Experimental and Quasi-Experimental Designs for Research[1]

DONALD T. CAMPBELL
Northwestern University

JULIAN C. STANLEY
Johns Hopkins University

In this chapter we shall examine the validity of 16 experimental designs against 12 common threats to valid inference. By experiment we refer to that portion of research in which variables are manipulated and their effects upon other variables observed. It is well to distinguish the particular role of this chapter. It is *not* a chapter on experimental design in the Fisher (1925, 1935) tradition, in which an experimenter having complete mastery can schedule treatments and measurements for optimal statistical efficiency, with complexity of design emerging only from that goal of efficiency. Insofar as the designs discussed in the present chapter become complex, it is because of the intransigency of the environment: because, that is, of the experimenter's lack of complete control. While contact is made with the Fisher tradition at several points, the exposition of that tradition is appropriately left to full-length presentations, such as the books by Brownlee (1960), Cox (1958), Edwards (1960), Ferguson (1959), Johnson (1949), Johnson and Jackson (1959), Lindquist (1953), McNemar (1962), and Winer (1962). (Also see Stanley, 1957b.)

[1] The preparation of this chapter has been supported by Northwestern University's Psychology-Education Project, sponsored by the Carnegie Corporation. Keith N. Clayton and Paul C. Rosenblatt have assisted in its preparation.

PROBLEM AND BACKGROUND

McCall as a Model

In 1923, W. A. McCall published a book entitled *How to Experiment in Education.* The present chapter aspires to achieve an up-to-date representation of the interests and considerations of that book, and for this reason will begin with an appreciation of it. In his preface McCall said: "There are excellent books and courses of instruction dealing with the statistical manipulation of experimental data, but there is little help to be found on the methods of securing adequate and proper data to which to apply statistical procedure." This sentence remains true enough today to serve as the leitmotif of this presentation also. While the impact of the Fisher tradition has remedied the situation in some fundamental ways, its most conspicuous effect seems to have been to

elaborate statistical analysis rather than to aid in securing "adequate and proper data."

Probably because of its practical and common-sense orientation, and its lack of pretension to a more fundamental contribution, McCall's book is an undervalued classic. At the time it appeared, two years before the first edition of Fisher's *Statistical Methods for Research Workers* (1925), there was nothing of comparable excellence in either agriculture or psychology. It anticipated the orthodox methodologies of these other fields on several fundamental points. Perhaps Fisher's most fundamental contribution has been the concept of achieving pre-experimental equation of groups through randomization. This concept, and with it the rejection of the concept of achieving equation through matching (as intuitively appealing and misleading as that is) has been difficult for educational researchers to accept. In 1923, McCall had the fundamental qualitative understanding. He gave, as his first method of establishing comparable groups, "groups equated by chance." "Just as representativeness can be secured by the method of chance, ... so equivalence may be secured by chance, provided the number of subjects to be used is sufficiently numerous" (p. 41). On another point Fisher was also anticipated. Under the term "rotation experiment," the Latin-square design was introduced, and, indeed, had been used as early as 1916 by Thorndike, McCall, and Chapman (1916), in both 5×5 and 2×2 forms, i.e., some 10 years before Fisher (1926) incorporated it systematically into his scheme of experimental design, with randomization.[2]

McCall's mode of using the "rotation experiment" serves well to denote the emphasis of his book and the present chapter. The rotation experiment is introduced not for reasons of efficiency but rather to achieve some degree of control where random assignment to equivalent groups is not possible. In a similar vein, this chapter will examine the imperfections of numerous experimental schedules and will nonetheless advocate their utilization in those settings where better experimental designs are not feasible. In this sense, a majority of the designs discussed, including the unrandomized "rotation experiment," are designated as *quasi*-experimental designs.

[2] Kendall and Buckland (1957) say that the Latin square was invented by the mathematician Euler in 1782. Thorndike, Chapman, and McCall do not use this term.

