

# **EXPERIMENTAL AND QUASI-EXPERIMENTAL DESIGNS FOR RESEARCH**

---

**Donald T. Campbell  
Julian C. Stanley**

# **EXPERIMENTAL AND QUASI-EXPERIMENTAL DESIGNS FOR RESEARCH**

**DONALD T. CAMPBELL**

**Syracuse University**

**JULIAN C. STANLEY**

**Johns Hopkins University**

**HOUGHTON MIFFLIN COMPANY BOSTON**  
**Dallas Geneva, Ill. Hopewell, N.J. Palo Alto London**

Reprinted from *Handbook of Research on Teaching*  
Copyright © 1963 by Houghton Mifflin Company  
All rights reserved  
Printed in U.S.A.  
Library of Congress Catalogue Card Number 81-80806  
ISBN: 0-395-30787-2

## Preface

This survey originally appeared in N. L. Gage (Editor), *Handbook of Research on Teaching*, published by Rand McNally & Company in 1963, under the longer title "Experimental and Quasi-Experimental Designs for Research on Teaching." As a result, the introductory pages and many of the illustrations come from educational research. But as a study of the references will indicate, the survey draws from the social sciences in general, and the methodological recommendations are correspondingly broadly appropriate.

For the convenience of the user we have added a table of contents, a list of supplementary references, a name index and a subject index.

DONALD T. CAMPBELL

JULIAN C. STANLEY

1966

# Contents

## PREFACE

PROBLEM AND BACKGROUND	1
McCall as a Model	1
Disillusionment with Experimentation in Education	2
Evolutionary Perspective on Cumulative Wisdom and Science	4
Factors Jeopardizing Internal and External Validity	5
THREE PRE-EXPERIMENTAL DESIGNS	6
1. <i>The One-Shot Case Study</i>	6
2. <i>The One-Group Pretest-Posttest Design</i>	7
3. <i>The Static-Group Comparison</i>	12
THREE TRUE EXPERIMENTAL DESIGNS	13
4. <i>The Pretest-Posttest Control Group Design</i>	13
Controls for Internal Validity	13
Factors Jeopardizing External Validity	16
Interaction of testing and X	18
Interaction of selection and X	19
Other interactions with X	20
Reactive arrangements	20
Tests of Significance for Design 4	22
A wrong statistic in common use	22
Use of gain scores and analysis of covariance	23
Statistics for random assignment of intact classrooms to treatments	23
Statistics for internal validity	23
5. <i>The Solomon Four-Group Design</i>	24
Statistical Tests for Design 5	25
6. <i>The Posttest-Only Control Group Design</i>	25
The Statistics for Design 6	26

<i>Factorial Designs</i>	27	
Interaction	27	
Nested Classifications	29	
Finite, Random, Fixed, and Mixed Models	31	
<i>Other Dimensions of Extension</i>	31	
Testing for Effects Extended in Time	31	
Generalizing to Other Xs: Variability in the Execution of X	32	
Generalizing to Other Xs: Sequential Refinement of X and Novel Control Groups	33	
Generalizing to Other Os	33	
 QUASI-EXPERIMENTAL DESIGNS		34
<i>Some Preliminary Comments on the Theory of Experimentation</i>	34	
7. <i>The Time-Series Experiment</i>	37	
Tests of Significance for the Time-Series Design	42	
8. <i>The Equivalent Time-Samples Design</i>	43	
Tests of Significance for Design 8	45	
9. <i>The Equivalent Materials Design</i>	46	
Statistics for Design 9	47	
10. <i>The Nonequivalent Control Group Design</i>	47	
11. <i>Counterbalanced Designs</i>	50	
12. <i>The Separate-Sample Pretest-Posttest Design</i>	53	
13. <i>The Separate-Sample Pretest-Posttest Control Group Design</i>	55	
14. <i>The Multiple Time-Series Design</i>	55	
15. <i>The Recurrent Institutional Cycle Design: A "Patched-Up"         Design</i>	57	
16. <i>Regression-Discontinuity Analysis</i>	61	
 CORRELATIONAL AND EX POST FACTO DESIGNS		64
Correlation and Causation	64	
The Retrospective Pretest	66	
Panel Studies	67	
The Lazarsfeld Sixteenfold Table	68	
Ex Post Facto Analysis	70	
 CONCLUDING REMARKS		71
 REFERENCES		71
 SOME SUPPLEMENTARY REFERENCES		77
 NAME INDEX		79

