From T.D. Cook + D.T. Campbell (1979)
Quasi-Experimentation

3

Quasi-Experiments: Nonequivalent Control Group Designs

NOTATIONAL SYSTEM

This chapter discusses some quasi-experimental designs which attempt to partition respondents into nonequivalent groups that receive different treatments or no explicit treatment at all. None of the designs presented here includes time-series data. In outlining the designs we shall use a notational system in which X stands for a treatment, O stands for an observation, subscripts 1 through n refer to the sequential order of implementing treatments ($X_1 \ldots X_n$) or of recording observations ($O_1 \ldots O_n$). A dashed line between experimental groups indicates that they were not randomly formed. A wavy line between nonequivalent groups indicates that they can be considered cohorts, a term that will be defined later.

This notational system gives all the information that is required for describing experimental designs. It does not, of course, give all the information required for designing research. Indeed, a monograph on research design would have to deal with how research questions are formulated, how sampling is carried out to achieve representativeness, when causal inferences can reasonably be drawn, how data should be analyzed, and how results should be communicated to various interested parties. In this book we shall concern ourselves in detail with how the scheduling of treatments and observations helps in making causal inferences, and how the experimental data might be analyzed statistically in order to support causal inferences. All other questions of research design are treated only in the detail necessary for making our points about causal inference.

THREE DESIGNS THAT OFTEN DO NOT PERMIT REASONABLE CAUSAL INFERENCES

The three designs below are frequently used in social science research. While they are often useful for suggesting new ideas, they are *normally* not sufficient for permitting strong tests of causal hypotheses because they fail to rule out a number of plausible alternative interpretations.

It should not be forgotten that experimental design is only one way to rule out alternative interpretations and that sometimes threats can be ruled out in nondesign ways. This is especially the case when particular threats seem implausible in light of accepted theory or common sense or when the threats are validly measured and it is shown in the statistical analysis that they are not operating. With the designs under consideration, we should therefore expect some instances where few, if any, threats to internal validity are plausible even though the scheduling of treatments and measures does not by itself rule out most of the relevant threats. Though we believe that the three designs we shall examine below are *generally* uninterpretable, we urge the reader not to conclude that studies using them are *invariably* uninterpretable. Indeed, we shall later cite one example where causal inference seems reasonable.

The One-Group Posttest-Only Design

This design is diagrammed below. As can be seen, it involves making observations only on persons who have undergone a treatment, and then only after they have received it. If this were all of the information we had about the variable and about the population, the design would be totally uninformative.

X	*O*

One basic deficiency is the lack of pretest observations from persons receiving the treatment. As a result, one cannot easily infer that the treatment is related to any kind of change. A second deficiency is the lack of a control group of persons who did not receive the treatment. Without this control it is difficult to conceptualize the relevant threats and to measure them individually. In most contexts one needs time-relevant data from no-treatment control groups to check on maturational trends. One also needs information from such groups to check on any other causally irrelevant factors that could affect posttest scores and prevent one from inferring what the posttest mean would have been in the treatment group had there been no treatment.

Our predecessors have probably been mistaken in identifying the design we are discussing as "The One-Shot Case Study." Single-setting, one-time-period case studies as used in the social and clinical sciences occur in settings where many variables are measured at the posttest; contextual knowledge is already rich, even if impressionistic; and intelligent presumptions can be made about what this group would have been like without *X*. These three factors can often serve the same roles that pretest measures and control groups do more formally in more elaborate experimental designs. It may even be that in hypothesis-testing case studies, the multiple implications of the thesis for the multiple observations available generate "degrees of freedom" analogous to those coming from numbers of persons and replications in an experiment (Campbell, 1975). However that may be, and while recognizing that the epistemology of humanistic scholarly approaches needs much further elaboration, certainly the case study as normally practiced should not be demeaned by identification with the one-group posttest-only design. However, one would often recommend with the case study that scholarly effort should be redistributed so as to provide explicit evidence about conditions prior to the presumed cause and about contemporary conditions in social settings without the treatment that are similar to the setting in which the case study is taking place. All inference is comparative, and it is usually optimal to have comparable sorts of evidence, comparable degrees of detail and precision, about conditions prior to the implementation of a treatment and about factors that occur simultaneously with the treatment.

A consideration of some of the conditions under which posttest observations on a single treatment group may result in reasonable causal inference may be of help at this point. Scriven's (1976) concept of the "modus operandi" approach provides one valuable perspective, based upon considering police detective approaches. When a thief commits a crime he leaves behind a series of clues which perhaps indicate when he entered a building, by which means he entered, which types of goods he stole or left behind, and so on. The experienced detective examines these clues and relates them to what is known about the preferred operating style of identified burglars, ruling out some potential suspects because their usual mode of breaking and entering differs from what is observed. Having reduced the universe of likely suspects (i.e., causes), the detective then searches for further clues in the hope of narrowing down the number of suspects to one or two. The researcher can sometimes function as a detective, noting the level of different variables and using this information to rule out some threats to both internal and construct validities. For example, in education a case study might reveal that a new mathematics curriculum stresses algebra over geometry and arithmetic and that, after completing the curriculum, a particular group of children scores well above national norms in algebra but not in geometry and arithmetic. To the extent the researcher can rule out alternative possibilities (e.g., the difference is due to chance, or existed before the new curriculum began), it is provisionally warranted to infer causation.

Three points need stressing about the modus operandi example we have just cited. First, the experimental design is no longer that of a simple case study with a single dependent variable measured at one time. The design has become more complex and has many dependent variables that are expected to have different levels. (Making the design even more complex by adding respondents' posttest recollections of how they were at the pretest can also help. If one is prepared to assume that such retrospective pretests are valid, one can then examine the data for differences in the amount of change between different constructs.)

Second, some causes, like some burglars, leave their unique "signature" on the effect. For example, if a specific poison is found in the victim's blood, compelling causal inferences can be established if the prevalence of that chemical is rare, if its sources are few or unique, and if spatial and temporal contiguity can be established that links the victim to sources of that particular chemical. On further examination, such cases turn out to be those in which the causal hypotheses involved are already established, and the "base rates" for their distribution known. To establish that a single factory using vinyl chloride is the culprit for neighborhood and employee cancers requires a much less elaborate experimental design than one set up to determine which air and water pollutants may be cancer producing. The notion of "signed causes" may be usefully extended even back into laboratory research. Consider the one-group posttest-only design in a study of the effects of rehearsing a nonsense syllable list when compared to a study of the effects of a movie on some social attitude. The nonsense syllable list is a ''signed cause,'' and the base rate for people at large being able to recite these 12 nonsense syllables in the specific order in which they were presented in an experiment can be assumed to be extremely low. But for the social attitude example, innumerable communications in ordinary daily life are potential causes. To give an attitude survey only to those who have just seen the movie would normally tell us nothing about the movie's impact on the attitude. (However, the illustrations that respondents use in open-ended responding and the justifications for the attitudes that respondents advance might turn up plausible ''signed causes'' testifying to the movie's impact as opposed to the impact of other attitude-influencing forces.)

Third, one point can be made about both simple case studies with a single novel dependent variable (e.g., a nonsense syllable list or death due to vinyl chloride poisoning) and more complex case studies with different predictions for different dependent variables. Their interpretability depends on being able to make confident inferences about change and on being able to rule out more threats to internal validity than is normally the case with a simple case study that has a single outcome variable which can be multiply influenced. The functional requirements are to assess change and to rule out alternative interpretations. Though the case study and other one-group posttest-only designs are generally poor at achieving these ends, they are not always so. At times multiple dependent variables are available and realistic assumptions about change and alternative interpretations of change can be made. Moreover, the case study is useful for purposes unrelated to inferring causation, such as assessing whether there was a treatment and how well it was delivered, or generating new hypotheses about the phenomenon under investigation.

The Posttest-Only Design with Nonequivalent Groups

Often a treatment is implemented before the researcher can prepare for it, and so the research design is worked out after the treatment has begun. Such research is often said to be ex post facto. However, research of this kind does not necessarily imply the absence of pretest observations, for archival records can often be used to establish what the pretest scores of the various experimental units were. We shall understand ex post facto here in a more restricted sense than is often used—as research where there are no pretest observations on the same or equivalent scales for which posttest observations are available.

If we add to the case study a single nonequivalent control group that does not receive the treatment, we arrive at the design diagrammed below.

```
X   O
- - - - -
    O
```

Its most obvious flaw is the absence of pretests, which leads to the possibility that any posttest differences between the groups can be attributed either to a treatment effect or to *selection* differences between the different groups. The plausibility of selection differences in research with nonequivalent groups usually renders the design uninterpretable.

The posttest-only design with nonequivalent groups can be more complicated than appears above, especially if multiple groups are involved that receive the treatment with different dosages. This would be the case, for instance, if one nonequivalent group of parolees had one year's counseling, another group had nine months, another group six months, another three months, and another none. Such a design would take the form below, with the subscript indicating the treatment period in months.

$$\begin{array}{cc} \hline X_{12} & O \\ \hdashline X_{9} & O \\ \hdashline X_{6} & O \\ \hdashline X_{3} & O \\ \hdashline X_{0} & O \\ \hline \end{array}$$

If the length of the treatment were related to posttest scores, this would be consistent with the possibility that the treatment had causally influenced the outcome measure, say, recidivism. But it would not be strong evidence, for one would have to rule out the possibility that the persons least likely to go back to prison were selected for the longer parole period. This could happen because the persons making the selection wanted the parole counseling to appear beneficial or because the ex-offenders who were likely to stay in the counseling treatment longer were those least likely to drop out of the experiment as a result of committing crimes.

A tradition has developed in economics and sociology of trying to overcome the lack of pretest measures by seeking out pretest measures which correlate with the posttest within experimental groups *but are not measured on the same scale as the posttest*. This means that more easily retrieved measures such as age, sex, social class, race, place of birth, or residence are substituted for the absent pretest. We shall leave discussion of this particular modification of the posttest-only design until we deal with designs that have both pretests and nonequivalent groups.

The One-Group Pretest-Posttest Design

This design is one of the more frequently used designs in the social sciences and is diagrammed below.

$$\begin{array}{ccc} \hline O_1 & X & O_2 \\ \hline \end{array}$$

It can be seen that pretest observations *(O_1)* are recorded on a single group of persons, who later receive a treatment *(X)*, after which posttest observations are made *(O_2)*. Since the use of this design is so widespread, we would like to illustrate its weaknesses using a hypothetical example.

Imagine the case where the supervisory style of foremen is altered in a work setting and where the change is expected to increase productivity. Imagine, further, that the posttest level of productivity is reliably higher than the pretest level. One might want to attribute such an increase to the change in supervisory style. However, the change might alternatively be due to *history* in the sense that other events could have happened between the pretest and posttest that affected productivity. Some of these could have occurred within the work setting (e.g., a new salary scale might have been implemented, union policies might have changed, or a new training program might have been introduced). Other events could have taken place outside of the work setting (e.g., a new export drive might have been started nationally, or the weather might have become warmer, allowing workers to feel better or become better acquainted). Any of these events, or others, *could* have affected productivity. In order to rule them out, the researcher has to make the case either that they are implausible in the particular context of a given study or that they are plausible but did not actually operate. Data are usually required for making the second case, and either common sense, or theory, or experience are required for buttressing the argument of implausibility. If he or she cannot rule out a particular history threat, then the researcher has to admit that he or she cannot draw confident causal conclusions because a frequently plausible threat cannot be ruled out.

Consider, next, *statistical regression.* Why should supervisory styles be changed? One reason might be that productivity is low and needs to be increased. Now, productivity in any one year can be low because of a genuinely stable decline in productivity or because productivity has been consistently low for some time. But it might also be *low in any one year* because of random factors like an atypical strike or delays in the delivery of raw materials or other "errors" in the method of measuring productivity. Such random factors mean that the impetus for change will be negatively correlated with productivity, and this is tantamount to deliberately choosing a year of extremely low productivity for conducting one's experiment. What will happen in such a case is that productivity will probably increase in the next year as it regresses towards the grand mean of the productivity trend. In other words, by choosing to change one's work practices when productivity is low one can capitalize upon random fluctuations in productivity.

A more common form of regression artifact for this design arises when a special program is given only to those with extreme scores on the pretest, as would happen with a compensatory education program that is given only to low-achieving children. Selecting out low scorers will produce a spurious improvement if the pretest-posttest correlation is less than 1.00. The magnitude of the spurious regression effect will depend on (a) how far the correlation is below unity (McNemar, 1940) and (b) how far the low scorers are below their population mean. In a similar vein, if an advanced training program were given to the best salespeople of one year and was in fact totally ineffective, it would nonetheless probably appear to reduce the sales volume of its graduates when their pretest-posttest performance is compared to that of other salespeople. What is important to note is that the amount of regression is negatively related to how highly each salesperson's sales volume is correlated from year to year, and is positively related to how far the pretest year deviates from the average sales volume of the average salesperson.