Disillusionment with Experimentation in Education

This chapter is committed to the experiment: as the only means for settling disputes regarding educational practice, as the only way of verifying educational improvements, and as the only way of establishing a cumulative tradition in which improvements can be introduced without the danger of a faddish discard of old wisdom in favor of inferior novelties. Yet in our strong advocacy of experimentation, we must not imply that our emphasis is new. As the existence of McCall's book makes clear, a wave of enthusiasm for experimentation dominated the field of education in the Thorndike era, perhaps reaching its apex in the 1920s. And this enthusiasm gave way to apathy and rejection, and to the adoption of new psychologies unamenable to experimental verification. Good and Scates (1954, pp. 716–721) have documented a wave of pessimism, dating back to perhaps 1935, and have cited even that staunch advocate of experimentation, Monroe (1938), as saying "the direct contributions from controlled experimentation have been disappointing." Further, it can be noted that the defections from experimentation to essay writing, often accompanied by conversion from a Thorndikian behaviorism to Gestalt psychology or psychoanalysis, have frequently occurred in persons well trained in the experimental tradition.

To avoid a recurrence of this disillusionment, we must be aware of certain sources of the previous reaction and try to avoid the false anticipations which led to it. Several aspects may be noted. First, the claims made for the rate and degree of progress which would result from experiment were grandi-

osely overoptimistic and were accompanied by an unjustified depreciation of nonexperimental wisdom. The initial advocates assumed that progress in the technology of teaching had been slow *just because* scientific method had not been applied: they assumed traditional practice was incompetent, just because it had not been produced by experimentation. When, in fact, experiments often proved to be tedious, equivocal, of undependable replicability, and to confirm prescientific wisdom, the overoptimistic grounds upon which experimentation had been justified were undercut, and a disillusioned rejection or neglect took place.

This disillusionment was shared by both observer and participant in experimentation. For the experimenters, a personal avoidance-conditioning to experimentation can be noted. For the usual highly motivated researcher the nonconfirmation of a cherished hypothesis is actively painful. As a biological and psychological animal, the experimenter is subject to laws of learning which lead him inevitably to associate this pain with the contiguous stimuli and events. These stimuli are apt to be the experimental process itself, more vividly and directly than the "true" source of frustration, i.e., the inadequate theory. This can lead, perhaps unconsciously, to the avoidance or rejection of the experimental process. If, as seems likely, the ecology of our science is one in which there are available many more wrong responses than correct ones, we may anticipate that most experiments will be disappointing. We must somehow inoculate young experimenters against this effect, and in general must justify experimentation on more pessimistic grounds—not as a panacea, but rather as the only available route to cumulative progress. We must instill in our students the expectation of tedium and disappointment and the duty of thorough persistence, by now so well achieved in the biological and physical sciences. We must expand our students' vow of poverty to include not only the willingness to accept poverty of finances, but also a poverty of experimental results.

More specifically, we must increase our time perspective, and recognize that continuous, multiple experimentation is more typical of science than once-and-for-all definitive experiments. The experiments we do today, if successful, will need replication and cross-validation at other times under other conditions before they can become an established part of science, before they can be theoretically interpreted with confidence. Further, even though we recognize experimentation as the basic language of proof, as the only decision court for disagreement between rival theories, we should not expect that "crucial experiments" which pit opposing theories will be likely to have clear-cut outcomes. When one finds, for example, that competent observers advocate strongly divergent points of view, it seems likely on a priori grounds that both have observed something valid about the natural situation, and that both represent a part of the truth. The stronger the controversy, the more likely this is. Thus we might expect in such cases an experimental outcome with mixed results, or with the balance of truth varying subtly from experiment to experiment. The more mature focus—and one which experimental psychology has in large part achieved (e.g., Underwood, 1957b)—avoids crucial experiments and instead studies dimensional relationships and interactions along many degrees of the experimental variables.

Not to be overlooked, either, are the greatly improved statistical procedures that quite recently have filtered slowly into psychology and education. During the period of its greatest activity, educational experimentation proceeded ineffectively with blunt tools. McCall (1923) and his contemporaries did one-variable-at-a-time research. For the enormous complexities of the human learning situation, this proved too limiting. We now know how important various contingencies—dependencies upon joint "action" of two or more experimental variables—can be. Stanley (1957a, 1960, 1961b, 1961c, 1962), Stanley and Wiley (1962), and others have stressed the assessment of such interactions.