## SUBJECT INDEX

83

## TABLES

1. Sources of Invalidity for Designs 1 through 6 8
2. Sources of Invalidity for Quasi-Experimental Designs 7 through 12 40
3. Sources of Invalidity for Quasi-Experimental Designs 13 through 16 56

## FIGURES

1. Regression in the Prediction of Posttest Scores from Pretest, and Vice Versa 10
2. Some Possible Outcomes of a 3 x 3 Factorial Design 28
3. Some Possible Outcome Patterns from the Introduction of an Experimental Variable at Point *X* into a Time Series of Measurements,  $O_1-O_8$  38
4. Regression-discontinuity Analysis 62

## CHAPTER 5

# Experimental and Quasi-Experimental Designs for Research<sup>1</sup>

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.)

### 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

<sup>1</sup> 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.



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  $5 \times 5$  and  $2 \times 2$  forms, i.e., some 10 years before Fisher (1926) incorporated it systematically into his scheme of experimental design, with randomization.<sup>2</sup>

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 imper-

fections 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.

### 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-

<sup>2</sup> 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.

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 nat-

ural 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.<sup>3</sup> 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

<sup>3</sup>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 *Xs* and *Os* in a given row are applied to the same specific persons. The left-to-right dimension indicates the temporal order, and *Xs* and *Os* 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   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,

etc. Such studies often involve tedious collection of specific detail, careful observation, testing, and the like, and in such instances involve the error of *misplaced precision*. How much more valuable the study would be if the one set of observations were reduced by half and the saved effort directed to the study in equal detail of an appropriate comparison instance. It seems well-nigh unethical at the present time to allow, as theses or dissertations in education, case studies of this nature (i.e., involving a single group observed at one time only). "Standardized" tests in such case studies provide only very limited help, since the rival sources of difference other than  $X$  are so numerous as to render the "standard" reference group almost useless as a "control group." On the same grounds, the many uncontrolled sources of difference between a present case study and potential future ones which might be compared with it are so numerous as to make justification in terms of providing a benchmark for future studies also hopeless. In general, it would be better to apportion the descriptive effort between both sides of an interesting comparison.

Design 1, if taken in conjunction with the implicit "common-knowledge" comparisons, has most of the weaknesses of each of the subsequent designs. For this reason, the spelling out of these weaknesses will be left to those more specific settings.

## 2. THE ONE-GROUP PRETEST-POSTTEST DESIGN

While this design is still widely used in educational research, and while it is judged as enough better than Design 1 to be worth doing where nothing better can be done (see the discussion of quasi-experimental designs below), it is introduced here as a "bad example" to illustrate several of the confounded extraneous variables that can jeopardize *internal validity*. These variables offer plausible hypotheses explaining an  $O_1$ — $O_2$  difference, rival to the hypothesis that  $X$  caused the difference:

$O_1$   $X$   $O_2$

The first of these uncontrolled rival hypotheses is *history*. Between  $O_1$  and  $O_2$  many other change-producing events may have occurred in addition to the experimenter's  $X$ . If the pretest ( $O_1$ ) and the posttest ( $O_2$ ) are made on different days, then the events in between may have caused the difference. To become a *plausible* rival hypothesis, such an event should have occurred to most of the students in the group under study, say in some other class period or via a widely disseminated news story. In Collier's classroom study (conducted in 1940, but reported in 1944), while students were reading Nazi propaganda materials, France fell; the attitude changes obtained seemed more likely to be the result of this event than of the propaganda itself.<sup>4</sup> *History* becomes a more plausible rival explanation of change the longer the  $O_1$ — $O_2$  time lapse, and might be regarded as a trivial problem in an experiment completed within a one- or two-hour period, although even here, extraneous sources such as laughter, distracting events, etc., are to be looked for. Relevant to the variable *history* is the feature of *experimental isolation*, which can so nearly be achieved in many physical science laboratories as to render Design 2 acceptable for much of their research. Such effective experimental isolation can almost never be assumed in research on teaching methods. For these reasons a minus has been entered for Design 2 in Table 1 under *History*. We will classify with *history* a group of possible effects of season or of institutional-event schedule, although these might also be placed with *maturational*. Thus optimism might vary with seasons and anxiety with the semester examination schedule (e.g., Crook, 1937; Windle, 1954). Such effects might produce an  $O_1$ — $O_2$  change confusable with the effect of  $X$ .

A second rival variable, or class of variables, is designated *maturational*. This term is used here to cover all of those biological or

<sup>4</sup> Collier actually used a more adequate design than this, designated Design 10 in the present system.

