are devices which focus the respondent upon describing the external world, or give him less latitude in describing himself (e.g., **5**). It seems likely that, apart from considerations of validity, the Rorschach test is less reactive than the TAT or MMPI. Where the reactive nature of the testing process results from the focusing of attention on the experimental variable, it may be reduced by imbedding the relevant content in a comprehensive array of topics, as has regularly been done in Hovland's attitude change studies (**14**). It seems likely that with attention to the problem, observational and measurement techniques can be developed which are much less reactive than those now in use.

Instrument decay provides a fourth uncontrolled source of variance which could produce an O_1–O_2 difference that might be mistaken for the effect of X. This variable can be exemplified by the fatiguing of a spring scales, or the condensation of water vapor in a cloud chamber. For psychology and the social sciences it becomes a particularly acute problem when human beings are used as a part of the measuring apparatus, as judges, observers, raters, coders, etc. Thus O_1 and O_2 may differ because the raters have become more experienced, more fatigued, have acquired a different adaptation level, or have learned about the purpose of the experiment, etc. However infelicitously, this term will be used to typify those problems introduced when shifts in measurement conditions are confounded with the effect of X, including such crudities as having a different observer at O_1 and O_2, or using a different interviewer or coder. Where the use of different interviewers, observers, or experimenters is unavoidable, but where they are used in large numbers, a sampling equivalence of interviewers is required, with the relevant N being the N of interviewers, not interviewees, except as refined through cluster sampling considerations (**18**).

A possible fifth extraneous factor deserves mention. This is statistical *regression*. When, in Design 2, the group under investigation has been selected for its extremity on O_1, O_1–O_2 shifts toward the mean will occur which are due to random imperfections of the measuring instrument or random instability within the population, as reflected in the test-retest reliability. In general, regression operates like maturation in that the effects increase systematically with the O_1–O_2 time interval. McNemar (**22**) has demonstrated the profound mistakes in interpretation which failure to control this factor can introduce in remedial research.

The Static Group Comparison. The third pre-experimental design is the Static Group Comparison.

$X \quad O_1$
- - - - - 3. The Static Group Comparison
$\quad\;\; O_2$

In this design, there is a comparison of a group which has experienced X with a group which has not, for the purpose of establishing the effect of X. In contrast with Design 6, there is in this design no means of certifying that the groups were equivalent at some prior time. (The absence of sampling equivalence of groups is symbolized by the row of dashes.) This design has its most typical occurrence in the social sciences, and both its prevalence and its weakness have been well indicated by Stouffer (**32**). It will be recognized as one form of the correlational study. It is introduced here to complete the list of confounding factors. If the Os differ, this difference could have come about through biased *selection* or recruitment of the persons making up the groups; i.e., they might have differed anyway without the effect of X. Frequently, exposure to X (e.g., some mass communication) has been voluntary and the two groups have an inevitable systematic difference on the factors determining the choice involved, a difference which no amount of matching can remove.

A second variable confounded with the effect of X in this design can be called experimental *mortality*. Even if the groups were equivalent at some prior time, O_1 and O_2 may differ now not because individual members have changed, but because a biased subset of members have dropped out. This is a typical problem in making inferences from comparisons of the attitudes of college freshmen and college seniors, for example.

True Experimental Designs

The Pretest-Posttest Control Group Design. One or another of the above considerations led psychologists between 1900 and 1925 (**2**, **30**) to expand Design 2 by the addition of a control group, resulting in Design 4.

$O_1 \; X \; O_2$ 4. Pretest-Posttest Control Group
$O_3 \quad\; O_4$ Design

Because this design so neatly controls for the main effects of history, maturation, testing, instrument decay, regression, selection, and mortality, these separate sources of variance are not usually made explicit. It seems well to state briefly the relationship of the design to each of these confounding factors, with particular attention to the application of the design in social settings.

If the differences between O_1 and O_2 were due to intervening historical events, then they should also show up in the O_3–O_4 comparison. Note, however, several complications in achieving this control. If respondents are run in groups, and if there is only one experimental session and one control session, then there is no control over the unique internal histories of the groups. The O_1–O_2 difference, even if not appearing in O_3–O_4, may be due to a chance distracting factor appearing in one or the other group. Such a design, while controlling for the shared history or event series, still confounds X with the unique session history. Second, the design implies a simultaneity of O_1 with O_3 and O_2 with O_4 which is usually impossible. If one were to try to achieve simultaneity by using two experimenters, one working with the experimental respondents, the other with the controls, this would confound experimenter differences with X (introducing one type of instrument decay). These considerations make it usually imperative that, for a true experiment, the experimental and control groups be tested and exposed individually or in small subgroups, and that sessions of both types be temporally and spatially intermixed.