Even when the pretest-posttest correlation is high, a posttest increase in productivity could be accounted for in terms of *maturation*. Typically, productivity levels do not stay constant from year to year, being subject to systematic fluctuations as well as to the random fluctuations that lead to statistical regression. Whenever productivity is systematically rising over the years or is subject to systematic cyclical fluctuations within any one year, a posttest increase over pretest levels can appear in the design under discussion. This increase would not be unambiguously attributable to the supervisory change. It could alternatively be attributed to workers becoming more experienced, or to machinery becoming more and more sophisticated, or to the average height and weight of United States males being on the increase, or whatever. This threat is particularly relevant because many of the factories that allow outside investigators to conduct research may be proud of the fact that they are getting better and better. Of course, productivity sometimes systematically decreases, and cyclical trends decrease as well as increase. Hence, with the design under discussion, maturation can sometimes lead to spurious decreases as well as spurious increases.

In some contexts there are nondesign ways of directly estimating whether maturation could plausibly account for pretest-posttest differences. Consider the productivity example again; imagine that the pretest and posttest are separated by a year and that the mean level of experience (years worked in an organization) increases by a year between pretest and posttest. If the pool of *pretest scores were sufficiently large,* one could regress productivity onto years worked. If the resulting regression line were flat and the correlation of experience and productivity was therefore zero, then the maturation hypothesis would be rendered implausible because it presupposes that the experience gained during a year's work affects productivity. However, if there were a simple linear relationship, one could then use the regression equation derived from the pretest scores in order to predict the productivity change expected in a year. This could then be used to assess if the one-year change predicted from the pretest scores alone was different from the actual change obtained during the year.

Though it is useful, care must be taken with such an estimation procedure. First, with obtrusive measures a problem arises because the expected posttest performance is derived from a pool of pretest scores where measurement has taken place only once. Consequently, if the expected and obtained scores differed, this might be either because the posttest was affected by the treatment or because knowledge gained at the first testing altered performance on the subsequent testing. Second, problems can arise if the expected maturation is assessed from a pool of persons who contribute pretest scores and if some of these persons then drop out of the study so that the pretest and posttest groups are less than totally comparable. Differences between expected and obtained scores might then be due to selection. Finally, it is worth noting that the mean posttest experience must inevitably be different from the mean pretest experience if maturation is occurring and the analyses are restricted to persons who provide both pretest and posttest data. As a result, the pretest data cannot be used to assess the relationship between experience and productivity for that part of the experience distribution which is represented in the posttest but not in the pretest. All one can do is assume that the relationship in this part of the distribution can be inferred by extrapolating from the way in which productivity and experience are related *for the particular range of pretest experience scores.*

Though we have focused on history, maturation, and regression as frequent competing explanations in the one-group pretest-posttest design, it would be incorrect to think that they are the only relevant ones. *Testing* is an obvious addition to the list. This is because exposure to an outcome measure at one time can lead to shifts in performance at another. For instance, in education a pretest can be the impetus to learning the correct answers to items and thus increase the posttest level of performance. *Instrumentation* can also be a threat if, for reasons that are relevant or irrelevant to the experiment, the definition of an outcome measure is changed. This might happen, for example, if the definition of what constitutes a "serious" traffic accident is modified in some record-keeping system.

Occasionally there will be specific settings where the threats of history, maturation, regression, instrumentation, and testing are implausible or can be convincingly ruled out via direct measurement. In those settings, the one-group pretest-posttest design will be interpretable. Unfortunately, these settings are likely to be rare. To rule out effects of history, the respondents would have to be *physically isolated* from historical forces that affect productivity. To rule out statistical regression, we would require a *series of pretest observations* from which it could be inferred that the introduction of the treatment was not associated with extreme values of the pretest. However, since the one-group pretest-posttest design is defined as having only one pretest observation, nothing can be known of the pretest trend except in very rare circumstances where the phenomenon under investigation is known to be nonchanging. Alternatively, regression would be implausible if the pretest-posttest correlation were close to 1.00, since the magnitude of regression depends upon the unreliability of the measures and unreliable measures have lower test-retest correlations. Though high test-retest correlations are not unknown, they are often closer to .40 than to 1.00 with social measures and time intervals of about a year. As for maturation, *a long series of pretest observations* would obviously permit testing the threat sensitively, but it is by definition not available. A less sensitive test would be to use a regression procedure like the one described for the purpose of estimating maturation numerically, though this requires accepting certain assumptions. If neither of these empirical strategies were feasible, one would then have to deal with a maturation threat by using common sense, theory, or experience to assess the plausibility of assuming whether maturation was or was not occurring in the specific context of a specific research project.

Consider how much more fortunate the physicist is when compared to the behavioral scientist who works in field settings. The physicist can use a laboratory to create physical isolation, and he or she often works with objects that do not change over the time period of an experiment. This being so, history, maturation, and regression are not problematic, and the data from single-group pretest-posttest changes are often causally interpretable. Consider, next, the cultural anthropologist who wants to investigate how a new tool (e.g., the axe) has affected a remote tribe that has been "untouched by the modern world." This researcher, too, can use physical isolation (e.g., a dense jungle setting) and the presumption of stable pretest trends with respect to outcomes in order to help causally interpret certain kinds of pretest-posttest shifts. The social scientist, on the other hand, has fewer of the advantages of the physicist or the cultural anthropologist working in "remote" areas, for the social scientist is trying to answer causal questions in more complex social settings where the entities being studied are clearly amenable to change for reasons that have nothing to do with the experiment.

The practical question is, therefore: When can physical isolation be achieved or assumed by the social scientist? When can he or she assume stable pretest trends given that no pretest data are available for empirically assessing the trend? In our opinion, social scientists working in field settings will rarely be able to give confident answers to these questions *unless they are working with novel outcome variables and short pretest-posttest time intervals*. This means that we should usually not expect hard-headed causal inferences from the simple before-after design when it is used by itself, though inferences may be possible under the special conditions noted above. However, we will often achieve some knowledge using the design, even when pretest-posttest intervals are long and the outcome variables are subject to multiple influence other than the treatment. This is because the design permits ruling out some competing threats to validity and it often suggests hypotheses worth further exploration. Our hope is that persons considering the use of this design will hesitate before resorting to it and will decide to incorporate into it some of the design adjuncts we shall mention later.

SOME GENERALLY INTERPRETABLE NONEQUIVALENT CONTROL GROUP DESIGNS

In this section, we shall distinguish eight kinds of generally interpretable nonequivalent control group designs. These can be labeled (a) no-treatment control group designs, (b) nonequivalent dependent variables designs, (c) removed-treatment control group designs, (d) repeated-treatment designs, (e) reversed-treatment nonequivalent control group designs, (f) cohort designs, (g) posttest-only designs with predicted higher-order interactions, and (h) regression-discontinuity designs. These designs are not in any way qualitatively different from the ones we labeled "generally interpretable." They merely make it possible to rule out more threats to internal validity. Readers would do themselves a disservice if they treated the designs to follow as qualitatively distinct from those already presented.

Our separate discussion of the designs that follow should not blind the reader to the importance of incorporating more than one of them into the work that he or she does. The designs have different strengths and weaknesses, and their creative mixture within a single study can significantly increase our confidence in making causal attributions. We hope, therefore, that the reader will pay particular attention to some of the studies we shall deal with in detail (e.g., Broadbent and Little, 1960; Lawler and Hackman, 1969; and Lieberman, 1956). These studies were improved by creatively mixing the design features that we shall treat separately for pedagogic convenience alone.

The Untreated Control Group Design with Pretest and Posttest

The design is diagrammed below. It is perhaps the most frequently used design in social science research and is fortunately often interpretable. It can, therefore, be recommended in situations where nothing better is available. Because of its frequent usage, we shall deal with it in considerable detail.

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

Our discussion will be in two parts. In this chapter, we shall detail how much the interpretation of outcomes from this design depends on the particular pattern of findings. To do this, we shall discuss five different outcomes, using the first one to illustrate the most likely threats. For heuristic reasons, the discussion will be phrased in terms of group differences in pretest-posttest gains. In the next chapter, four ways of statistically analyzing the data from the basic design will be outlined, and it will be concluded that a simple analysis of pretest-posttest gain is normally inappropriate. Thus, our heuristic focus on gains in this chapter should not be interpreted to imply that the data from the pretest-posttest design with nonequivalent groups should be analyzed as simple gain scores.

Outcome 1. The basic design under discussion usually controls for all but four threats to internal validity. One uncontrolled threat is that of *selection-maturation*. This arises when the respondents in one group are growing more experienced, more tired, or more bored than the respondents in another group. To help understand this, imagine the situation where a new practice is introduced into one of two settings where identical tasks are being performed and where the treatment group outperforms the controls at the pretest. If the treatment increased productivity, we would expect a posttest difference between groups that was larger than the pretest difference, as Figure 3.1 illustrates. But we would also expect this pattern of data if the treatment and control groups differed because the former were, say, brighter on the average and were using their aptitude to gain new knowledge at a faster rate than the controls.

Figure 3.1 has been drawn to illustrate the special situation where there is no growth at all among the controls who have reached a stable level of performing by the pretest. When the data are of this form, the major issue that the investigator has to face is: How plausible is it to postulate causally irrelevant growth patterns that only affect the experimental group? In many instances, it will be much easier to think of reasons why the experimentals and controls should be maturing at different rates *in the same direction* than it will be to think of reasons why one group should be changing in one direction while the other group is not changing at all. For instance, different growth rates in the same direction are common in education contexts or in contexts where people are expected to gain through experience. In other instances, it will be easier to think of reasons why neither group should be growing than to think of reasons why one should be growing and the other not. If plausibility or, preferably, the direct analysis of pretest data to assess growth rates within conditions indicates that no growth would be expected in either group during the experiment, then when the experimental outcomes are as depicted in Figure 3.1, one need not worry too much about selection-maturation.

However, if the threat cannot be ruled out, despite the absence of all measured pretest-posttest growth in the controls, then it must be taken seriously.

A second problem arises with *instrumentation.* It is not clear with many scales that the intervals are equal, and change is often easier to detect at some points on a scale than others. Scaling problems are presumably more acute the greater the nonequivalence of the experimental groups and the farther apart they are on the scale, especially if any of the group means approaches one end of the scale where ceiling or floor effects are likely. An inspection of the pretest and posttest frequency distributions within each treatment group will suggest whether instrumentation problems are plausible. If they are, the distributions will be skewed and/or the group means and variances will be correlated. Sometimes, the raw data can be rescaled so as to reduce the problem, while at other times a careful choice must be made of intact and unmatched groups that score at about the middle of a scale and close to each other.

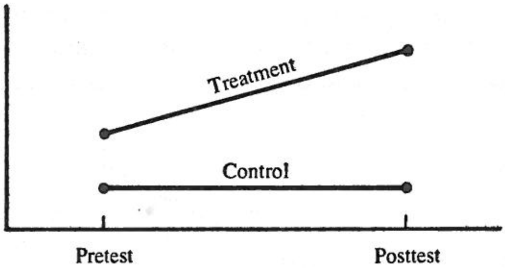


Figure 3.1. First outcome of the no-treatment control group design with pretest and posttest.

A third problem has to do with *differential statistical regression.* Consider the consequences of having a predetermined experimental group of low scorers (say, children eligible for Head Start) and then selecting no-treatment controls from the children nearby so that their scores will be as close as possible to the experimental mean, yet still above it on the average since they are not eligible for Head Start. Knowing this, the researcher might put in a special (but misguided) effort to select as controls children with particularly low scores, arguing that the best approximation for creating equivalent groups is to form groups with minimal differences. But the deliberate selection of low scorers is a form of matching which will result in the control group mean regressing to its population baseline, as in Figure 3.1. The experimentals, on the other hand, would not be expected to change since they were not selected for their low scores. Such differential regression could obviously result in the Figure 3.1 pattern of outcomes.

A fourth problem relates to the interaction of selection and history. We shall call this *local history,* events other than the treatment which affect the experimental group but not the control group, or vice versa. Imagine introducing some participative decision-making procedure to a group of day workers while leaving decision making as it was among night workers. Imagine, further, that the investigator was interested in seeing whether participative decision making induced higher work morale. Now, if the experiment started in the early spring and ended by midsummer we might expect the increasingly warm weather to have a greater effect on the work morale of day workers than of night workers. This is only one of many possible examples, of course, and the plausibility of a local history explanation has to be examined within the particular context of specific research settings when the nonequivalent control group design is used.

Outcome 2. There is a pattern of selection-maturation interaction which is both more common and more lawful than the one that would be represented by Figure 3.1. This pattern occurs when nonequivalent groups are growing at different average rates in a common direction, as in Figure 3.2. Such differences in growth rate will usually be reflected in pretest differences. When the differential growth continues for the total course of an experiment (in the absence of other forces which affect observed growth, such as ceiling effects), it will result in larger posttest than pretest differences between groups. This pattern will have nothing to do with the effects of a treatment, but it may seem to the unwary as if it does.

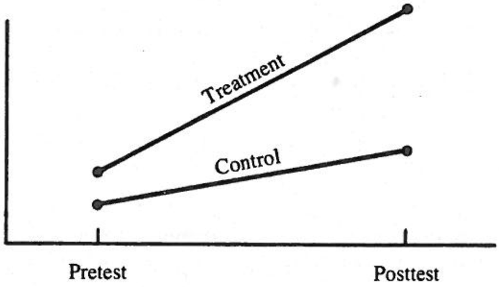


Figure 3.2. Second outcome of the no-treatment control group design with pretest and posttest.