TABLE 1  
SOURCES OF INVALIDITY FOR DESIGNS 1 THROUGH 6

	Sources of Invalidity								
	Internal								External
	History	Maturation	Testing	Instrumentation	Regression	Selection	Mortality	Interaction of Selection and Maturation, etc.	Interaction of Testing and X Interaction of Selection and X Reactive Arrangements Multiple-X Interference
<i>Pre-Experimental Designs:</i>									
1. One-Shot Case Study X O	-	-				-	-		-
2. One-Group Pretest-Posttest Design O X O	-	-	-	-	?	+	+	-	- - ?
3. Static-Group Comparison X O O	+	?	+	+	+	-	-	-	-
<i>True Experimental Designs:</i>									
4. Pretest-Posttest Control Group Design R O X O R O O	+	+	+	+	+	+	+	+	- ? ?
5. Solomon Four-Group Design R O X O R O O R X O R O	+	+	+	+	+	+	+	+	+
6. Posttest-Only Control Group Design R X O R O	+	+	+	+	+	+	+	+	+

Note: In the tables, a minus indicates a definite weakness, a plus indicates that the factor is controlled, a question mark indicates a possible source of concern, and a blank indicates that the factor is not relevant.

It is with extreme reluctance that these summary tables are presented because they are apt to be "too helpful," and to be depended upon in place of the more complex and qualified presentation in the text. No + or - indicator should be respected unless the reader comprehends why it is placed there. In particular, it is against the spirit of this presentation to create uncomprehended fears of, or confidence in, specific designs.

psychological processes which systematically vary with the passage of time, independent of specific external events. Thus between  $O_1$  and  $O_2$  the students may have grown older, hungrier, more tired, more bored, etc., and the obtained difference may reflect this process rather than  $X$ . In remedial education,

which focuses on exceptionally disadvantaged persons, a process of "spontaneous remission," analogous to wound healing, may be mistaken for the specific effect of a remedial  $X$ . (Needless to say, such a remission is not regarded as "spontaneous" in any causal sense, but rather represents the cumulative

effects of learning processes and environmental pressures of the total daily experience, which would be operating even if no  $X$  had been introduced.)

A third confounded rival explanation is the effect of *testing*, the effect of the pretest itself. On achievement and intelligence tests, students taking the test for a second time, or taking an alternate form of the test, etc., usually do better than those taking the test for the first time (e.g., Anastasi, 1958, pp. 190-191; Cane & Heim, 1950). These effects, as much as three to five IQ points on the average for naïve test-takers, occur without any instruction as to scores or items missed on the first test. For personality tests, a similar effect is noted, with second tests showing, in general, better adjustment, although occasionally a highly significant effect in the opposite direction is found (Windle, 1954). For attitudes toward minority groups a second test may show more prejudice, although the evidence is very slight (Rankin & Campbell, 1955). Obviously, conditions of anonymity, increased awareness of what answer is socially approved, etc., all would have a bearing on the direction of the result. For prejudice items under conditions of anonymity, the adaptation level created by the hostile statements presented may shift the student's expectations as to what kinds of attitudes are tolerable in the direction of greater hostility. In a signed personality or adjustment inventory, the initial administration partakes of a problem-solving situation in which the student attempts to discover the disguised purpose of the test. Having done this (or having talked with his friends about their answers to some of the bizarre items), he knows better how to present himself acceptably the second time.

With the introduction of the problem of test effects comes a distinction among potential measures as to their *reactivity*. This will be an important theme throughout this chapter, as will a general exhortation to use *nonreactive* measures wherever possible. It has long been a truism in the social sciences that the process of measuring may

change that which is being measured. The test-retest gain would be one important aspect of such change. (Another, the interaction of testing and  $X$ , will be discussed with Design 4, below. Furthermore, these reactions to the pretest are important to avoid even where they have different effects for different examinees.) The reactive effect can be expected whenever the testing process is in itself a stimulus to change rather than a passive record of behavior. Thus in an experiment on therapy for weight control, the initial weigh-in might in itself be a stimulus to weight reduction, even without the therapeutic treatment. Similarly, placing observers in the classroom to observe the teacher's pretraining human relations skills may in itself change the teacher's mode of discipline. Placing a microphone on the desk may change the group interaction pattern, etc. In general, the more novel and motivating the test device, the more reactive one can expect it to be.