As to the other factors: if maturation or testing contributed an O_1–O_2 difference, this should appear equally in the O_3–O_4 comparison, and these variables are thus controlled for their main effects. To make sure the design controls for instrument decay, the same individual or small-session approximation to simultaneity needed for history is required. The occasional practice of running the experimental group and control group at different times is thus ruled out on this ground as well as that of history. Otherwise the observers may have become more experienced, more hurried, more careless, the maze more redolent with irrelevant cues, the lever-tension and friction diminished, etc. Only when groups are effectively simultaneous do these factors affect experimental and control groups alike. Where more than one experimenter or observer is used, counterbalancing experimenter, time, and group is recommended. The balanced Latin square is frequently useful for this purpose (4).

While regression is controlled in the design as a whole, frequently secondary analyses of effects are made for extreme pretest scorers in the experimental group. To provide a control for effects of regression, a parallel analysis of extremes should also be made for the control group.

Selection is of course handled by the sampling equivalence ensured through the randomization employed in assigning persons to groups, perhaps supplemented by, but not supplanted by, matching procedures. Where the experimental and control groups do not have this sort of equivalence, one has a compromise design rather than a true experiment. Furthermore, the O_1–O_3 comparison provides a check on possible sampling differences.

The design also makes possible the examination of experimental mortality, which becomes a real problem for experiments extended over weeks or months. If the experimental and control groups do not differ in the number of lost cases nor in their pretest scores, the experiment can be judged internally valid on this point, although mortality reduces the generalizability of effects to the original population from which the groups were selected.

For these reasons, the Pretest-Posttest Control Group Design has been the ideal in the social sciences for some thirty years. Recently, however, a serious and avoidable imperfection in it has been noted, perhaps first by Schanck and Goodman (**29**). Solomon (**30**) has expressed the point as an *interaction* effect of testing. In the terminology of analysis of variance, the effects of history, maturation, and testing, as described so far, are all *main* effects, manifesting themselves in mean differences independently of the presence of other variables. They are effects that could be added on to other effects, including the effect of the experimental variable. In contrast, interaction effects represent a joint effect, specific to the concomitance of two or more conditions, and may occur even when no main effects are present. Applied to the testing variable, the interaction effect might involve not a shift due solely or directly to the measurement process, but rather a sensitization of respondents to the experimental variable so that when X was preceded by O there would be a change, whereas both X and O would be without effect if occurring alone. In terms of the two types of validity, Design 4 is internally valid, offering an adequate basis for generalization to other sampling-equivalent *pretested* groups. But it has a serious and systematic weakness in representativeness in that it offers, strictly speaking, no basis for generalization to the *unpretested* population. And it is usually the *unpretested* larger universe from which these samples were taken to which one wants to generalize.

A concrete example will help make this clearer. In the NORC study of a United Nations information campaign (**31**), two equivalent samples, of a thousand each, were drawn from the city's population. One of these samples was interviewed, following which the city of Cincinnati was subjected to an intensive publicity campaign using all the mass media of communication. This included special features in the newspapers and on the radio, bus cards, public lectures, etc. At the end of two months, the second sample of 1,000 was interviewed and the results compared with the first 1,000. There were no differences between the two groups except that the second group was somewhat more pessimistic about the likelihood of Russia's cooperating for world peace, a result which was attributed to history rather than to the publicity campaign. The second sample was no better informed about the United Nations nor had it noticed in particular the publicity campaign which had been going on. In connection with a program of research on panels and the reinterview problem, Paul Lazarsfeld and the Bureau of Applied Social Research arranged to have the initial sample reinterviewed at the same time as the second sample was interviewed, after the publicity campaign. This reinterviewed group showed significant attitude changes, a high degree of awareness of the campaign and important increases in information. The inference in this case is unmistakably that the initial interview had sensitized the persons interviewed to the topic of the United Nations, had raised in them a focus of awareness which made the subsequent publicity campaign effective for them but for them only. This study and other studies clearly document the possibility of interaction effects which seriously limit our capacity to generalize from the pretested experimental group to the unpretested general population. Hovland (**15**) reports a general finding which is of the opposite nature but is, nonetheless, an indication of an interactive effect. In his Army studies the initial pretest served to reduce the effects of the experimental variable, presumably by creating a commitment to a given position. Crespi's (**9**) findings support this expectation. Solomon (**30**) reports two studies with school children in which a spelling pretest reduced the effects of a training period. But whatever the direction of the effect, this flaw in the Pretest-Posttest Control Group Design is serious for the purposes of the social scientist.