It is common in quasi-experimental research in many substantive areas to find differences in growth rates, particularly when respondents self-select themselves into receiving a treatment. For when self-selection occurs, treatments are more likely to become available to the specially meritorious or to persons with keen desires to "improve themselves." Since the "meritorious" or the "keen" will usually be intrinsically more able or more exposed to opportunities for change, the "meritorious" or "keen" will change faster over time. Thus a selection-maturation difference can masquerade as a treatment effect.

Several clues are available for assessing whether nonequivalent groups may be maturing at different rates over time. First, if the group mean differences are a result of biased social aggregation or selection only, then the differential growth between groups should also be occurring within groups. To help understand this, imagine two groups of children, one of which receives an educational treatment and is composed of children with more ability, on the average, than the controls. The experimentals might very well gain more than the controls over time for reasons associated with the children's innate abilities. But more importantly for our present purposes, on many measures of educational achievement we would expect the more able among the experimentals to gain more than the less able *among the experimentals*. The point is that many patterns of selection-maturation should lead to increased *within-group* variances at the posttest when compared to the pretest. In estimating the plausibility of a selection-maturation threat, it is always important to inspect the variances, particularly when the obtained outcomes are like those depicted in Figure 3.2. Second, the plausibility of selection-maturation can be estimated by plotting *pretest* outcome scores against the maturational variable (e.g., age or years of experience) for the experimental and control groups separately. If the regression lines differ in linear slope, this is presumptive evidence of differential average growth rates.

The growth pattern we have been describing up to this point is one associated with phrases like "the rich get richer" or "the able become more able." This pattern is characterized by a constant increase over time in the mean difference between nonequivalent groups and by within-group variances that increase with their respective group means. But there are many, many ways in which two or more groups can differ from each other over time in the absence of a treatment effect, and there is no *logical* need for the difference between groups to be increasing at a constant rate. Instead, much more complex patterns of growth differences can be imagined (e.g., simple linear growth in one condition and quadratic growth in another, or quadratic growth in one condition and cubic in another). However, in our experience (largely influenced by educational experiments), it is more common to have linearly increasing group differences over the time span of an experiment than it is to have more complex forms of group differences. But what is most common in our experience is not synonymous with what is possible in the context of a particular research problem.

Since it is important to be able to identify the particular pattern of maturation that would be expected in each group in the basic research design, attempts should be made to estimate at least the gross pattern of maturational differences. As we have seen, the pretest data are sometimes appropriate for this. So, too, is background theory, for descriptive longitudinal data sometimes exist from populations similar to those in an experiment. Consider the case in education. Here, many data sets describe how those who initially had a higher level of achievement came to grow further ahead of their lower-scoring contemporaries on most achievement measures, resulting in a constant increase in group differences. But for certain skills (e.g., the notion of conservation) we suspect that learning gains may come about abruptly as children reach the "stage" where the skill can be learned. For skills like conservation, constant increases would not be expected. Instead, sharp discontinuities in the growth pattern would be expected in each group and at different times across groups.

Unfortunately, the necessary longitudinal descriptive information is not available in other substantive areas, and estimates of group differences in maturationa patterns will have to be made on the basis of experience or guesswork. Neither o these is, alas, suitable for numerically estimating particular group growth rates i a specific study (e.g., estimates of the form: the change in income for each montl of age is X dollars). Nor are they suitable for even estimating gross differences i pattern (e.g., the average expected change in income is greater in group A than i group B).

Thus far, we have treated selection-maturation as though it were the onl threat to internal validity that could operate if the outcome pattern in Figure 3. resulted. This is clearly not the case. All the other threats listed for the outcome i Figure 3.1 apply to Figure 3.2. Take local history as an example. One group ca obviously experience different local events between the pretest and posttest, an these can masquerade as treatment effects. Take instrumentation as another example. Since nonequivalent groups usually differ at the pretest, there is the distinc possibility of floor or ceiling effects that affect one group differently from another. In some educational experiments the higher-scoring group's true achievement shif from pretest to posttest may be underestimated because of a floor effect. Actually, these threats do more than make the interpretation of the data pattern in Figure 3. hazardous. They also make it difficult to use the pretest data for estimating th gross expected maturation pattern within each treatment group.

Basically, Figures 3.1 and 3.2 differ in their implications for how plausible i the threat of differential growth. Figure 3.1 suggests that no change would b expected between the pretest and the posttest, and this increases the likelihood tha the observed change in the experimental group is due to the treatment. However, to accept this assumption about the control group the researcher has to ask: (1 "Might there have been change in the control group to which the measures wer not sensitive?" If the answer to this is negative, the researcher also must ask (2 "Though no change occurred among the controls, how sure can I be that ther would also have been no change in the nonequivalent treatment group in th absence of a treatment?" Figure 3.2 suggests that the no-treatment controls migh well be changing over time and that the research is not tapping into a no-chang situation. It does more than that, however. It also suggests that the group wit greater pretest advantages may be changing at a faster rate than the group wit fewer pretest advantages. This being so, the researcher has to ask: "How can test whether there is differential growth?" and "How can I estimate the pattern o the differential growth?" We shall see in the next chapter that differential chang patterns of the form "the rich get richer" (i.e., a constant increase in group mea differences over time) are more amenable to statistical analysis than other forms o differential change. Whatever the thoughts one has about selection-maturation, one cannot afford to ignore local history, differential instrumentation shifts, differential testing, and differential statistical regression as other possible interpretations of the outcome patterns depicted in both Figure 3.1 and Figure 3.2.

Outcome 3. Our discussion of the nonequivalent no-treatment control grou design has thus far focused on the outcome where the treatment group is superio to the controls at the pretest and appears to be even more superior at the posttest. Let us now look at Figure 3.3 which shows the related outcome where the pretest superiority is diminished or eliminated by the posttest.

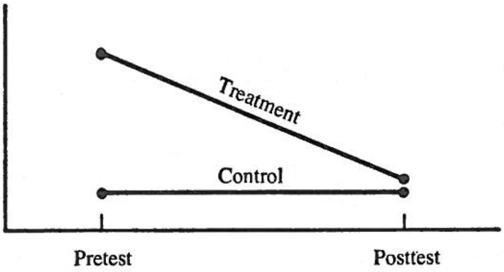


Figure 3.3. Third outcome of the no-treatment control group design with pretest and posttest.

This particular outcome was obtained from a sample of black third, fourth, and fifth graders in a study of the effects of school integration on academic self-concept (Weber, Cook and Campbell, 1971). At the pretest, black children who attended all-black schools had a higher academic self-concept mean than black children who attended integrated schools in the same school district. But after formal school integration had taken place, the initially segregated and initially integrated black children did not differ. While the basic logic of experimental design with control groups involves starting with equality between groups and finishing with differences between them, we should be alert to ''catch up'' designs in which the ''control group'' already has the treatment which the experimental group receives between pretest and posttest. Of course, all of the problems described for Figures 3.1 and 3.2 are still relevant, especially the possibility of a selection-maturation interaction.

Outcome 4. A fourth possible outcome of the no-treatment control group design with pretest and posttest is depicted below. Its salient characteristics are that the controls initially outperform the experimentals and that the difference between experimentals and controls is greater at the pretest than the posttest. This is a particularly interesting outcome since it is the one desired when organizations introduce compensatory inputs to increase the performance of groups who have started out at a disadvantage (as in some educational contexts) or where performance did not seem up to par for other reasons (as happens in industry when changes are made to improve poor performances).

This outcome is subject to the typical scaling (i.e., instrumentation) and local history (i.e., selection × history) threats that were discussed earlier. But two special aspects stand out. First, regression is more of a threat than it is when the outcome is as depicted in Figure 3.1. When no attempt has been made to match groups and one is instead dealing with stable group differences, regression is normally not a threat if the research outcome is similar to the one depicted in Figure 3.1. This is because we would have no reason to expect respondents in the treatment condition to regress upwards from their higher pretest scores.

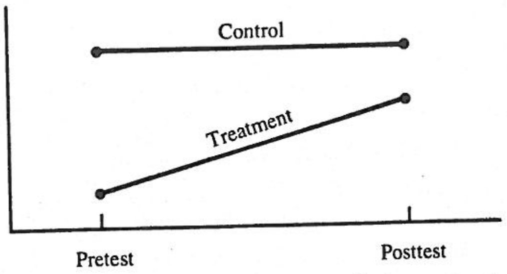


Figure 3.4. Fourth outcome of the no-treatment control group design with pretest and posttest.

But the situation can be very different, as in the case illustrated in Figure 3.4, if the treatment is deliberately given to experimental group members *because their scores are unexpectedly low*. This will lead to regression upwards by the posttest and to the data pattern shown in Figure 3.4. We would expect such regression because there is likely to be an especially large error component determining an unexpectedly low pretest mean (as opposed to a low pretest mean which reflects a stable pattern of poor performance and measures having a large true score and a small error variance component). To assess the likelihood of statistical regression, it is important to explore *why* low-scoring groups are assigned a particular treatment.

Second, the outcome in Figure 3.4 is particularly useful since it rules out the previously mentioned cumulative pattern of selection-maturation where persons scoring lower at the pretest could be expected to be even further behind at the posttest. When this common maturational pattern is presumed to operate, the outcome in Figure 3.4 would imply that the treatment had had an effect *despite* the lower expected pretest-posttest change among respondents in the treatment group. In other words, the treatment was so powerful that it overcame a countervailing force which tended to obscure a true treatment effect. Of course, it should not be assumed that all maturational trends follow the pattern of the higher-scoring group spontaneously changing faster than the lower-scoring group. This assumption has to be checked against the growth patterns reflected in the pretest scores or against any general laws that might be applicable to a given research area. Nonetheless, outcome four rules out all possibilities of differential linear growth that are based upon members of the higher-scoring pretest group growing at a faster average rate than members of the lower-scoring pretest group. Less common patterns of selection-maturation would have to be invoked as alternative explanations of the findings in Figure 3.4.

Outcome 5. Bracht and Glass (1968) have noted the desirability of basing causal inferences on interaction patterns like that in Figure 3.5. Here the trend lines cross over and the means are significantly different from each other in one direction at the pretest and in the opposite direction at the posttest. The important point is not the crossover per se, since any interaction tells us that trend lines differ. The important point concerns the pattern of switching mean differences, for this tells us that the low-scoring pretest group (the "experimentals") has overtaken the high-scoring control group. None of the other interaction patterns that we have presented thus far does this, nor is it done if the trend lines cross but the two posttest means do not differ.

There are several other reasons why Figure 3.5 is usually more interpretable than other outcomes of the nonequivalent control group design. First, the plausibility of an alternative scaling interpretation is reduced, for no logarithmic or other transform will remove the interaction. Moreover, any reference to a "ceiling" effect mediating the crossover is inappropriate. While this effect might explain why a lower-scoring pretest group comes to score as high as a higher-scoring group, it would not explain how the lower-scoring group then drew ahead. A more convincing scaling artifact would have to be based on the notion that there is some true change in the lower-scoring group but that it is inflated because the scale intervals make change easier for both low- and high-scoring units than for units scoring closer to the grand mean. Note, though, that this entails postulating the exacerbation of a true effect and not the mediation of an artifactual effect.

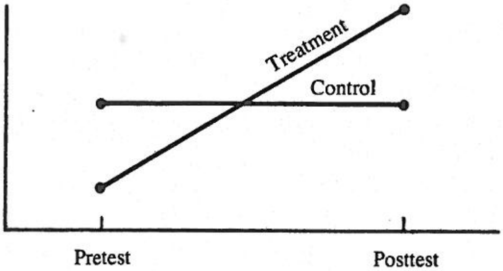


Figure 3.5. Fifth outcome of the no-treatment control group design with pretest and posttest.

Second, the Figure 3.5 outcome renders a regression alternative explanation less likely. When groups are selected on the basis of pretest scores or variables related to pretest scores, there is reason to suspect that a low treatment mean might be regressing to a higher grand mean. However, it is rarely reasonable to expect that this grand mean will be higher than that of the higher-scoring control group. It would have to be, however, if statistical regression were to explain why the experimental group overtakes the control group and significantly differs from it at the posttest.

Third, Cook et al. (1975) have commented on the interpretability of Figure 3.5 when a selection-maturation threat is feared. They reanalyzed some of the Educational Testing Service (ETS) data on the effectiveness of "Sesame Street" and found that children in Winston-Salem who had been encouraged to view the show knew reliably less at the pretest than children who had not been so encouraged. By the posttest, however, the treatment group knew reliably more than the control group, thereby resulting in a data pattern that resembled Figure 3.5. The selection-maturation problem is reduced in this case because so few documented maturation patterns can be described in terms of trends that meet and cross over as opposed to trends that never meet and grow continuously further apart. Of course, complicated forms of selection-maturation cannot be ruled out if the data are as in Figure 3.5. For instance, Cook et al. had to probe the possibility that the encouraged children were both younger and brighter than those in the control group—that they scored lower at the pretest because they were younger and changed more over time because they were brighter. Fortunately, data analysis indicated that the encouraged and nonencouraged groups did not differ in age or pretest measures of cognitive aptitudes that are often considered stable.

