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, 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 bench mark 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:

$$O_1 \quad X \quad O_2$$

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.[4] *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 *maturation*. 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 *maturation*. This term is used here to cover all of those biological or

[4] 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
Pre-Experimental Designs: 1. One-Shot Case Study *X O*	−	−				−	−			−		
2. One-Group Pretest-Posttest Design *O X O*	−	−	−	−	?	+	+	−	−	−	?	
3. Static-Group Comparison *X O* ---------- *O*	+	?	+	+	+	−	−	−		−		
True Experimental Designs: 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.

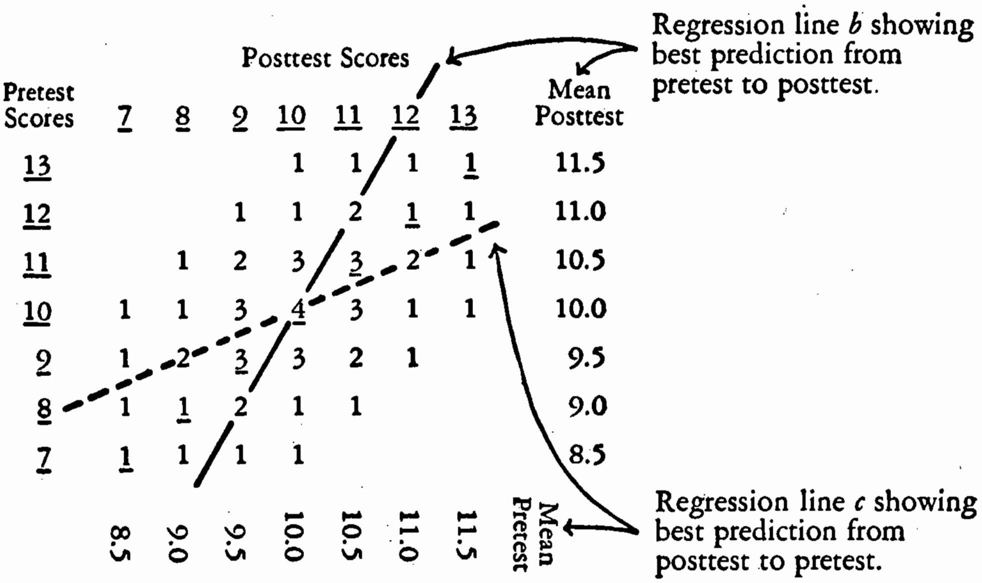
psychological processes which systematically vary with the passage of time, independent of specific external events. Thus between O_1 and O_2 the students may have grown older, hungrier, more tired, more bored, etc., and the obtained difference may reflect this process rather than *X*. In remedial education,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



Pretest Scores	7	8	9	10	11	12	13	Mean Posttest
13				1	1	1	1	11.5
12			1	1	2	1	1	11.0
11		1	2	3	3	2	1	10.5
10	1	1	3	4	3	1	1	10.0
9	1	2	3	3	2	1		9.5
8	1	1	2	1	1			9.0
7	1	1	1	1				8.5
Mean Pretest	8.5	9.0	9.5	10.0	10.5	11.0	11.5	

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

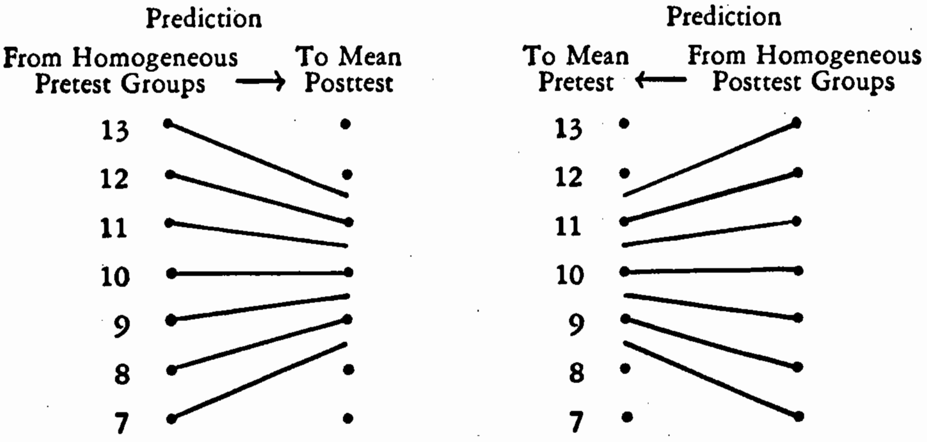


Fig. 1*b*.

Fig. 1*c*.

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

selected to make the location of the row and column means obvious upon visual inspection. The value of .50 is similarly chosen for presentation convenience.) In this hypothetical instance, no true change has taken place, but as is usual, the fallible test scores show a retest correlation considerably less than unity. If, as suggested in the example initiated above, one starts by looking only at those with very low scores on the pretest, e.g., scores of 7, and looks only to the scores of these students on the posttest, one finds the posttest scores scattered, but in general better, and on the average "regressed" halfway (i.e., the regression or correlation coefficient is .50) back to the group mean, resulting in an average of 8.5. But instead of this being evidence of progress it is a tautological, if specific, restatement of the fact of imperfect correlation and its degree.