The Solomon Four-Group Design. It is Solomon's (**30**) suggestion to control this problem by adding to the traditional two-group experiment two unpretested groups as indicated in Design 5.

$$\begin{array}{lll} O_1 & X & O_2 \\ O_3 & & O_4 \\ & X & O_5 \\ & & O_6 \end{array}$$

5. Solomon Four-Group Design

This Solomon Four-Group Design enables one both to control and measure both the main and interaction effects of testing and the main effects of a composite of maturation and history. It has become the new ideal design for social scientists. A word needs to be said about the appropriate statistical analysis. In Design 4, an efficient single test embodying the four measurements is achieved through computing for each individual a pretest-posttest difference score which is then used for comparing by t test the experimental and control groups. Extension of this mode of analysis to the Solomon Four-Group Design introduces an inelegant awkwardness to the otherwise elegant procedure. It involves assuming as a pretest score for the unpretested groups the mean value of the pretest from the first two groups. This restricts the effective degrees of freedom, violates assumptions of independence, and leaves one without a legitimate base for testing the significance of main effects and interaction. An alternative analysis is available which avoids the assumed pretest scores. Note that the four posttests form a simple two-by-two analysis of variance design:

	No X	*X*
Pretested	O_4	O_2
Unpretested	O_6	O_5

The column means represent the main effect of X, the row means the main effect of pretesting, and the interaction term the interaction of pretesting and X. (By use of a t test the combined main effects of maturation and history can be tested through comparing O_6 with O_1 and O_3.)

The Posttest-Only Control Group Design. While the statistical procedures of analysis of variance introduced by Fisher (**10**) are dominant in psychology and the other social sciences today, it is little noted in our discussions of experimental arrangements that Fisher's typical agricultural experiment involves no pretest: equivalent plots of ground receive different experimental treatments and the subsequent yields are measured.[4] Applied to a social experiment

[4] This is not to imply that the pretest is totally absent from Fisher's designs. He suggests the use of previous year's yields, etc., in covariance analysis. He notes, however, "with

as in testing the influence of a motion picture upon attitudes, two randomly assigned audiences would be selected, one exposed to the movie, and the attitudes of each measured subsequently for the first time.

$A\ X\ O_1$ 6. Posttest-Only Control Group Design
$A\quad\ O_2$

In this design the symbol A had been added, to indicate that at a specific time prior to X the groups were made equivalent by a random sampling *assignment*. A is the point of selection, the point of allocation of individuals to groups. It is the existence of this process that distinguishes Design 6 from Design 3, the Static Group Comparison. Design 6 is not a static cross-sectional comparison, but instead truly involves control and observation extended in time. The sampling procedures employed assure us that at time A the groups were equal, even if not measured. A provides a point of prior equality just as does the pretest. A point A is, of course, involved in all true experiments, and should perhaps be indicated in Designs 4 and 5. It is essential that A be regarded as a specific point in time, for groups change as a function of time since A, through experimental mortality. Thus in a public opinion survey situation employing probability sampling from lists of residents, the longer the time since A, the more the sample underrepresents the transient segments of society, the newer dwelling units, etc. When experimental groups are being drawn from a self-selected extreme population, such as applicants for psychotherapy, time since A introduces maturation (spontaneous remission) and regression factors. In Design 6 these effects would be confounded with the effect of X if the As as well as the Os were not contemporaneous for experimental and control groups.

Like Design 4, this design controls for the effects of maturation and history through the practical simultaneity of both the As and the Os. In superiority over Design 4, no main or interaction effects of pretesting are involved. It is this feature that recommends it in particular. While it controls for the main and interaction effects of pretesting as well as does Design 5, the Solomon Four-Group Design, it does not measure these effects, nor the main effect of history-maturation. It can be noted that Design 6 can be considered as the two unpretested "control" groups from the Solomon Design, and that Solomon's two traditional pretested groups have in this sense the sole purpose of measuring the effects of pretesting and history-maturation, a purpose irrelevant to the main aim of studying the effect of X (25). However, under normal conditions of not quite perfect sampling control, the four-group design provides in addition greater assurance against mistakenly attributing to X effects which are not due it, inasmuch as the effect of X is documented in three different fashions (O_1 vs. O_2, O_2 vs. O_4, and O_5 vs. O_6). But, short of the four-group design, Design 6 is often to be preferred to Design 4, and is a fully valid experimental design.