Though the outcome in Figure 3.5 is usually interpretable, any attempt to set up a design to achieve it involves considerable risk and should not be undertaken lightly. This is especially true in growth situations where a true treatment effect would have to countervail against a lower expected growth rate in an experimental group. Where this possibility exists, a no-difference finding would not make it clear whether a treatment effect had or had not been obtained for two countervailing forces could have cancelled each other out. Even if there were a difference, this would much more readily take the form of Figure 3.4 than Figure 3.5. Figure 3.4 is much less interpretable than Figure 3.5 on a variety of grounds. It is one thing to comment on the interpretative advantages provided by a cross-over interaction with reliable and switching pretest and posttest differences, and it is quite another to obtain the data pattern.

We have discussed these five outcomes of the no-treatment control group design because the basic design is widely used and its interpretability depends, in part, on the particular outcomes obtained in a research project. The investigator who plans to employ this design would do well to ponder, before data collection, which outcomes are interpretable and which forces may countervail against obtaining such outcomes. In particular, the researcher has to consider the risk of equivocal findings resulting from studies where the no-treatment controls outperform the experimentals at the pretest. To be sure, outcomes of low equivocality *can* result if pretest differences favor the controls (see Figure 3.5), but they are less likely than the more ambiguous Figure 3.4 outcomes.

The Untreated Control Group Design with Proxy Pretest Measures

Sometimes it is not possible to collect pretest measures from respondents either on the same instrument as is used at the posttest or on a parallel form of the instrument. Instead, the researcher has to seek out pretest measures which he or she hopes will correlate with posttest scores within each group, despite being different in form from the posttest scores. The need for different measures is most striking when novel responses are involved—say, in an experiment where the consequences of a preliminary algebra curriculum are to be evaluated, for it would not make sense to give a preexperimental algebra test to children who have never had algebra. Instead, one might give them a general test of mathematical aptitude. Different pretest measures are also needed when evaluating the consequence of an ongoing practice, since in this case it may not be possible to collect any measures other than those found in the archives or those where changes are not likely to be affected by the treatment (e.g., stable characteristics like age, sex, or socioeconomic status).

Such variables function in the design and analysis as proxies for the pretest. The hope is that proxy pretests will be correlated with the posttest within each treatment group, thereby serving two purposes. First, statistical power will be increased if scores on the proxy pretest are related to posttest scores. Second, a preliminary indication will be evident of the way in which selection operates, for the proxy variables may suggest some of the specific initial differences between groups. The basic difficulty with proxy pretests is that they usually correlate less well with posttest scores than do pretests that are collected on the same instrument as the posttest. Consequently, proxies are less adequate than similar pretests for increasing statistical power and for understanding the particular ways in which the posttest scores may be related to initial group differences. (The reader should not infer from this that pretests collected on the same instrument are perfect—far from it.) Pretest-posttest correlations are inevitably imperfect even when the same treatment is used and the measures are corrected for attenuation. For instance, the factorial composition of a measure can change, from one time to another—as when an algebra test gives weight to both algebraic ability and reading skill when used with younger children at a pretest but taps only into algebraic skill when used at the posttest when the children are older. Though the algebra test may have the same apparent form at each testing session, it is not quite the same test each time. The advantage of similar-appearing pretests over proxy pretests are only relative, as we shall see in the next chapter.

The proxy pretest design is diagrammed below. The new subscripts, *A* and *B*, refer to different measures. It can be seen from the figure that the design is identical in all ways to the simple pretest-posttest design with nonequivalent groups except for the *A* and *B* subscripts.

$$\begin{array}{ccc} \hline O_{\mathrm{A1}} & X & O_{\mathrm{B2}} \\ \hdashline O_{\mathrm{A1}} & & O_{\mathrm{B2}} \\ \hline \end{array}$$

Let us consider a concrete example of the design. Imagine a large firm that offers a year-long business leadership evening course to all of its first-year executives. Some 30 of the new junior executives take it, another 50 cannot fit it in, are not interested, or have other reasons. At the end of the year, all of the executives are given a test of Business Leadership Skills (O_{B2}), which hopefully has been developed independently of the curriculum materials. Let us imagine further that statistically significant differences are found on the O_{B2} measure at the posttest and they favor the alumni of the course. Now, skeptics might allege that the alumni would have had better leadership skills even without the course. So, an effort might be made to control for this probability by using personnel selection test scores in each man's file. Imagine, further, that within each group, tests of General Ability, Social Intelligence, and Interpersonal Dominance are found to correlate substantially with the scores on Business Leadership Skills, and so an equally weighted composite of standard scores is formed which correlates .70 within each group. This composite is O_{A1}.

If the course alumni and their controls turn out not to have differed on O_{A1} (or any of its components), this provides reasonable assurance that the groups did not differ on the shared components producing the correlation between O_{A1} and O_{B2}. In such a case, the use of the proxy pretest would increase the interpretability of the quasi-experiment, although the possibility of pretreatment differences on unmeasured components of O_{B2} remains. If, however, a group difference is found on O_{A1}, the usual procedures of matching, covariance, partial r, or multiple regression will almost always *under-adjust*, and the practitioner of these procedures will misleadingly package this under-adjustment as a treatment effect. Indeed, when the correlation between O_{A1} and O_{B2} is .00, no adjustment will take place, and all of the pretest differences will remain in the posttest. When the correlation is substantial and less than 1.00, some of the pretest differences will be removed but not all. Hence, we would erroneously conclude from posttest group differences in leadership skills that the course was effective when, in fact, the posttest difference might be due to selection. It is only when the correlation of O_{A1} and O_{B2} is 1.00 (when corrected for unreliability in O_{B2}) that the adjustment will be adequate and all the group pretest differences will be removed from the posttest. Such a high correlation is unlikely between a posttest and a proxy pretest, and it is this fact that makes the proxy pretest design so difficult to interpret. (An extended discussion of this issue can be found in chapter 7.)

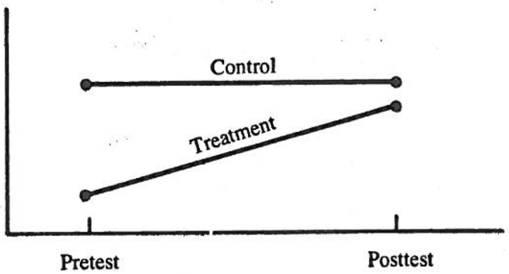
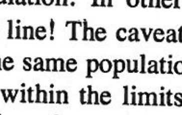


Figure 3.6. Hypothetical pretest and posttest means of a treatment and nonequivalent control group.

The hypothetical example we have just discussed refers to the situation where there is a reliable posttest difference between nonequivalent groups. Sometimes, when there is no such difference, the ex post facto design under discussion has been used to infer that there is no treatment effect. Here, too, one can be in error. Figure 3.6 gives the hypothetical pretest and posttest means for two nonequivalent groups measured on the same scale at each time interval. The group scoring the lowest at the pretest should be considered the treatment group. If we only had posttest information we would conclude from Figure 3.6 that the posttest means did not differ. Now, if we had pretest scores that correlated highly with posttest scores, a statistical adjustment like covariance analysis would reveal a difference between the adjusted posttest scores, with the treatment mean being higher. But if the pretest scores did not correlate highly with the posttest, which is more likely with proxy variables, there will be little adjustment of the posttest means and no reliable differences would be obtained. However, such differences might have been found with the same posttest data if we had better pretest measures. What this implies is that any no-difference conclusion based on the use of proxy pretest variables might be false, reflecting the inadequacies of the design rather than the ineffectiveness of the treatment. (Campbell and Erlebacher, 1970, have discussed this special case of the ex post facto design in greater detail, and the interested reader should consult that reference.)

The Untreated Control Group Design with Separate Pretest and Posttest Samples

The basic pretest-posttest design with nonequivalent groups is sometimes used with a further modification. Instead of the same units being measured at each time interval in the nonequivalent treatment groups, separate samples are used in each group at each time interval. The design is most often needed when the researcher strongly suspects that pretest measurement will affect posttest responses in a way that could easily lead to incorrect inferences about cause. The actual design is diagrammed below, with the vertical line indicating noncomparability across time.



The strongest context where this design can be used is when the pretest and posttest groups in each treatment condition are selected at random so that they are representative of the same population. In other words, the design is strongest where there is no vertical dotted line! The caveat, even when pretest and posttest are each random samples from the same population, is that the pretest and posttest groups are only comparable (1) within the limits of sampling error, so that with smaller and heterogeneous samples of respondents, comparability is problematic; (2) the pretest and posttest groups have to be comparable *after* the process of selecting, so that comparability is not achieved if the selection process is *implemented* in biased fashion or if the persons refusing to be measured are, on the average, different at each time interval; and (3) even if the pretest and posttest groups are comparable, the treatment and control groups are not. Hence most of the threats to internal validity that are relevant to the nonequivalent control group design where the same units are measured at the pretest and posttest still apply when different units are measured at each time.

The difficulties of implementing the design under question can be illustrated by considering an experiment on how lawn mower use is affected by public advertisement campaigns and face-to-face contacts, each of which is designed to inform people about how to use their power lawn mowers more safely (Kerpelman et al., 1978). Part of the total design involved the use of independent pretest and posttest samples in each of two towns, one of which received the ads plus face-to-face contact and the other served as no-treatment controls. Within each town, the selection of respondents was made randomly at both the pretest and posttest, and implementation of the sampling design was left to different research subcontractors at each site. Analysis of the demographic background characteristics of respondents revealed a highly similar pretest-posttest profile at one site, being presumptive evidence that the major assumption of the design was met. However, the pretest and posttest samples had a different profile at another site. The cause of this site difference was not clear, but its implications were. It meant that the greater changes in knowledge of lawn mower safety found at the treatment site when compared to the control site were causally ambiguous: were they due to the treatment or to differences in the composition of the pretest and posttest samples at the one site?

Anyone considering the use of a nonequivalent group design with unique pretest and posttest samples should first assess, in very critical fashion, whether the need for separate pretest and posttest groups is compelling. If *and only if* it is, the researcher should then pay special attention to (1) how the samples are drawn, (2) how the sampling design is implemented and (3) which variables are measured as indirect checks on the comparability of samples. Variables should be chosen because one would not expect them to be affected by the treatment and because such variables would be expected to influence posttest performance. Needless to say, such variables should be reliably measured. However, even if the measures meet these conditions, their use as tests of comparability is still indirect, since it is logically impossible in the posttest-only group to measure the most crucial aspect of comparability—pretest performance. Anyone using this design should check on the background comparability of the pretest and posttest samples as soon as the posttest data are in, for it may well be that the sampling design has been incorrectly implemented and can be reimplemented before the data collection staff has been disbanded. Alternatively, if the data collection phase has come to a complete halt, the only thing that can be done when the pretest and posttest samples appear noncomparable at any one site is to examine the data to find out whether the noncomparability is general or is limited to a specific subsample. In the study of Kerpelman et al. an impressive correspondence was obtained within each treatment group between the pretest and posttest profiles of all the respondents who had not gone to college, but no such correspondence was found with respondents who had attended college. In such a case, it may be reasonable to assume that the sampling requirement of the design has been carried out as planned with *some* of the total set of respondents (i.e., those who had not gone to college).

The design with separate pretest and posttest samples is weak, and should not be attempted unless a clear indication exists that resources permit no stronger design. Even then, the design has to be implemented in such a way that great care is devoted to the sampling and data collection phases, for without independent evidence of the comparability of pretest and posttest samples the design is nearly worthless for purposes of inferring cause. This is because the simplest form of selection will usually provide a plausible alternative interpretation. Moreover, the design will often be low in statistical power given the absence of information about the relationship of the pretest to the posttest within each treatment group.

The Untreated Control Group Design with Pretest Measures at More Than One Time Interval

This design is a variant on the most commonly used untreated control group design, and is diagrammed below.

$$\begin{array}{ccc}
\hline
O_1 & O_2\ X & O_3 \\
\hdashline
O_1 & O_2 & O_3 \\
\hline
\end{array}$$

It can be seen that the only addition to the basic design is an antecedent pretest of the same form used at the posttest.

The advantages of pretests at two (or more) time points are considerable. We saw earlier how a significant threat to internal validity with the untreated control group design is selection-maturation. Adding a pretest permits one to see whether the nonequivalent groups are in fact growing apart at different rates between O_1 and O_2, at a time when the treatment could not have affected the scores. Of course, one has to be careful in using the two pretests to estimate possible differences in growth rates for three reasons. First, the growth rates will be fallibly estimated, given measurement error. Second, scaling artifacts might make the measured growth between O_1 and O_2 unrepresentative of what we would expect between O_2 and O_3. And third—even in the absence of fallible measurement and imperfect scales—we would still have to assume that the rate of growth between O_1 and O_2 in each group would have continued between O_2 and O_3. This is testable for the untreated control group but not for the crucial treatment group. These difficulties notwithstanding, the second pretest can help considerably in assessing the plausibility of selection-maturation.

This is not the only benefit of this design. If the O_2 observations were atypical in either of the groups and we have no O_1 measures, a spurious treatment effect could emerge because of statistical regression. But with a prior pretest one can quickly see whether the O_2 level is inexplicably high or low when compared to O_1. A third benefit of the design cannot perhaps be fully appreciated at this point, for it has more to do with inferential statistical analysis than design. In some instances, it is desirable to be able to estimate the correlation between observations taken from a single group across a known time interval. To compute the correlation between O_2 and O_3 in the treated group gives an unclear estimate of what the correlation would have been had no treatment been given, which is the crucial information needed for the statistical analysis. Consequently, the O_2–O_3

correlation from the untreated group is often used as the best estimate of the corresponding correlation in the treated group. However, a good case can be made that the O_1–O_2 correlation in the treated group will usually be a better estimate than the O_2–O_3 correlation in the untreated group. Since correlations are sensitive to the length of the test-retest interval, it is desirable to have the same interval between O_1 and O_2 as between O_2 and O_3. This is why the design with multiple pretests has been drawn with equal intervals.