Because time passed and events occurred between pretest and posttest, one is tempted to relate this change causally to the specific direction of time passage. But note that a time-reversed analysis is possible here, as by starting with those whose posttest scores are 7, and looking at the scatter of their pretest scores, from which the reverse implication would be drawn—i.e., that scores are getting worse. The most mistaken causal inferences are drawn when the data are presented in the form of Fig. 1b (or the top or bottom portion of 1b). Here the bright appear to be getting duller, and the dull brighter, as if through the stultifying and homogenizing effect of an institutional environment. While this misinterpretation implies that the population variability on the posttest should be less than on the pretest, the two variabilities are in fact equal. Furthermore, by entering the analysis with pure groups of posttest scores (as in regression line c and Fig. 1c), we can draw the opposite inference. As McNemar (1940) pointed out, the use of time-reversed control analyses and the direct examination for changes in population variabilities are useful precautions against such misinterpretation.

We may look at regression toward the mean in another, related way. The more deviant the score, the larger the error of measurement it probably contains. Thus, in a sense, the typical extremely high scorer has had unusually good "luck" (large positive error) and the extremely low scorer bad luck (large negative error). Luck is capricious, however, so on a posttest we expect the high scorers to decline somewhat on the average, the low scorers to improve their relative standing. (The same logic holds if one begins with the posttest scores and works back to the pretest.)

Regression toward the mean is a ubiquitous phenomenon, not confined to pretesting and posttesting with the same test or comparable forms of a test. The principal who observes that his highest-IQ students tend to have less than the highest achievement-test score (though quite high) and that his lowest-IQ students are usually not right at the bottom of the achievement-test heap (though quite low) would be guilty of the regression fallacy if he declared that his school is understimulating the brightest pupils and overworking the dullest. Selecting those students who scored highest and lowest on the achievement test and looking at their IQs would force him by the same illogic to conclude the opposite.

While regression has been discussed here in terms of errors of measurement, it is more generally a function of the degree of correlation; the lower the correlation, the greatei the regression toward the mean. The lack of perfect correlation may be due to "error" and/or to systematic sources of variance specific to one or the other measure.

Regression effects are thus inevitable accompaniments of imperfect test-retest correlation for groups *selected for their extremity*. They are not, however, necessary concomitants of extreme scores wherever encountered. If a group *selected for independent reasons* turns out to have an extreme mean, there is less a priori expectation that the group mean will regress on a second testing, for the random or extraneous sources of variance have been allowed to affect the initial scores in both directions. But for a group selected *because* of its extremity on a fallible variable, this is not the case. Its extremity is artificial and it will regress toward the mean of the population from which it was selected.

Regression effects of a more indirect sort can be due to selection of extreme scorers on measures other than the pretest. Consider a case in which students who "fail" a classroom examination are selected for experimental coaching. As a pretest, Form A of a standard achievement test is given, and as a posttest, Form B. It is probable that the classroom test correlates more highly with the immediate Form A administration than with the Form B administration some three months later (if the test had been given to the whole class on each occasion). The higher the correlation, the less regression toward the mean. Thus the classroom failures will have regressed upward less on the pretest than on the posttest, providing a pseudogain which might have been mistaken for a successful remedial-education effort. (For more details on gains and regression, see Lord, 1956, 1958; McNemar, 1958; Rulon, 1941; R. L. Thorndike, 1942.)

This concludes the list of weaknesses of Design 2 which can be conveniently discussed at this stage. Consulting Table 1 shows that there is one more minus under internal validity, for a factor which will not be examined until the discussion of Design 10 (see page 217) in the quasi-experimental designs section, and two minuses for external validity, which will not be explained until the discussion of Design 4 (see page 186).

3. The Static-Group Comparison

The third pre-experimental design needed for our development of invalidating factors is the static-group comparison. This is a design in which a group which has experienced X is compared with one which has not, for the purpose of establishing the effect of X.

$$\begin{array}{cc} X & O_1 \\ \hdashline & O_2 \end{array}$$

Instances of this kind of research include, for example, the comparison of school systems which require the bachelor's degree of teachers (the X) versus those which do not; the comparison of students in classes given speed-reading training versus those not given it; the comparison of those who heard a certain TV program with those who did not, etc. In marked contrast with the "true" experiment of Design 6, below, there are in these Design 3 instances no formal means of certifying that the groups would have been equivalent had it not been for the X. This absence, indicated in the diagram by the dashed lines separating the two groups, provides the next factor needing control, i.e., *selection.* If O_1 and O_2 differ, this difference could well have come about through the differential recruitment of persons making up the groups: the groups might have differed anyway, without the occurrence of X. As will be discussed below under the ex post facto analysis, matching on background characteristics other than O is usually ineffective and misleading, particularly in those instances in which the persons in the "experimental group" have sought out exposure to the X.

A final confounded variable for the present list can be called experimental *mortality,* or the production of O_1—O_2 differences in groups due to the differential drop-out of persons from the groups. Thus, even if in Design 3 the two groups had once been identical, they might differ now not because of any change on the part of individual members, but rather because of the selective drop-out of persons from one of the groups. In educational research this problem is most frequently met in those studies aimed at ascertaining the effects of a college education by comparing measures on freshmen (who have not had the X) with seniors (who have). When such studies show freshman women to be more beautiful than senior women, we recoil from the implication that our harsh course of training is debeautifying, and instead point to the hazards in the way of a beautiful girl's finishing college before getting married. Such an effect is classified here as experimental *mortality*. (Of course, if we consider the *same* girls when they are freshmen and seniors, this problem disappears, and we have Design 2.)