Design 6 has indeed been used in the social sciences, perhaps first of all in the classic experiment by Gosnell, *Getting Out the Vote* (11). Schanck and Goodman (29), Hovland (15) and others (1, 12, 23, 24, 27) have also

annual agricultural crops, knowledge of yields of the experimental area in a previous year under uniform treatment has not been found sufficiently to increase the precision to warrant the adoption of such uniformity trials as a preliminary to projected experiments" (10, p. 176).

employed it. But, in spite of its manifest advantages of simplicity and control, it is far from being a popular design in social research and indeed is usually relegated to an inferior position in discussions of experimental designs if mentioned at all (e.g., **15**, **16**, **32**). Why is this the case?

In the first place, it is often confused with Design 3. Even where *S*s have been carefully assigned to experimental and control groups, one is apt to have an uneasiness about the design because one "doesn't know what the subjects were like before." This objection must be rejected, as our standard tests of significance are designed precisely to evaluate the likelihood of differences occurring by chance in such sample selection. It is true, however, that this design is particularly vulnerable to selection bias and where random assignment is not possible it remains suspect. Where naturally aggregated units, such as classes, are employed intact, these should be used in large numbers and assigned at random to the experimental and control conditions; cluster sampling statistics (**18**) should be used to determine the error term. If but one or two intact classrooms are available for each experimental treatment, Design 4 should certainly be used in preference.

A second objection to Design 6, in comparison with Design 4, is that it often has less precision. The difference scores of Design 4 are less variable than the posttest scores of Design 6 if there is a pretest-posttest correlation above .50 (**15**, p. 323), and hence for test-retest correlations above that level a smaller mean difference would be statistically significant for Design 4 than for Design 6, for a constant number of cases. This advantage to Design 4 may often be more than dissipated by the costs and loss in experimental efficiency resulting from the requirement of two testing sessions, over and above the considerations of representativeness.

Design 4 has a particular advantage over Design 6 if experimental mortality is high. In Design 4, one can examine the pretest scores of lost cases in both experimental and control groups and check on their comparability. In the absence of this in Design 6, the possibility is opened for a mean difference resulting from differential mortality rather than from individual change, if there is a substantial loss of cases.

A final objection comes from those who wish to study the relationship of pretest attitudes to kind and amount of change. This is a valid objection, and where this is the interest, Design 4 or 5 should be used, with parallel analysis of experimental and control groups. Another common type of individual difference study involves classifying persons in terms of amount of change and finding associated characteristics such as sex, age, education, etc. While unavailable in this form in Design 6, essentially the same correlational information can be obtained by subdividing both experimental and control groups in terms of the associated characteristics, and examining the experimental-control difference for such subtypes.

For Design 6, the Posttest-Only Control Group Design, there is a class of social settings in which it is optimally feasible, settings which should be more used than they now are. Whenever the social contact represented by *X* is made to single individuals or to small groups, and where the response to that stimulus can be identified in terms of individuals or type of *X*, Design 6 can be applied. Direct mail and door-to-door contacts represent such settings. The alternation of several appeals from door-to-door in a fund-raising campaign can be organized as a true experiment without increasing the cost of the solicitation. Experimental variation of persuasive materials in a direct-mail sales campaign can provide a better experimental laboratory for the study of mass communication and persuasion than is available in any university. The well-established, if little-used, split-run technique in comparing alternative magazine ads is a true experiment of this type, usually limited to coupon returns rather than sales because of the problem of identifying response with stimulus type (**20**). The split-ballot technique (**7**) long used in public opinion polls to compare different question wordings or question sequences provides an excellent example which can obviously be extended to other topics (e.g., **12**). By and large these laboratories have not yet been used to study social science theories, but they are directly relevant to hypotheses about social persuasion.