*Instrumentation* or "instrument decay" (Campbell, 1957) is the term used to indicate a fourth uncontrolled rival hypothesis. This term refers to autonomous changes in the measuring instrument which might account for an  $O_1$ — $O_2$  difference. These changes would be analogous to the stretching or fatiguing of spring scales, condensation in a cloud chamber, etc. Where human observers are used to provide  $O_1$  and  $O_2$ , processes of learning, fatiguing, etc., within the observers will produce  $O_1$ — $O_2$  differences. If essays are being graded, the grading standards may shift between  $O_1$  and  $O_2$  (suggesting the control technique of shuffling the  $O_1$  and  $O_2$  essays together and having them graded without knowledge of which came first). If classroom participation is being observed, then the observers may be more skillful, or more blasé, on the second occasion. If parents are being interviewed, the interviewer's familiarity with the interview schedule and with the particular parents may produce shifts. A change in observers between  $O_1$  and  $O_2$  could cause a difference.



A fifth confounded variable in some instances of Design 2 is *statistical regression*. If, for example, in a remediation experiment, students are picked for a special experimental treatment because they do particularly poorly on an achievement test (which becomes for them the  $O_1$ ), then on a subsequent testing using a parallel form or repeating the same test,  $O_2$  for this group will almost surely average higher than did  $O_1$ . This dependable result is not due to any genuine effect of  $X$ , any test-retest practice effect, etc. It is rather a tautological aspect of the imperfect correlation between  $O_1$  and

$O_2$ . Because errors of inference due to overlooking regression effects have been so troublesome in educational research, because the fundamental insight into their nature is so frequently missed even by students who have had advanced courses in modern statistics, and because in later discussions (e.g., of Design 10 and the ex post facto analysis) we will assume this knowledge, an elementary and old-fashioned exposition is undertaken here. Figure 1 presents some artificial data in which pretest and posttest for a whole population correlate .50, with no change in the group mean or variability. (The data were

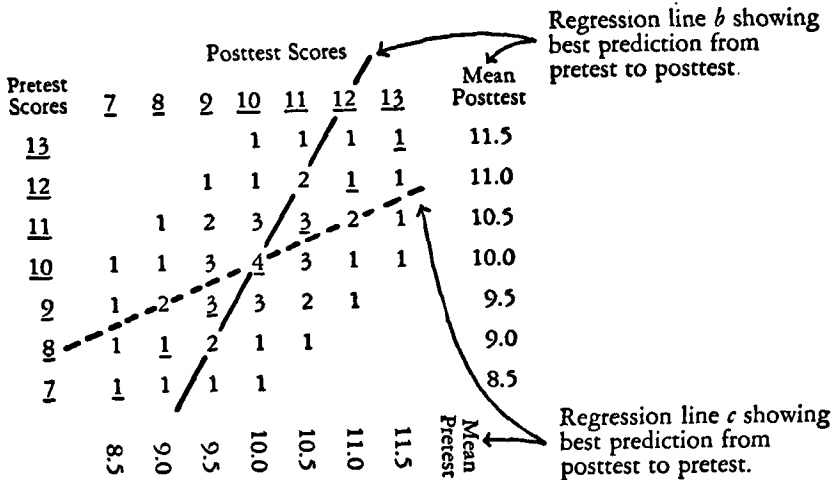


Fig. 1a. Frequency Scatter of Posttest Scores for Each Class of Pretest Scores, and Vice Versa.

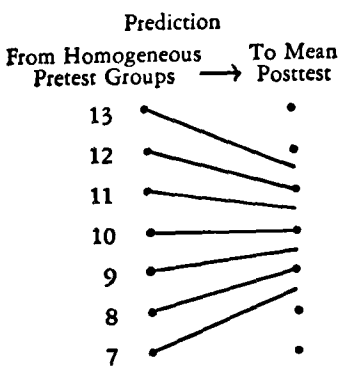


Fig. 1b.

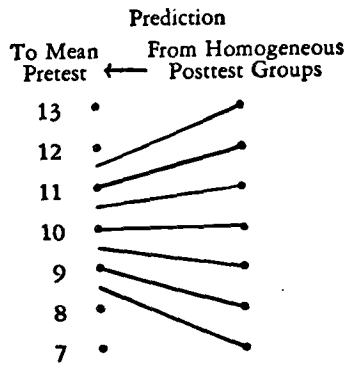


Fig. 1c.

Fig. 1. Regression in the Prediction of Posttest Scores from Pretest, and Vice Versa.