Adding a pretest clearly helps the interpretation of possible causal relationships. Why, then, is the design not used more often? One reason may be relative ignorance, but another is surely that the design is often unfeasible. In many situations, one is fortunate to be able to delay the treatment implementation in order to obtain a single pretest let alone two. Sometimes, archives will make the second pretest possible, though with archives the researcher can often extend the pretest ''time series'' for many more than two intervals. Unfortunately, time is not the only feasibility constraint. Some persons responsible for authorizing research expenditures are loath to see money spent for any features of experimental design other than the posttest measurement of persons who have received particular treatments. It is difficult to get such persons to spend money on untreated control group respondents; and we suspect that it will also be even more difficult to get them to spend money on more than one pretest. Nonetheless, where the time frame or the archival record system permits, it is useful to try to persuade authorities to allow two pretests on the same measure taken at different times.

The Nonequivalent Dependent Variables Design

By itself, this is one of the weakest interpretable quasi-experiments. We discuss it for two reasons. First, it can be implemented when resources allow measurement of only a single treatment group. Second, nonequivalent dependent variables can strengthen causal interpretation when *added* to other quasi-experimental designs. The design is diagrammed below in a form which highlights its similarity to the untreated control group design:

O_{1A}	X	O_{2A}	*A* and *B* represent different measures from a single group.
O_{1B}		O_{2B}	

However, the essence of the design is that *a single group* of persons is involved. These are pretested on two scales, one of which is expected to change because of the treatment *(O_A)* and the other is not *(O_B)*. Hence, use of the design is restricted to theoretical contexts where differential change is predicted. If the research is conducted without hypotheses, the design reduces to being a simple one-group pretest-posttest design with multiple dependent variables. Any pattern of differential change might be due to chance, ceiling/basement effects, or to differences in reliability between the measures. These last points are important, for when change and no-change are predicted, it is imperative to demonstrate that the predicted no-change variable has been reliably measured and would probably have registered true effects had they occurred.

Findings using the nonequivalent dependent variables design are only interpretable when the two outcome variables are conceptually similar and each would be affected by most of the plausible alternative interpretations of the obtained effect, other than the treatment. To take an exaggerated example, it would be trivial to demonstrate that a new industrial machine was related to a pretest-posttest difference in productivity (O_A) but not to differences in hair styles (O_B). Rather, one would want to show that the machine caused a difference in, say, the quantity of production during its hours of operation (O_A) but not during the hours when it was down and different machines were used (O_B). The importance of the two related but different dependent variables comes from the fact that alternative interpretations, such as history, would be expected to affect productivity whether the experimental machine was operating or not.

Let us illustrate by an actual example how the credibility of this design depends on initial expectations that both variables *A* and *B* should be affected by any plausible alternative interpretations of an apparent treatment effect. Broadbent and Little (1960) surveyed the literature from laboratory experiments on the effects of noise on industrial productivity. They concluded that "the effect of noise is to increase the frequency of momentary lapses in efficiency rather than to produce decline in rate of work, gross failures of coordination, or similar inefficiency." They set out to test this in an industrial setting where personnel have the job of perforating the edge of film. This was a particularly fortunate work setting since, for payment reasons, measures of the rate of work and the number of broken films were routinely collected and archived. The number of broken films was one of the operational definitions of "momentary lapse," a variable that should be affected by noise. The amount of work performed while machines were working was one measure of "rate of work," a variable that should not be affected by noise. Moreover, it was possible to obtain the relevant archival data both before and after a workroom was experimentally treated and the noise level was reduced. A comparison of before-after changes showed that there were fewer "momentary lapses" after the noise was reduced—and that neither rate of work nor absenteeism was affected when noise was reduced. Without the literature review and the hypotheses it generated, it would probably not have been possible to predict that noise should affect the number of momentary lapses but not the rate of work.

The use of the nonequivalent dependent variables design is not restricted to industrial contexts, of course. It can be used in education where one is attempting to assess, for example, the effectiveness of a new curriculum on geometry. One would expect geometry scores to increase if the curriculum were successful, but one would not expect scores to change on tests measuring the ability to manipulate fractions. If such differential change was obtained, simple maturation or testing effects would be ruled out, since these would presumably increase scores on each test, not just on one of them. In marketing contexts, one would expect a drive to increase sales of a particular washing machine detergent not to affect the sales of cosmetic soaps. However, an increase in the general level of affluence, or a general increase in the concern for cleanliness, would be expected to affect sales of both washing machine detergents and cosmetic soaps.

In our previous discussion of the Broadbent and Little (1960) quasi-experiment, we distorted their design somewhat in order to make the general point.

Consideration of what they actually did and found is useful because it reveals the fundamental weakness of the simple nonequivalent dependent variables design. The investigators found that the rate of work increased in the room where the noise had been experimentally reduced, and they attributed this to historical factors. They were able to do this because their design wisely included a nonequivalent control group work area where the noise had not been experimentally reduced. Thus, the investigators were able to demonstrate that the noisy and less noisy rooms did not differ in the rate of work but did differ in the number of momentary lapses. Without this control for the effects of history, Broadbent and Little would have had some difficulty in explaining why the rate of work increased and lapses decreased in the less noisy work areas. This was, after all, the very outcome that they did not want because it would not have supported the propositions about the *differential* effects of noise that were derived from laboratory experiments. The nonequivalent dependent variables design is entirely dependent on contrasting patterns of change and no-change, and it cannot handle the pattern of general change that Broadbent and Little obtained in their reduced-noise condition alone. Thus, nonequivalent dependent variables are probably better used as part of a larger design rather than a complete design in itself.

As we have presented the design thus far, it has only two variables and two waves of measurement. We would like now to set the nonequivalent dependent variables design into a broader "pattern-matching" context. The design is obviously strongest where differentiated patterns of change are predicted that allow many alternative interpretations to be ruled out. The probability of ruling out threats depends in part on the specificity of the predicted data pattern so that interpretability increases (1) with the number of dependent variables for which predictions are made—the two-variable/two-wave case is merely the simplest case of the more general design—and (2) with the specificity of numerical or sign predictions made.

The most important point to be noted is that the *prospective* consideration of plausible alternative interpretations will sometimes incline the researcher to predict a data pattern involving multiple variables against which the obtained data can be matched to assess how well the data corroborate the expected pattern and how well alternative interpretations are ruled out. (Since some effects would be expected by chance in the multivariable-multiwave case, it is advisable, where possible, to have two measures of each relationship.) In any event, nonequivalent dependent variables designs based on multiple variables and multiple measurement waves will often permit ruling out all plausible threats to internal validity. But if these threats are not made explicit before data collection begins, it is unlikely that all the variables will be measured that are required for matching obtained data with a pattern of relationships that logically rules out threats to valid causal inference.

The Removed-Treatment Design with Pretest and Posttest

It is sometimes not feasible to obtain even a nonequivalent control group. In such a situation one is forced to create conditions that closely approximate meeting the conceptual requirements of a no-treatment control group. The design that we outline below does this in many instances.

$$O_1 \quad X \quad O_2 \qquad O_3 \quad \overline{X} \quad O_4$$

In essence, the design calls for a simple one-group pretest-posttest design (see from O_1 to O_2). But a third wave of data collection is added (see O_3), after which the treatment is removed from the treatment group ($\overline{X}$ symbolizes this), and a final measure is taken after the treatment has been removed (O_4). Thus, from O_1 to O_2 is the experimental sequence, as it were, while the sequence from O_3 to O_4 serves as a no-treatment control for the sequence from O_1 to O_2. Note that the same group of units is involved throughout the whole design.

In this design, we would expect an effective treatment to cause a difference between O_1 and O_2 that is opposite in direction to the difference between O_3 and O_4. However, since it is possible that the initial effects of the treatment might increase or even dissipate between O_2 and O_3, it is important to add that there has to be a noticeable discontinuity after $\overline{X}$ (as in Figure 3.7). If there is not, and if there is a smooth trend from O_2 to O_4, then any difference between O_3 and O_4 that was different from the experimental $O_1 - O_2$ difference might be due to the treatment having no long-term effect rather than to the treatment effect dissipating because it was removed.

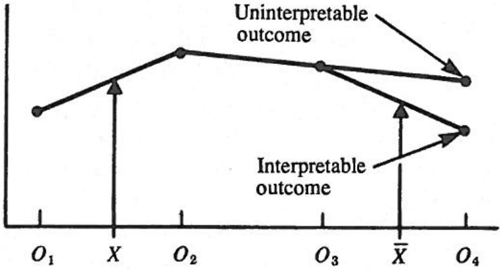


Figure 3.7. Generally interpretable outcome of the removed-treatment design.

Only four problems seem to arise with the interpretable outcome in Figure 3.7. First, it might be difficult to obtain the pattern of statistical effects necessary for statistical conclusion validity, since this would require both that $(O_1 - O_2) \neq (O_3 - O_4)$ and that $(O_2 - O_3) \neq (O_3 - O_4)$.

A second problem concerns construct validity of the cause. Since many treatments are ameliorative in nature, removing them might not only be hard to defend ethically but may also arouse frustration in respondents that should be correlated with indexes of aggression, satisfaction, and perhaps performance. Such considerations indicate that sometimes it will not be possible to implement the design, especially if a deliberate choice has to be made by research personnel to remove an ameliorative treatment.

Third, there are many instances where respondents voluntarily decide to discontinue their exposure to a treatment for reasons that have no obvious relationship to the fact that social research is taking place. The design is most likely to be used, therefore, when subjects self-select themselves out of the treatment group. But very special care has to be taken when this happens. To illustrate the need for care, imagine someone who becomes a foreman (X), develops promanagerial attitudes between O_1 and O_2, dislikes his new contact with managers, and becomes less promanagerial by O_3. This person would be a likely candidate for resigning from his position or for being relieved of it ($\bar{X}$). Any continuation of his less promanagerial attitudes after changing from a foreman back to a factory worker would result in an $O_3 - O_4$ difference that differed from the $O_1 - O_2$ difference. Thus, the researcher has to decide whether the $O_3 - O_4$ difference reflects spontaneous maturation or the change of jobs. The maturation explanation would be more likely if the $O_3 - O_4$ difference were similar to the $O_2 - O_3$ difference (see the outcome marked "uninterpretable" in Figure 3.7) but would be less likely if the $O_3 - O_4$ difference were greater than the $O_2 - O_3$ difference (see the "interpretable" outcome in Figure 3.7). A rise in promanagerial attitudes between O_1 and O_2 and a decline in promanagerial attitudes between O_3 and O_4 that was greater than the $O_2 - O_3$ decline would strongly suggest that entering a new role causes one to adopt the attitudes appropriate to that role.

Fourth, it is advantageous with this design that the observations be made at equal time intervals. This permits a control for any spontaneous linear changes that take place over a given time period. A simple comparison of the differences between $O_2 - O_3$ and $O_3 - O_4$ would be meaningless if the $O_3 - O_4$ time interval were longer than the $O_2 - O_3$ interval. This is because a constant rate of decay would reveal larger $O_3 - O_4$ differences than $O_2 - O_3$ differences. As we shall later see with time-series designs, it is often possible to estimate rates of change per time interval so that the equal spacing of observations loses many of its advantages. But a sensitive estimate of the spontaneous rate of change is not possible with the design under consideration since two of the time differences might be affected either by the treatment or its removal. This increases the reliance on equal intervals, though they control only for simple linear patterns of spontaneous change.

Lieberman (1956) used a simpler version of the removed-treatment design in his examination of the attitude change that follows role change. He obtained samples of foremen who lost their new positions and reverted to being workers again. Lieberman had three waves of measurement: before becoming a foreman, after becoming a foreman, and after reverting back to worker status. The part of his design under discussion differed, therefore, from the one we have outlined since only one measurement was made between the treatment and the removal of the treatment. Hence, we are unable to attribute any differences between his $O_1 - O_2$ and his $O_2 - O_3$ measures as due to a shift in the attitude of former foremen who became workers again or to the fact that foremen whose attitudes were becoming less managerial were selected for demotion. In addition, the statistical analysis of the three-wave design could involve contrasting the $O_1 - O_2$ difference with the

$O_2 - O_3$ difference, a procedure that uses the O_2 observations twice. If, through sampling error, the O_2 mean were raised, this would necessarily be reflected in an $O_1 - O_2$ difference of a different algebraic sign (and hence implied causal direction) from the $O_2 - O_3$ difference. Such a difference would occur even if nothing had happened as a result of the treatment! Having two observations between the treatment and removal of the treatment rules out these possibilities in the design we have advocated.