Multiple X *designs.* In presenting the above designs, X has been opposed to No-X, as is traditional in discussions of experimental design in psychology. But while this may be a legitimate description of the stimulus-isolated physical science laboratory, it can only be a convenient shorthand in the social sciences, for any No-X period will not be empty of potentially change-inducing stimuli. The experience of the control group might better be categorized as another type of X, a control experience, an X_C instead of No-X. It is also typical of advance in science that we are soon no longer interested in the qualitative fact of effect or no-effect, but want to specify degree of effect for varying degrees of X. These considerations lead into designs in which multiple groups are used, each with a different X_1, X_2, X_3, X_n, or in multiple factorial design, as X_{1a}, X_{1b}, X_{2a}, X_{2b}, etc. Applied to Designs 4 and 6, this introduces one additional group for each additional X. Applied to 5, The Solomon Four-Group Design, two additional groups (one pretested, one not, both receiving X_n) would be added for each variant on X.

In many experiments, X_1, X_2, X_3, and X_n are all given to the same group, differing groups receiving the Xs in different orders. Where the problem under study centers around the effects of order or combination, such counterbalanced multiple X arrangements are, of course, essential. Studies of transfer in learning are a case in point (**34**). But where one wishes to generalize to the effect of each X as occurring in isolation, such designs are not recommended because of the sizable interactions among Xs, as repeatedly demonstrated in learning studies under such labels as proactive inhibition and learning sets. The use of counterbalanced sets of multiple Xs to achieve experimental equation, where natural groups not randomly assembled have to be used, will be discussed in a subsequent paper on compromise designs.

Testing for effects extended in time. The researches of Hovland and his associates (**14, 15**) have indicated repeatedly that the longer range effects of persuasive Xs may be qualitatively as well as quantitatively different from immediate effects. These results emphasize the importance of designing experiments to measure the effect of X at extended periods of time. As the misleading early research on reminiscence and on the consolidation of the memory trace indicate (**36**), repeated measurement of the same persons cannot be trusted to do this if a reactive measurement process is involved. Thus, for Designs 4 and 6, two separate groups must be added for each posttest period. The additional control group cannot be omitted, or the effects of intervening history, maturation, instrument decay, regression, and mortality are confounded with the delayed effects of *X*. To follow fully the logic of Design 5, four additional groups are required for each posttest period.

True experiments in which O *is not not under* E*'s control.* It seems well to call the attention of the social scientist to one class of true experiments which are possible without the full experimental control over both the "when" and "to whom" of both *X* and *O*. As far as this analysis has been able to go, no such true experiments are possible without the ability to control *X*, to withhold it from carefully randomly selected respondents while presenting it to others. But control over *O* does not seem so indispensable. Consider the following design.

$A\ X\ O_1$
$A\ \ \ \ \ O_2$ 6. Posttest Only Design, where *O* cannot be withheld from any respondent
$\quad\ (O)$
$\quad\ (O)$
$\quad\ (O)$

The parenthetical *O*s are inserted to indicate that the studied groups, experimental and control, have been selected from a larger universe all of which will get *O* anyway. An election provides such an *O*, and using "whether voted" rather than "how voted," this was Gosnell's design (**11**). Equated groups were selected at time *A*, and the experimental group subjected to persuasive materials designed to get out the vote. Using precincts rather than persons as the basic sampling unit, similar studies can be made on the content of the voting (**6**). Essential to this design is the ability to create specified randomly equated groups, the ability to expose one of these groups to *X* while withholding it (or providing X_2) from the other group, and the ability to identify the performance of each individual or unit in the subsequent *O*. Since such measures are natural parts of the environment to which one wishes to generalize, they are not reactive, and Design 4, the Pretest-Posttest Control Group Design, is feasible if *O* has a predictable periodicity to it. With the precinct as a unit, this was the design of Hartmann's classic study of emotional vs. rational appeals in a public election (**13**). Note that 5, the Solomon Four-Group Design, is not available, as it requires the ability to withhold *O* experimentally, as well as *X*.

Further Problems of Representativeness

The interaction effect of testing, affecting the external validity or representativeness of the experiment, was treated extensively in the previous section, inasmuch as it was involved in the comparison of alternative designs. The present section deals with the effects upon representativeness of other variables which, while equally serious, can apply to any of the experimental designs.