Experiments may be multivariate in either or both of two senses. More than one "independent" variable (sex, school grade, method of teaching arithmetic, style of printing type, size of printing type, etc.) may be incorporated into the design and/or more than one "dependent" variable (number of errors, speed, number right, various tests, etc.) may be employed. Fisher's procedures are multivariate in the first sense, univariate in the second. Mathematical statisticians, e.g., Roy and Gnanadesikan (1959), are working toward designs and analyses that unify the two types of multivariate designs. Perhaps by being alert to these, educational researchers can reduce the usually great lag between the introduction of a statistical procedure into the technical literature and its utilization in substantive investigations.

Undoubtedly, training educational researchers more thoroughly in *modern* experimental statistics should help raise the quality of educational experimentation.

Evolutionary Perspective on Cumulative Wisdom and Science

Underlying the comments of the previous paragraphs, and much of what follows, is an evolutionary perspective on knowledge (Campbell, 1959), in which applied practice and scientific knowledge are seen as the resultant of a cumulation of selectively retained tentatives, remaining from the hosts that have been weeded out by experience. Such a perspective leads to a considerable respect for tradition in teaching practice. If, indeed, across the centuries many different approaches have been tried, if some approaches have worked better than others, and if those which worked better have therefore, to some extent, been more persistently practiced by their originators, or imitated by others, or taught to apprentices, then the customs which have emerged may represent a valuable and tested subset of all possible practices.

But the selective, cutting edge of this process of evolution is very imprecise in the natural setting. The conditions of observation, both physical and psychological, are far from optimal. What survives or is retained is determined to a large extent by pure chance. Experimentation enters at this point as the means of sharpening the relevance of the testing, probing, selection process. Experimentation thus is not in itself viewed as a source of ideas necessarily contradictory to traditional wisdom. It is rather a refining process superimposed upon the probably valuable cumulations of wise practice. Advocacy of an experimental science of education thus does not imply adopting a position incompatible with traditional wisdom.

Some readers may feel a suspicion that the analogy with Darwin's evolutionary scheme becomes complicated by specifically human factors. School principal John Doe, when confronted with the necessity for deciding whether to adopt a revised textbook or retain the unrevised version longer, probably chooses on the basis of scanty knowledge. Many considerations besides sheer efficiency of teaching and learning enter his mind. The principal can be right in two ways: keep the old book when it is as good as or better than the revised one, or adopt the revised book when it is superior to the unrevised edition. Similarly, he can be wrong in two ways: keep the old book when the new one is better, or adopt the new book when it is no better than the old one.

"Costs" of several kinds might be estimated roughly for each of the two erroneous choices: (1) financial and energy-expenditure cost; (2) cost to the principal in complaints from teachers, parents, and school-board members; (3) cost to teachers, pupils, and society because of poorer instruction. These costs in terms of money, energy, confusion, reduced learning, and personal threat must be weighed against the probability that each will occur and also the probability that the error itself will be detected. If the principal makes his decision without suitable research evidence concerning Cost 3 (poorer instruction), he is likely to overemphasize Costs 1 and 2. The cards seem stacked in

favor of a conservative approach—that is, retaining the old book for another year. We can, however, try to cast an experiment with the two books into a decision-theory mold (Chernoff & Moses, 1959) and reach a decision that takes the various costs and probabilities into consideration explicitly. How nearly the careful deliberations of an excellent educational administrator approximate this decision-theory model is an important problem which should be studied.

Factors Jeopardizing Internal and External Validity

In the next few sections of this chapter we spell out 12 factors jeopardizing the validity of various experimental designs.[3] Each factor will receive its main exposition in the context of those designs for which it is a particular problem, and 10 of the 16 designs will be presented before the list is complete. For purposes of perspective, however, it seems well to provide a list of these factors and a general guide to Tables 1, 2, and 3, which partially summarize the discussion. Fundamental to this listing is a distinction between *internal validity* and *external validity. Internal validity* is the basic minimum without which any experiment is uninterpretable: Did in fact the experimental treatments make a difference in this specific experimental instance? *External validity* asks the question of *generalizability:* To what populations, settings, treatment variables, and measurement variables can this effect be generalized? Both types of criteria are obviously important, even though they are frequently at odds in that features increasing one may jeopardize the other. While *internal validity* is the *sine qua non,* and while the question of *external validity,* like the question of inductive inference, is never completely answerable, the selection of designs strong in both types of validity is obviously our ideal. This is particularly the case for research on teaching, in which generalization to applied settings of known character is the desideratum. Both the distinctions and the relations between these two classes of validity considerations will be made more explicit as they are illustrated in the discussions of specific designs.