The Repeated-Treatment Design

When the investigator has access to only a single research population it will sometimes be possible to introduce the treatment, fade it out, and then reintroduce it at a later date. Obviously, this design is most viable in contexts where the initial effects of the treatment are transient or do not prevent the treatment from having an even stronger effect when it is reintroduced. The design is diagrammed below.

$$O_1 \quad X \quad O_2 \quad \overline{X} \quad O_3 \quad X \quad O_4$$

The most interpretable outcome of this design is when O_1 differs from O_2, O_3 differs from O_4, and the $O_3 - O_4$ difference is in the same direction as the $O_1 - O_2$ difference. The design is of the general type associated with Skinnerians, and the basic logic behind it was used in the original Hawthorne studies (Roethlisberger and Dickson, 1939). It may be remembered that in some of those studies women factory workers were separated from their larger work groups and were given different kinds of rest periods at different times so that the experimenters could investigate the effects of rest on productivity. In some cases, the same rest period was introduced at two different times; if we were to regard only these repeated rest periods, we would have the basic design under discussion here.

One threat to internal validity comes from the possibility of cyclical maturation—that is, productivity is being affected by regularly occurring systematic factors. For example, if O_2 and O_4 were recorded on Tuesday morning and O_1 and O_4 on Friday afternoon, any differences in productivity might be related to differences in daily performance rather than to a treatment. It would be preferable, therefore, if such cyclical factors could be ruled out. A second threat to internal validity can arise if there is resentment when the treatment is removed between O_2 and O_3. If this were to happen, O_3 would be decreased and an $O_3 - O_4$ difference might be erroneously attributed to a replication of the treatment's effect when it was in fact due to removing a source of frustration by reinstating the treatment.

When the basic design is used as it was in the Hawthorne studies, it is particularly vulnerable on grounds of external and statistical conclusion validity. For example, many of the performance graphs in Roethlisberger and Dickson (1939) are of individual women workers, and in the Relay Assembly Row Experiment there was a grand total of only six women! Moreover, there appears to be considerable variability in how the women reacted to treatments (particularly the Mica Splitting Room Experiment). We cannot be sure to what extent results would be statistically significant if the analyses were based on summing across all the women. (We cannot help but note in passing how closely the Hawthorne studies parallel the design of Skinnerian experiments. There is the same preference for few subjects and repeated reintroduction of the treatment, and there is the same disdain for statistical tests.) Of course, the repeated treatment design does not *require* that there be a small population or an absence of statistical tests. These are merely correlates of the use of this design in the past. We would strongly urge the use of larger samples and statistical tests.

Construct validity is a major threat because respondents may well notice the introduction, removal, and reintroduction of the treatment with the consequence that they can guess and respond to a hypothesis about the purpose of the study. It is worth noting that this can occur even when there is none of the obtrusive observation or special group status that was involved in the original Hawthorne experiments. When respondents are reacting to their special status in an experiment or to a hypothesis they might have guessed, we cannot be sure how the treatment should be labeled. This design is better, therefore, when there are unobtrusive treatments and a long delay between the treatment and its reintroduction. It is also desirable that there be no confounding of cycles and reintroductions of the treatment. Hence, the design is best of all when the reintroductions are frequent and randomly distributed across time blocks. (This last point will be discussed later when we deal with randomized experiments, particularly the Equivalent Time Samples Design.)

The Reversed-Treatment Nonequivalent Control Group Design with Pretest and Posttest

This design can be diagrammed

$$\begin{array}{ccc} \hline O_1 & X+ & O_2 \\ \hdashline O_1 & X- & O_2 \\ \hline \end{array}$$

where $X+$ represents a treatment that is expected to influence an effect in one direction and $X-$ represents the conceptually opposite treatment that would be expected to reverse the pattern of findings in the $X+$ group.

Morse and Reimer (1956) probably used this design to investigate how decision-making procedures that were either "democratic" (i.e., participative) or "hierarchically controlled" affected productivity and job satisfaction. The hypothesis was that the "democratic" procedure would increase productivity and satisfaction but the hierarchically controlled procedure would decrease them. To test this, Morse and Reimer developed a design that involved the use of four divisions in an organization, two of which were assigned to each experimental condition. While it is not clear from the report whether the assignment of treatments to the four divisions was done on a random basis, the small number of experimental units makes it difficult to believe that the treatment and control groups would have been comparable at the pretest even if random assignment had taken place. Indeed, at one point in their report Morse and Reimer commented that the two experimental groups tended to differ in pretest satisfaction, and tables in the report indicate a possible group pretest difference in respondents' perception of the locus of decision making. Thus, for our present illustrative purposes, we shall consider the Morse and Reimer study as a quasi-experiment, the results of which indicated that satisfaction increased between pretest and posttest in the "democratic" decision-making group and decreased in the "hierarchically controlled" group. It is precisely such a pattern of change in opposite directions that is indicative of a treatment effect. But, as always, alternative interpretations have to be considered.

Respondents in the Morse and Reimer study probably did not select themselves into work divisions, and the work divisions probably did not select themselves into being in one experimental group or the other. If these suppositions are correct, selection-maturation would probably not be a major threat to the internal validity of the interpretation of findings. We would have no reason to suspect that the two groups were spontaneously maturing in different directions. What makes a selection-maturation interaction less likely in this design than in many others is that the only interaction pattern that can alternatively explain the findings is the rare one where maturation operates *in different directions* in each group rather than the more typical maturational pattern where change occurs *at different rates in the same direction* in each group. The reversed-treatment control group design is often stronger with respect to internal validity than its simpler alternatives because selection-maturation cannot usually explain changes in opposite directions.

The reversed-treatment design with nonequivalent groups is stronger than the no-treatment control group design with respect to construct validity. This is because the theoretical causal variable has to be rigorously specified if a test is made that depends on one version of the cause affecting one group one way and another group the other way. Moreover, many of the irrelevancies associated with one treatment will be different from those associated with the reversed treatment. To understand these points better, what would have happened if Morse and Reimer's design had involved only a "democratic" decision-making group and a no-treatment control group. A steeper pretest-posttest satisfaction slope in the "democratic" group could have been attributed to the new locus of decision making or to a Hawthorne effect. But the plausibility of a Hawthorne effect is lessened when we note the pretest-posttest decrease in satisfaction in the "hierarchically controlled" group. This is because awareness of being in a research study is typically considered to elicit socially desirable responses (higher productivity or greater satisfaction, rather than less desirable responses such as decreased satisfaction). It is the high construct validity of the cause which makes the reversed-treatment design potentially more appropriate for theory-testing research than the no-treatment control group design.

The last statement should not be taken to mean that the reversed-treatment design is flawless with regard to specifying the causal construct. Morse and Reimer found that productivity was greater at the posttest than the pretest in both the "democratic" and "hierarchical" decision-making groups. If we accept for the moment that these data indicate an increase in productivity (which is not clear in the absence of a no-treatment control group), the possibility arises that a Hawthorne effect may have caused the productivity change. This is because the productivity measure, unlike satisfaction, did not show the expected changes in opposite directions. The moral of this is that the potentially high construct validity of the reversed-treatment design depends on the research revealing changes in opposite directions. When change is in the same direction in both groups, we are left in the same position as with the relatively uninterpretable one-group pretest-posttest design, i.e., there is change, but we do not know whether the same type of change would have occurred even in the absence of a planned treatment. To be maximally interpretable, the reversed-treatment design needs (1) a placebo control group which receives a treatment that is not expected to influence productivity or satisfaction except through a Hawthorne effect and (2) a no-treatment control group which would provide a no-cause baseline.

We should also not forget that, in many organizational contexts, ethical and practical drawbacks prevent the implementation of some kinds of reversed treatments. It is not, after all, useful or humane to introduce treatments that will harm people or productivity. Indeed, the understandable preference for ameliorative and prosocial treatments makes reversing such treatments problematic. Given this, it will often not be possible to reverse treatments deliberately, and we should instead expect to use the design in question mostly when reversals are unplanned. Since reversed treatments will often be unpopular, it may be difficult to unconfound effects attributable to the planned aspects of the reversal from effects attributable to any unplanned affective consequences that follow from receiving an unpopular treatment.

A problem of statistical conclusion validity is associated with the reversed-treatment design. Imagine that a simple analysis of variance resulted in a statistically significant interaction of experimental groups and time of testing, with the posttest differences being greater than the pretest ones. We would not know from this whether the effect was the result of (a) sampling error, (b) only one of the treatments causing its expected effect or (c) both treatments causing effects in opposite directions. Significant pretest-posttest differences *within each group* would decrease the chances of (a) above but would not discriminate between (b) and (c) in many cases. This is because at times it would be reasonable to assume that at least one of the pretest-posttest differences was due to group-specific maturation and not to the treatment or its reversal. Interpreting the direction of change would be even more problematic if only one of the two differences were significant. While it would seem at first glance that the significant difference represents directional change, this need not be so, for the single difference might be due to maturation. To remove ambiguity about whether there is treatment-correlated change in one but not the other treatment group we need a no-treatment control group. Indeed, researchers who are particularly interested in the direction of change should think twice before using a reversed-treatment design in the absence of a no-treatment control group. However, with a mixed design of a treatment/reversed-treatment, and no-treatment group, researchers would be in a very strong position. It would be even stronger if they could add a fourth group of placebo controls.

Cohort Designs in Formal and Informal Institutions with Cyclical Turnover

Many formal institutions are characterized by regular turnover as one group of persons graduates to another level of the institution. Schools provide an obvious example of this as children move from grade to grade. So, too, do many businesses as one group of trainees follows another. Such systematic turnover patterns are not confined to institutional settings in the normal use of that expression. For instance, children follow each other within families, sharing much the same home environment and differing genetically from each other only in random fashion. We shall use the term "cohort" to denote groups of respondents who follow each other through formal institutions or informal institutions like the family. Such cohorts are useful for experimental purposes because (1) some cohorts receive a particular treatment while preceding or following cohorts that do not, (2) it is often reasonable to assume that a cohort differs in only minor ways from its contiguous cohorts, and (3) it is often possible to use archival records for comparing cohorts who have received a treatment with cohorts who were in the same institutions before the treatment began or after it was discontinued.

The crucial feature that makes cohort designs particularly useful for drawing causal inferences is that a "quasi-comparability" can often be assumed between the cohorts that do and do not receive a treatment—between, say, the fourth grade class one year and the fourth grade class the next year. How reasonable it is to assume quasi-comparability in a particular research setting depends on how similar the cohort groups are on the average in background characteristics, including organizational history. The degree of achieved comparability will never be as high with cohorts as with random assignment. Indeed, the virtue of cohort designs is only relative, based on the fact that cohort groups are likely to be more similar to each other than are treatment groups which do not share the same home or work environment. Since cohort designs do not automatically rule out selection, they gain additional strength if the data analysis shows that cohort groups with and without the treatment do not systematically differ on reliably measured third variables that are believed to be possible mediators of a treatment effect. While the strong measurement of such variables obviously cannot rule out selection threats associated with unmeasured variables, it can at least rule out some particular types of selection.

Our discussion of cohort designs will be divided into two sections. The first deals with designs that recognize differences in the degree to which various respondents experience a treatment. The second deals with designs that can be used if the treatment is more homogeneous and respondents have little opportunity to regulate how much of the treatment they receive. The latter would be the case, for example, when evaluating the effects of a curriculum in schools or training centers. We shall not deal with the growing use of cohort designs to try to unconfound the effects of age (i.e., growing older), birth cohort (i.e., being born within a given time span), and period of history (i.e., the events occurring between any two time intervals). Using the data from sample surveys of cohorts to unconfound these effects is a task that demographers and developmentalists in particular have set themselves, and the inferential pitfalls that make their task difficult are outlined in Glenn (1977).

The Cohort Design in Which Treatment Partitioning Is Possible

Minton (1975) wanted to examine how the first season of "Sesame Street" would affect the Metropolitan Readiness Test scores of a socially heterogeneous sample of kindergarten children. She located a kindergarten where the test was used at the end of the child's first year. However, she had no data from a control group of children who did not watch the show during the season. Fortunately, though, she had access to the Metropolitan scores of the children's older siblings from earlier years when they had been the same age as the "Sesame Street" viewers but the show had not been on the air. She was able, therefore, to compare the postkindergarten knowledge level of children who were "Sesame Street" viewers with the knowledge level of their siblings when they terminated kindergarten in the years before "Sesame Street." The essential design can be diagrammed below, with the wavy line indicating a restricted degree of selection nonequivalence. The purpose of the design is to test whether the two observed groups differ.

$$\begin{array}{ccc}\hline O_1 & & \\ \sim\sim\sim & \sim\sim\sim & \sim\sim\sim \\ & X & O_1 \\ \hline\end{array}$$

The design as it stands is not strong. First of all, there might be a selection problem since the older siblings are more likely to be first-borns and any differences between O_1 and O_2 might be due to comparing groups with different percentages of first-borns. It would be desirable, therefore, if effects of ordinal birth condition could be either assessed, reduced, or eliminated. One way of reducing the threat would be by analyzing the data separately for second-born older children and their third-born siblings, for third-born children and their fourth-born siblings, and so forth on the assumption that ordinal position makes more of a difference with first- and second-borns or penultimate and last-borns than with children in the middle (Zajonc and Markus, 1975).

The design is also weak with respect to history, for the older and younger siblings in the design could have experienced events other than "Sesame Street" which affected knowledge levels in one cohort more than the other. An indirect way of partially examining this threat would be to break down the cohorts into those whose kindergarten experience was separated by one, two, three, or more years to see if the greater learning of the younger group held over these particular sets of unique historical events. This procedure would be less than optimal, of course, because there would be no control for the historical events other than "Sesame Street" that took place in the same year the show was introduced.