The interaction effects of selection. Even though the true experiments control selection and mortality for internal validity purposes, these factors have, in addition, an important bearing on representativeness. There is always the possibility that the obtained effects are specific to the experimental population and do not hold true for the populations to which one wants to generalize. Defining the universe of reference in advance and selecting the experimental and control groups from this at random would guarantee representativeness if it were ever achieved in practice. But inevitably not all those so designated are actually eligible for selection by any contact procedure. Our best survey sampling techniques, for example, can designate for potential contact only those available through residences. And, even of those so designated, up to 19 per cent are not contactable for an interview in their own homes even with five callbacks (**37**). It seems legitimate to assume that the more effort and time required of the respondent, the larger the loss through nonavailability and noncooperation. If one were to try to assemble experimental groups away from their own homes it seems reasonable to estimate a 50 per cent selection loss. If, still trying to extrapolate to the general public, one further limits oneself to docile preassembled groups, as in schools, military units, studio audiences, etc., the proportion of the universe systematically excluded through the sampling process must approach 90 per cent or more. Many of the selection factors involved are indubitably highly systematic. Under these extreme selection losses, it seems reasonable to suspect that the experimental groups might show reactions not characteristic of the general population. This point seems worth stressing lest we unwarrantedly assume that the selection loss for experiments is comparable to that found for survey interviews in the home at the respondent's convenience. Furthermore, it seems plausible that the greater the cooperation required, the more the respondent has to deviate from the normal course of daily events, the greater will be the possibility of nonrepresentative reactions. By and large, Design 6 might be expected to require less cooperation than Design 4 or 5, especially in the natural individual contact setting. The interactive effects of experimental mortality are of similar nature. Note that, on these grounds, the longer the experiment is extended in time the more respondents are lost and the less representative are the groups of the original universe.

Reactive arrangements. In any of the experimental designs, the respondents can become aware that they are participating in an experiment, and this awareness can have an interactive effect, in creating reactions to *X* which would not occur had *X* been encountered without this "I'm a guinea pig" attitude. Lazarsfeld (**19**), Kerr (**17**), and Rosenthal and Frank (**28**), all have provided valuable discussions of this problem. Such effects limit generalizations to respondents having this awareness, and preclude generalization to the population encountering *X* with nonexperimental attitudes. The direction of the effect may be one of negativism, such as an unwillingness to admit to any persuasion or change. This would be comparable to the absence of any immediate effect from discredited communicators, as found by Hovland (**14**). The result is probably more often a cooperative responsiveness, in which the respondent accepts the experimenter's expectations and provides psueudoconfirmation. Particularly is this positive response likely when the respondents are self-selected seekers after the cure that *X* may offer. The Hawthorne studies (**21**), illustrate such sympathetic changes due to awareness of experimentation rather than to the specific nature of *X*. In some settings it is possible to disguise the experimental purpose by providing plausible façades in which *X* appears as an incidental part of the background (e.g., **26**, **27**, **29**). We can also make more extensive use of experiments taking place in the intact social situation, in which the respondent is not aware of the experimentation at all.

The discussion of the effects of selection on representativeness has argued against employing intact natural preassembled groups, but the issue of conspicuousness of arrangements argues for such use. The machinery of breaking up natural groups such as departments, squads, and classrooms into randomly assigned experimental and control groups is a source of reaction which can often be avoided by the use of preassembled groups, particularly in educational settings. Of course, as has been indicated, this requires the use of large numbers of such groups under both experimental and control conditions.

The problem of reactive arrangements is distributed over all features of the experiment which can draw the attention of the respondent to the fact of experimentation and its purposes. The conspicuous or reactive pretest is particularly vulnerable, inasmuch as it signals the topics and purposes of the experimenter. For communications of obviously persuasive aim, the experimenter's topical intent is signaled by the *X* itself, if the communication does not seem a part of the natural environment. Even for the posttest-only groups, the occurrence of the posttest may create a reactive effect. The respondent may say to himself, "Aha, now I see why we got that movie." This consideration justifies the practice of disguising the connection between *O* and *X* even for Design 6, as through having different experimental personnel involved, using different façades, separating the settings and times, and embedding the *X*-relevant content of *O* among a disguising variety of other topics.[5]

Generalizing to other Xs. After the internal validity of an experiment has been established, after a dependable effect of *X* upon *O* has been found, the next step is to establish the limits and relevant dimensions of generalization not only in terms of populations and settings but also in terms of categories and aspects of *X*. The actual *X* in any one experiment is a specific combination of stimuli, all confounded for interpretative purposes, and only some relevant to the experimenter's intent and theory. Subsequent experimentation should be designed to purify *X*, to discover that aspect of the original conglomerate *X* which is responsible for the effect. As Brunswik (**3**) has emphasized, the representative sampling of *X*s is as relevant a problem in linking experiment to theory as is the sampling of respondents. To define a category of *X*s along some dimension, and then to sample *X*s for experimental purposes from the full range of stimuli meeting the specification while other aspects of each specific stimulus complex are varied, serves to untie or unconfound the defined dimension from specific others, lending assurance of theoretical relevance.