Relevant to *internal validity,* eight different classes of extraneous variables will be presented; these variables, if not controlled in the experimental design, might produce effects confounded with the effect of the experimental stimulus. They represent the effects of:

1. *History,* the specific events occurring between the first and second measurement in addition to the experimental variable.
2. *Maturation,* processes within the respondents operating as a function of the passage of time per se (not specific to the particular events), including growing older, growing hungrier, growing more tired, and the like.
3. *Testing,* the effects of taking a test upon the scores of a second testing.
4. *Instrumentation,* in which changes in the calibration of a measuring instrument or changes in the observers or scorers used may produce changes in the obtained measurements.
5. *Statistical regression,* operating where groups have been selected on the basis of their extreme scores.
6. Biases resulting in differential *selection* of respondents for the comparison groups.
7. *Experimental mortality,* or differential loss of respondents from the comparison groups.
8. *Selection-maturation interaction,* etc., which in certain of the multiple-group quasi-experimental designs, such as Design 10, is confounded with, i.e., might be mistaken for, the effect of the experimental variable.

The factors jeopardizing *external validity* or *representativeness* which will be discussed are:

9. The *reactive* or *interaction effect* of *testing,* in which a pretest might increase or

[3] Much of this presentation is based upon Campbell (1957). Specific citations to this source will, in general, not be made.

decrease the respondent's sensitivity or responsiveness to the experimental variable and thus make the results obtained for a pretested population unrepresentative of the effects of the experimental variable for the unpretested universe from which the experimental respondents were selected.

10. The *interaction* effects of *selection* biases and the *experimental variable.*

11. *Reactive effects of experimental arrangements,* which would preclude generalization about the effect of the experimental variable upon persons being exposed to it in nonexperimental settings.

12. *Multiple-treatment interference,* likely to occur whenever multiple treatments are applied to the same respondents, because the effects of prior treatments are not usually erasable. This is a particular problem for one-group designs of type 8 or 9.

In presenting the experimental designs, a uniform code and graphic presentation will be employed to epitomize most, if not all, of their distinctive features. An *X* will represent the exposure of a group to an experimental variable or event, the effects of which are to be measured; *O* will refer to some process of observation or measurement; the *X*s and *O*s in a given row are applied to the same specific persons. The left-to-right dimension indicates the temporal order, and *X*s and *O*s vertical to one another are simultaneous. To make certain important distinctions, as between Designs 2 and 6, or between Designs 4 and 10, a symbol *R,* indicating random assignment to separate treatment groups, is necessary. This randomization is conceived to be a process occurring at a specific time, and is the all-purpose procedure for achieving pretreatment equality of groups, within known statistical limits. Along with this goes another graphic convention, in that parallel rows unseparated by dashes represent comparison groups equated by randomization, while those separated by a dashed line represent comparison groups not equated by random assignment. A symbol for matching as a process for the pretreatment equating of comparison groups has not been used, because the value of this process has been greatly oversold and it is more often a source of mistaken inference than a help to valid inference. (See discussion of Design 10, and the final section on correlational designs, below.) A symbol *M* for materials has been used in a specific way in Design 9.

THREE PRE-EXPERIMENTAL DESIGNS

1. The One-Shot Case Study

Much research in education today conforms to a design in which a single group is studied only once, subsequent to some agent or treatment presumed to cause change. Such studies might be diagramed as follows:

$$X \quad O$$

As has been pointed out (e.g., Boring, 1954; Stouffer, 1949) such studies have such a total absence of control as to be of almost no scientific value. The design is introduced here as a minimum reference point. Yet because of the continued investment in such studies and the drawing of causal inferences from them, some comment is required. Basic to scientific evidence (and to all knowledge-diagnostic processes including the retina of the eye) is the process of comparison, of recording differences, or of contrast. Any appearance of absolute knowledge, or intrinsic knowledge about singular isolated objects, is found to be illusory upon analysis. Securing scientific evidence involves making at least one comparison. For such a comparison to be useful, both sides of the comparison should be made with similar care and precision.

In the case studies of Design 1, a carefully studied single instance is implicitly compared with other events casually observed and remembered. The inferences are based upon general expectations of what the data would have been had the *X* not occurred,