Another control for history would have been to split the children into viewers and nonviewers or, if this were not possible, into heavy and light viewers. If "Sesame Street" were effective, we would then expect a statistical interaction—with larger knowledge differences between the heavy and light viewers than between their respective siblings. The reason for this is that, in the absence of a treatment effect, there would be no reason to assume that the difference in knowledge between heavy and light viewers of "Sesame Street" should vary in any way from the difference between their siblings who would presumably have become heavy and light viewers respectively if "Sesame Street" had been available to them. In particular, we would expect the heavy and light viewers to experience the same general history. Partitioning respondents into treatment groups based on the extent of their experience with the treatment greatly strengthens the internal validity of this particular cohorts design. It is difficult to come up with plausible alternative interpretations when the data look like Figure 3.8.

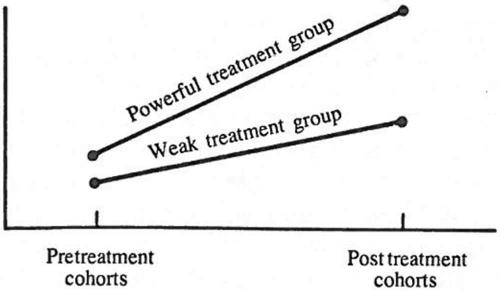
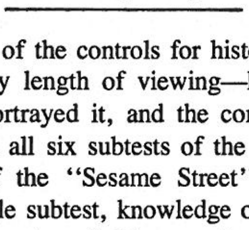


Figure 3.8. Interpretable outcome of a posttest-only cohort design with two treatment levels.

Partitioning has a further advantage for internal validity. If the conditions of testing differ between the earlier and later cohorts, then testing alone might cause higher scores in the later experimental group than the earlier control one. Partitioning respondents by the length of exposure to the treatment rules out a simple testing threat, for there is no reason why testing should have a greater effect in the longer exposure treatment group when compared to the shorter exposure group. For a number of reasons, then, we advocate partitioning respondents and implementing the modified design below, where X_1 and X_2 indicate quantitative differences in the extent of treatment implementation.



Though Minton had none of the controls for history and testing listed above—since she did not partition by length of viewing—her design was slightly more complicated than we have portrayed it, and the complications increase interpretability. She collected data on all six subtests of the Metropolitan Readiness Test, and found that the mean of the "Sesame Street" cohorts differed from their siblings' mean on only a single subtest, knowledge of letters. Since content analyses of the first year's programming of "Sesame Street" (Ball and Bogatz, 1970) have shown that more time was spent teaching letters than anything else, it might be assumed that a test of letter skills would be more sensitive to program effects than other tests. In a sense, then, the design that Minton used involved *both* cohort groups *and* nonequivalent dependent variables. In her case, however, the different status of each dependent variable was ascertained after the show's first season was over and not before. It is clearly more problematic to infer the different conceptual status of dependent variables after the fact than before it!

Instead of using siblings as cohorts, Ball and Bogatz (1970) in their evaluation of "Sesame Street" used older children from the local neighborhood. Their experimental design involved taking a sample of children and testing them before "Sesame Street" went on the air and six months later. Some of the children's ages ranged between 47 and 52 months at the pretest and between 53 and 58 months at the posttest. These were called the posttest cohort. Other children's ages ranged from 53-58 months at the pretest and 59-65 months at the posttest. These were called the pretest cohort. By considering just the posttest scores of the posttest cohort (then aged 53-58 months) and just the pretest scores of the pretest cohort (then also aged 53-58 months) Ball and Bogatz created a design where it was hoped that an effect of "Sesame Street" would be inferred if the posttest cohort of 53-58 months was more knowledgeable than the pretest cohort of 53-58 months.

When this design is used, maturation cannot easily explain any differences between the means of the pretest and the posttest cohorts. This is because they are of equivalent age and so are presumably at comparable maturational stages. A selection effect is also not likely, provided that the analysis included data from *all* the available children in each age cohort. As a check on the comparability of pretest and posttest cohorts, background characteristics can be measured. Indeed, Ball and Bogatz measured a variety of demographic background characteristics and showed that the cohorts did not differ from each other on any of the variables. As we have outlined it thus far, the design differs from Minton's only in using neighborhood children of different birthdates as cohorts instead of siblings of different birthdates. Consequently, we might anticipate that the design is vulnerable to a history interpretation. It might be, for example, that something which affects learning occurred at about the same time that "Sesame Street" was introduced or older pretest cohorts may have experienced unique events before the younger cohorts were born that affected the older cohorts' knowledge. Or the older children may have been at particularly sensitive maturational stages when they learned information also available to the younger cohorts but less meaningful for them. Also, the design as portrayed here and as implemented by Ball and Bogatz has a unique testing problem, since the scores of the pretest cohort came from a first pretest measurement wave while the scores of the posttest cohort were from a second wave of measurement (i.e., from the posttest after having been pretested six months earlier). Thus, it would not be clear whether any obtained differences between cohorts were due to the treatment or to differences in the frequency of measurement.

The problems of history and testing can be eliminated by following Ball and Bogatz's extension to their basic design. They used measures of the frequency of viewing "Sesame Street" to partition each cohort into four separate groups that differed in the level of reported viewing. The results are displayed in Figure 3.9, and an analysis of variance showed that the differences in knowledge between the various viewing groups were greater among the posttest cohort than the pretest one. Since the cohorts were of the same mean age, of comparable social background within the different viewing groups, and experienced the same history and testing sequences (all posttest cohorts were pretested), an interaction outcome like the one that Ball and Bogatz obtained can account for all the threats to internal validity that we have discussed thus far.

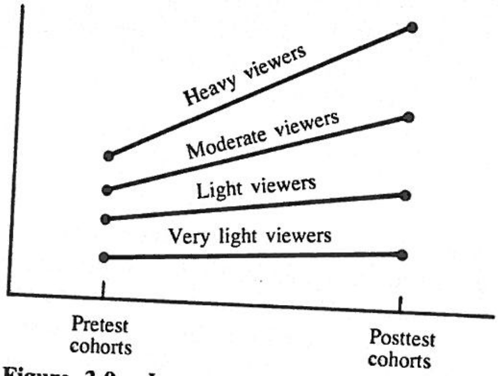


Figure 3.9. Interpretable outcome of a selection cohorts design with pretest and posttest cohorts.

The selection cohorts designs we have outlined are very useful (1) when age or experience can alternatively account for results in a pretest-posttest design or (2) when no pretest measures of experimentals are available. The data from such designs are especially interpretable in causal terms if there are different levels of a treatment and the data analysis reveals that these statistically interact with the cohort groups.

The Cohort Design in Which No Treatment Partitioning Is Possible

We have seen previously how history, selection, and some forms of testing are especially likely to be threats to internal validity in the simplest cohort designs. In much research these threats can be dealt with by separating out individuals and giving some a treatment while withholding it from others. But such separation is rarely possible in research with intact organizations. Here, a treatment has to be made available to all, and hard-headed causal inferences have to depend on making use of the fact that all the members of some organization received a treatment one year while none of their cohorts received it in previous years. The problem then becomes: How does one explicate and control for all the threats that operate when comparing cohorts?

We have previously noted Saretsky's (1972) claim that the no-treatment control group children in the Performance Contracting Experiment performed better than would have been expected on the basis of previous years. What Saretsky was trying to demonstrate was that something associated with the child or teacher knowing that he or she was in a control group enhanced learning. It is not clear how Saretsky tested this causal hypothesis. Let us assume for pedagogic purposes that he compared the average grade equivalent gain in control group classes with the average gain from classes of the same grade taught in previous years when there was no awareness of being in a study. Thus, the design would be of the form below where O_1 and O_2 represent scores for the earlier cohort and O_3 and O_4 represent scores for the other cohort at a later date. Obviously, this design—which we might call "the institutional cycles design"—can be extended back over time to include multiple "control" cohorts rather than the single one pictured here. Indeed, it seems that Saretsky reported data for two preexperimental years.

$$\begin{array}{ll} O_1 \quad O_2 & \\ \sim\sim\sim\sim\sim\sim\sim\sim & \sim\sim\sim\sim\sim\sim \\ & O_3\ X\ O_4 \end{array}$$

How would one control for *selection* deflating the mean difference for the earlier control cohort or inflating the difference for the later experimental cohort? One way of approximating a test would be to examine the background characteristics of students in each cohort to see if they are comparable on the available measures. If they are, and if the measures are plausibly related to the outcome variables, then a selection threat is all the less likely. Another way would be to draw all of the siblings out of the total sample and restrict the data analyses to them. Another would be to test whether the teachers were the same in each cohort, restricting the analyses to teachers who were equally represented in each cohort.

How could one attempt to rule out *testing*? The particular testing threat involved with the design under discussion is that at least one of the four tests was atypically administered or scored. Sometimes, there are data available about this, but often the assumption is made that school testing is a routine affair with which students and teachers are familiar so that testing should not differ by years. This argument is, of course, based on plausibility alone; and it would be useful if additional data on the issue could be made available.

History is the most salient and troublesome threat in the institutional cycle design when evaluating the effects of an innovation, for history might either have decreased the performance of the earlier control cohorts or increased the performance of the later treatment cohorts. Here, it can help to have a *series* of control cohort years, for if the pretreatment years are comparable to each other and only the posttreatment year differs from any other, then it is not plausible to claim that performance in the pretreatment cohorts was atypically low. The search for an alternative interpretation based on history can thus be limited to events that occurred in the treatment year and not earlier. Sometimes, no particular history threats will come to mind. But, in general, we would be in a stronger position to deal with history during the treatment year if a nonequivalent, no-treatment control group could be found and measured during the critical year. Or, failing this, the design could be greatly strengthened if nonequivalent dependent variables were specified, some of which should be affected by history while others should not be.

The plausibility of a history threat can be examined if the research question is slightly different and if, instead of wanting to know the effects of a new practice (e.g., participating in a performance contracting experiment as control respondents), one wants to know the effects of a long-established practice (e.g., what is the effect of asking second graders to do one-half hour of homework each school night). If one had access to past school records, or if one had a two-year period in which to do the research, then a design like the one sketched below might be possible. The first cohort group consists of children measured at the end of the second grade; the second consists of children from the next year's cohort when they reached the second grade and were measured at the beginning and end of that school year; and the third group consists of children who enter second grade the year after the second cohort group. The spacing of the observations is meant to represent the fact that O_1 and O_2 are not simultaneously observed, nor are O_3 and O_4. Interpretation is easier, of course, if measurement is simultaneous. This particular design is discussed in greater detail by Campbell and Stanley (1963), who call it the "Recurrent Institutional Cycle Design."

$$XO_1$$

$$\qquad O_2 \quad X \quad O_3$$

$$\qquad\qquad\qquad O_4$$

An approximate control for history is provided if, in addition to O_3 being greater than O_2, O_1 surpasses O_2, and O_3 surpasses O_4. These last comparisons suggest that a treatment has been effective at two different times so that any historical force would have had to operate twice if it is to explain both $O_1 > O_2$ and $O_3 > O_4$. Selection is also ruled out in this version of.a cohort design since, quite apart from the similarity (but nonequivalence) of cohorts, the same persons are involved in the $O_2 - O_3$ comparisons. The remaining problem that affects cohort designs—testing—is not ruled out since all the comparisons involve contrasting a first testing (O_2 or O_4) with a second testing (O_1 and O_3). This is why Campbell and Stanley recommended extending the design further by splitting the middle group that is both pretested and posttested into random halves, one of which receives a pretest, treatment, and posttest sequence while the other receives a treatment and posttest but no pretest. Any differences between these two groups at the posttest would presumably be due to repeated measurement; the failure to obtain differences or even strong but unreliable trends suggesting differences would indicate that repeated measurement is not a problem.

Though the three-group design that we have outlined is practical for use in institutional settings where everyone has to receive a treatment, it has one major drawback over and above testing (which in some concrete situations will seem an implausible threat). The drawback is that interpretability depends on a complex pattern of outcomes in which three contrasts are all statistically reliable in similar ways. Since two of these contrasts involve O_2, a chance elevation of O_2 would have disastrous implications. This implies that the design should only be used with reliable measures and large samples.

A Posttest-Only Design with Predicted Higher-Order Interactions

In some circumstances no pretest information is available and it is desirable to test a causal relationship. Unfortunately, there are few quasi-experimental designs which permit this, and pretests are an absolute necessity for most designs unless some form of cohort or interaction strategy is used. (It is a different matter with randomized experiments. Pretests can be dispensed with since randomization ensures the probabilistic equivalence of the different treatment groups at the pretest. However, it is advisable to collect pretest data nonetheless, for without it difficulties may arise in designing a ''fallback'' quasi-experiment that is interpretable. The need for such a fallback is acute when the comparability of treatment groups is not maintained over the course of an experiment, as would be the case if there were higher attrition from the experiment in some treatments than in others.)