In a sense, the placebo problem can be understood in these terms. The

[5] For purposes of completeness, the interaction of *X* with history and maturation should be mentioned. Both affect the generalizability of results. The interaction effect of history represents the possible specificity of results to a given historical moment, a possibility which increases as problems are more societal, less biological. The interaction of maturation and *X* would be represented in the specificity of effects to certain maturational levels, fatigue states, etc.

experiment without the placebo has clearly demonstrated that some aspect of the total X stimulus complex has had an effect; the placebo experiment serves to break up the complex X into the suggestive connotation of pill-taking and the specific pharmacological properties of the drug—separating two aspects of the X previously confounded. Subsequent studies may discover with similar logic which chemical fragment of the complex natural herb is most essential. Still more clearly, the sham operation illustrates the process of X purification, ruling out general effects of surgical shock so that the specific effects of loss of glandular or neural tissue may be isolated. As these parallels suggest, once recurrent unwanted aspects of complex Xs have been discovered for a given field, control groups especially designed to eliminate these effects can be regularly employed.

Generalizing to other Os. In parallel form, the scientist in practice uses a complex measurement procedure which needs to be refined in subsequent experimentation. Again, this is best done by employing multiple Os all having in common the theoretically relevant attribute but varying widely in their irrelevant specificities. For Os this process can be introduced into the initial experiment by employing multiple measures. A major practical reason for not doing so is that it is so frequently a frustrating experience, lending hesitancy, indecision, and a feeling of failure to studies that would have been interpreted with confidence had but a single response measure been employed.

Transition experiments. The two previous paragraphs have argued against the *exact* replication of experimental apparatus and measurement procedures on the grounds that this continues the confounding of theory-relevant aspects of X and O with specific artifacts of unknown influence. On the other hand, the confusion in our literature generated by the heterogeneity of results from studies all on what is nominally the "same" problem but varying in implementation, is leading some to call for exact replication of initial procedures in subsequent research on a topic. Certainly no science can emerge without dependably repeatable experiments. A suggested resolution is the *transition experiment,* in which the need for varying the theory-independent aspects of X and O is met in the form of a multiple X, multiple O design, one segment of which is an "exact" replication of the original experiment, exact at least in those major features which are normally reported in experimental writings.

Internal vs. external validity. If one is in a situation where either internal validity or representativeness must be sacrificed, which should it be? The answer is clear. Internal validity is the prior and indispensable consideration. The optimal design is, of course, one having both internal and external validity. Insofar as such settings are available, they should be exploited, without embarrassment from the apparent opportunistic warping of the content of studies by the availability of laboratory techniques. In this sense, a science is as opportunistic as a bacteria culture and grows only where growth is possible. One basic necessity for such growth is the machinery for selecting among alternative hypotheses, no matter how limited those hypotheses may have to be.

SUMMARY

In analyzing the extraneous variables which experimental designs for social settings seek to control, seven categories have been distinguished: history, maturation, testing, instrument decay, regression, selection, and mortality. In general, the simple or main effects of these variables jeopardize the internal validity of the experiment and are adequately controlled in standard experimental designs. The interactive effects of these variables and of experimental arrangements affect the external validity or generalizability of experimental results. Standard experimental designs vary in their susceptibility to these interactive effects. Stress is also placed upon the differences among measuring instruments and arrangements in the extent to which they create unwanted interactions. The value for social science purposes of the Posttest-Only Control Group Design is emphasized.

REFERENCES

1. Annis, A. D., & Meier, N. C. The induction of opinion through suggestion by means of planted content. *J. soc. Psychol.*, 1934, **5**, 65–81.
2. Boring, E. G. The nature and history of experimental control. *Amer. J. Psychol.*, 1954, **67**, 573–589.
3. Brunswik, E. *Perception and the representative design of psychological experiments.* Berkeley: Univer. of California Press, 1956.
4. Bugelski, B. R. A note on Grant's discussion of the Latin square principle in the design and analysis of psychological experiments. *Psychol. Bull.*, 1949, **46**, 49–50.
5. Campbell, D. T. The indirect assessment of social attitudes. *Psychol. Bull.*, 1950, **47**, 15–38.
6. Campbell, D. T. On the possibility of experimenting with the "Bandwagon" effect. *Int. J. Opin. Attitude Res.*, 1951, **5**, 251–260.
7. Cantril, H. *Gauging public opinion.* Princeton: Princeton Univer. Press, 1944.
8. Collier, R. M. The effect of propaganda upon attitude following a critical examination of the propaganda itself. *J. soc. Psychol.*, 1944, **20**, 3–17.
9. Crespi, L. P. The interview effect in polling. *Publ. Opin. Quart.*, 1948, **12**, 99–111.
10. Fisher, R. A. *The design of experiments.* Edinburgh: Oliver & Boyd, 1935.
11. Gosnell, H. F. *Getting out the vote: an experiment in the stimulation of voting.* Chicago: Univer. of Chicago Press, 1927.
12. Greenberg, A. Matched samples. *J. Marketing*, 1953–54, **18**, 241–245.
13. Hartmann, G. W. A field experiment on the comparative effectiveness of "emotional" and "rational" political leaflets in determining election results. *J. abnorm. soc. Psychol.*, 1936, **31**, 99–114.
14. Hovland, C. E., Janis, I. L., & Kelley, H. H. *Communication and persuasion.* New Haven: Yale Univer. Press, 1953.
15. Hovland, C. I., Lumsdaine, A. A., & Sheffield, F. D. *Experiments on mass communication.* Princeton: Princeton Univer. Press, 1949.
16. Jahoda, M., Deutsch, M., & Cook, S. W. *Research methods in social relations.* New York: Dryden Press, 1951.
17. Kerr, W. A. Experiments on the effect of music on factory production. *Appl. Psychol. Monogr.*, 1945, No. 5.
18. Kish, L. Selection of the sample. In L. Festinger and D. Katz (Eds.), *Research methods in the behavioral sciences.* New York: Dryden Press, 1953, 175–239.
19. Lazarsfeld, P. F. Training guide on the controlled experiment in social research. Dittoed. Columbia Univer., Bureau of Applied Social Research, 1948.
20. Lucas, D. B., & Britt, S. H. *Advertising psychology and research.* New York McGraw-Hill, 1950.
21. Mayo, E. *The human problems of an industrial civilization.* New York: Macmillan, 1933.
22. McNemar, Q. A critical examination of the University of Iowa studies of environmental influences upon the IQ. *Psychol. Bull.*, 1940, **37**, 63–92.
23. Menefee, S. C. An experimental study of strike propaganda. *Soc. Forces*, 1938, **16**, 574–582.
24. Parrish, J. A., & Campbell, D. T. Measuring propaganda effects with direct and indirect attitude tests. *J. abnorm. soc. Psychol.*, 1953, **48**, 3–9.
25. Payne, S. L. The ideal model for con-

trolled experiments. *Publ. Opin. Quart.*, 1951, **15**, 557-562.

26. POSTMAN, L., & BRUNER, J. S. Perception under stress. *Psychol. Rev.*, 1948, **55**, 314-322.
27. RANKIN, R. E., & CAMPBELL, D. T. Galvanic skin response to Negro and white experimenters. *J. abnorm. soc. Psychol.*, 1955, **51**, 30-33.
28. ROSENTHAL, D., & FRANK, J. O. Psychotherapy and the placebo effect. *Psychol. Bull.*, 1956, **53**, 294-302.
29. SCHANCK, R. L., & GOODMAN, C. Reactions to propaganda on both sides of a controversial issue. *Publ. Opin. Quart.*, 1939, **3**, 107-112.
30. SOLOMON, R. W. An extension of control group design. *Psychol. Bull.*, 1949, **46**, 137-150.
31. STAR, S. A., & HUGHES, H. M. Report on an educational campaign: the Cincinnati plan for the United Nations. *Amer. J. Sociol.*, 1949-50, **55**, 389.
32. STOUFFER, S. A. Some observations on study design. *Amer. J. Sociol.*, 1949-50, **55**, 355-361.
33. STOUFFER, S. A. Measurement in sociology. *Amer. sociol. Rev.*, 1953, **18**, 591-597.
34. UNDERWOOD, B. J. *Experimental psychology.* Appleton-Century-Crofts, 1949.
35. UNDERWOOD, B. J. Interference and forgetting. *Psychol. Rev.*, 1957, **64**, 49-60.
36. UNDERWOOD, B. J. *Psychological research.* New York: Appleton-Century-Crofts, 1957.
37. WILLIAMS, R. Probability sampling in the field: a case history. *Publ. Opin. Quart.*, 1950, **14**, 316-330.

Received October 7, 1956