Let us illustrate how, in the absence of pretest information, interaction predictions can be used with intact groups for providing relatively strong inferences about cause. Nisbett and Kanouse (1969) were interested in testing the idea that overweight persons lack the ability to discriminate the internal body cues that indicate hunger. Hence, the authors hypothesized that among overweight persons there would be no relationship between the time of last eating and the amount of grocery purchases, but that there would be such a relationship among persons of normal weight who do pay attention to internal cues which indicate how hungry they are. To test this, Nisbett and Kanouse asked customers who entered a supermarket when they had last eaten, and they also observed the customers' weights and the size of their grocery bills. The data analysis revealed that body weight (coded as overweight or normal) and the reported number of hours since last eating (a variable with six levels) statistically interacted to determine the size of the grocery bill. As predicted, there was positive correlation between purchases and time since last eating among normals but, unlike the prediction, there was a *negative* correlation of these variables among the overweight.

A major difficulty with this design is selection. Assume for the moment that persons of normal weight who wait the longest time between meals are more likely to have jobs. (After all, it is more difficult for social and practical reasons to eat at work than at home.) If this were the case, normal persons who had gone longer without eating might well be more affluent and have more money to spend on food. This would explain the pattern of discrimination that Nisbett and Kanouse predicted for normals. But it would not explain the pattern among the overweight. However, if we further assume that the overweight persons who go longer without eating may do so because they are less affluent, then they should have less to spend on groceries than their overweight counterparts who have recently eaten. This would explain the negative relationship among the overweight. Alternatively, the overweight persons who have not eaten for a comparatively long time might be abstaining in order to diet, and this might also be related to lower grocery purchases. The point is that various selection mechanisms *could* explain the interaction of body weight and time since last eating, though a different selection mechanism has to be invoked for each weight group.

A second potential problem with this design relates to the specificity of the predicted outcomes and the difficulty of obtaining such specific patterns of data.

Nisbett and Kanouse predicted that there would be no relationship between purchases and time since last eating among the overweight who are relatively insensitive to internal hunger cues. But they unexpectedly obtained a negative relationship among the overweight. The authors needed, therefore, to explain this unexpected pattern. Since they had creatively collected estimates of intended purchases from shoppers as they entered the supermarket, they were able to show that the overweight persons who had gone longer since last eating both intended to buy less and actually did buy less. This was interpreted as demonstrating that the purchasing behavior of overweight persons was probably determined by their expectations about purchasing rather than by their internal hunger cues. The corollary of this is that normal persons' behavior should be determined more by their internal hunger cues than by their expected purchasing. However, the evidence for this was ambiguous. While the difference between what normals expected to buy and what they actually bought increased over five levels of time-since-last-food, it deviated markedly from this pattern among normals who had not eaten for more than five and one-fourth hours. Persons in this last group actually bought *less* than they intended, even though the theory predicted that they should have been more sensitive to their internal hunger cues than others and that they should have been the most prone of all to buy on impulse, thereby buying more than they intended.

The moral is clear and is illustrated in our earlier discussion of case studies with multiple "outcome" variables and predictions about which areas should be affected in which ways if a certain variable were the "cause." The moral is that causal interpretation tends to be facilitated as the predicted interaction between nonequivalent groups grows more complex. But the chance of obtaining so many data points in the predicted order decreases with the number of data points predicted. There are many reasons for this, including chance, selection differences in intact groups which influence data patterns but are irrelevant to theory, and theories that are partially or totally incorrect. Replication is crucial when making higher-order interaction predictions. This helps control for chance fluctuations.

The importance of the relative complexity of the interaction predictions can be further illustrated from an archival quasi-experiment by Seaver (1973) who was interested in examining the effects of a teacher's performance expectancies on students' academic achievement. To do this, Seaver located from school records a group of children whose older siblings had obtained high or low achievement scores and grades in school. He then divided the two groups of younger children into those who had had the same teacher as their sibling and those who had had a different teacher. This resulted in a 2 × 2 design (same or different teacher crossed with high- or low-performing sibling). Seaver predicted that children with high-performing siblings would out-perform children with low-performing siblings by a greater amount if they had had the same teachers as their siblings than if they had had a different teacher.

Seaver obtained the predicted interaction on several subsets of the Stanford Achievement Test, and the means indicated support for the teacher expectancy hypothesis. Moreover, it is not easy to invoke a selection alternative interpretation. The one that springs most readily to mind is that children who had low-performing siblings might be assigned to teachers with a reputation for "handling difficult children'' who also happen to teach very little. Alternatively, children with high-performing siblings might be assigned to teachers with a reputation for stimulating potential "stars." But this simple selection explanation cannot be correct since children who had had different teachers were also labeled as low or high performers and so should have also been sent to a particular kind of teacher. The only selection interpretation which can be invoked is rather complicated and will not strike some readers as very plausible. It is that the children who had had different teachers were those who would have gone to teachers with reputations for dealing with high or low performers if this had been possible, but that they did not go because it was not possible. The best ways to examine this last threat in detail would be to have definite information that the assignment of teachers to children was haphazard or to have teachers equally represented in all cells of the design.

Two questions about the construct validity of the treatment in the Seaver study can be raised. First, it is assumed that it was the teacher's expectancy about the child's performance that influenced that performance. It is also possible, though probably less plausible, that it was the child's expectancy about the teacher's skill or friendliness toward the child and his or her family that influenced the child's learning. It would not be easy to dissociate these two interpretations unless one had either an experimenter-controlled manipulation of the child's and the teacher's expectancy or one had questionnaire or interview data which showed that children who manifested an expectancy effect did not expect teachers to teach them differently because they had taught their siblings. Second, there is no evidence from the study indicating why expectancy influenced performance. Was the apparent effect due to teachers calling less on children with poor-performing siblings, or to teachers reinforcing them differently, or to teachers publicly attributing less ability or motivation to them, or some other reason? Of course, the Seaver study was designed to answer questions about whether an expectancy effect could be demonstrated at all in a nonreactive archival quasi-experiment, and an examination of process variables was not intended. This was probably just as well, for archival experiments tend to be weak on process since they are typically set up to record performance outcomes and not the processes mediating performance.

The investigator who has only posttest data is indeed fortunate if he or she can translate the research hypothesis into an interaction in which one group of respondents is superior to some other group in one experimental condition and is inferior in another. Nisbett and Kanouse succeeded in doing this, as did Seaver. The major threat to the single-interaction design is that of selection, and the basic design's interpretability depends in large measure on how well selection artifacts can be explicitly ruled out or rendered less plausible. One technique for reducing the plausibility of selection is to make the interaction hypothesis involve a second- or third-order interaction. However, it is ironic that, on the one hand, interpretability increases with the *specificity* of predictions about particular statistics or particular interaction patterns; on the other hand, the *probability of obtaining* specific and expected data outcomes decreases with the very specificity of the predictions! Nonetheless the interaction prediction designs we have just outlined are very useful if carefully interpreted.

The Regression-Discontinuity Design

When people or groups are given awards or those in special need are given extra help, one would like to discover the consequences of such provisions. A regression-discontinuity design is often appropriate for these situations. The logic behind the design is simple. Imagine that respondents can be classified according to scores on a quantified continuum with a specified cutting point. Persons who score above the point gain, say, an award, while those who score below it do not. If the award had any influence on a particular outcome variable, a discontinuity should appear at the cutting point when separate regression lines are computed for the two groups. This is because the persons above the cutting point should have had their outcome scores increased by the award while those below the point should not have.

To illustrate this, imagine the situation in which one has a continuous interval scale measure of pretest organizational performance, perhaps output level or grades. One then gives a bonus to those persons who score above a particular output or school grade level—to be followed by an evaluation of the award's effects on subsequent performance. One could draw a graph with pretest output level along the horizontal and posttest output level along the vertical. A scat-

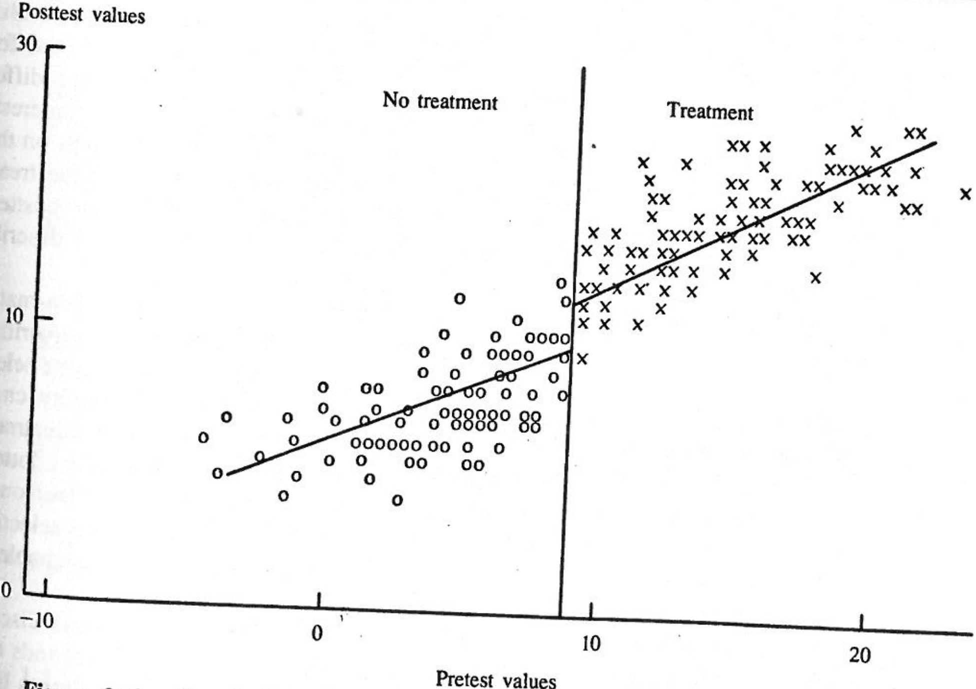


Figure 3.10. Hypothetical outcome of a pretest-posttest regression-discontinuity quasi-experiment.

terplot could then be computed which relates each person's posttest score to his or her pretest score. If the award were effective, there should be a discontinuity when fitting separate regression lines to the individuals above and below the pretest performance cutting point. Such a hypothetical case is portrayed in Figure 3.10.

Portraying the data from the regression-discontinuity design in this fashion highlights the similarity to the way some treatment effects would appear in a randomized experiment when posttest scores are plotted as a function of pretest scores. In the case of the randomized experiment, a main effect of the treatment would make the regression line of the treated group higher than that of the control group. If there were no interaction of treatment and pretest, this would hold for all values of the pretest. In the regression-discontinuity case where awards are dispensed, the treatment is successful, and only a main effect is obtained, a treatment effect would look the same except that all the control cases above the cutting point, and all the experimental cases below it, would be missing.

The similarity in regression plots indicating a simple main effect of the treatment is not the principal resemblance between the regression-discontinuity design and the randomized experiment. In the nonequivalent group designs discussed in this chapter the selection and selection-maturation processes were at best imperfectly known. Sometimes, descriptive data from populations similar to but not identical to those being examined can help ascertain gross selection differences and, where such data are longitudinal, can help ascertain selection-maturation. But such descriptive data rarely allow exact estimates of the selection confound for particular samples. In other cases, pretest data are available from units that differ in age or experience and who themselves will provide the posttest data of interest. Estimates of selection-maturation can then be generated for different groups on the assumption that within-group changes for a specific time period before the treatment would be identical to changes during the (equivalent length) pretest-posttest time period. In many other cases, only imperfect proxies are available for describing selection differences such as sex, race, place of birth, and the like.

In all the above cases, knowledge of the selection and related selection-maturation processes is at best imperfect and dependent on accepting unverified assumptions. With regression-discontinuity, as with random assignment, the selection process is known perfectly in theory. In the regression-discontinuity case, selection is based on the fallible scores used for assigning persons to the treatment or control groups. In the randomized experiment, assignment is based on a lottery and the average person in one treatment group is similar to the average person in another. As with the randomized experiment, it is knowledge of the selection process that makes the regression-discontinuity design so potentially amenable to causal interpretation.

It may occur to some readers that effects can be obtained other than a discontinuity in the level of the regressions at the cutting point—which corresponds to a simple main effect of the treatment. For example, the slope of the regression lines might differ on each side of the cutting point. At first glance, this difference in slope would imply an interaction of the pretest and treatment such that the persons on the treatment side of the point do better or worse depending on where their scores fall above the cutting point. For instance, suppose the National Science Foundation awarded individual fellowships to graduate students solely on the basis of Graduate Record Examination scores—which is not the case. An evaluation of the awards' effectiveness using a regression-discontinuity design might seem to imply, when regression slopes differ and the steeper one is on the awards side of the cutting point, that fellowships have more of an impact on students whose GRE scores are among the highest than on students whose scores just qualified them for fellowships. However, the interpretation of differences in slope is extremely difficult, if not impossible, in the absence of intercept differences.

The following discussion of the regression-discontinuity design is divided into two sections. The first deals with the use of pretest scores to classify units and assign them to treatments, and the second with the use of quantitative scores other than the pretest. This distinction has no theoretical importance, and is made only to remind readers that the regression-discontinuity design can be used when no pretest measures are available.

Regression-Discontinuity with Similar-Appearing Pretest Measures

Seaver and Quarton (1976) used the regression-discontinuity design to examine how college students' grades in one quarter were affected by making the Dean's list on the basis of their grades from the previous quarter. The investigators obtained grades for 1,002 students for the quarters before and after the list was published; their sample included persons who did and did not qualify for the

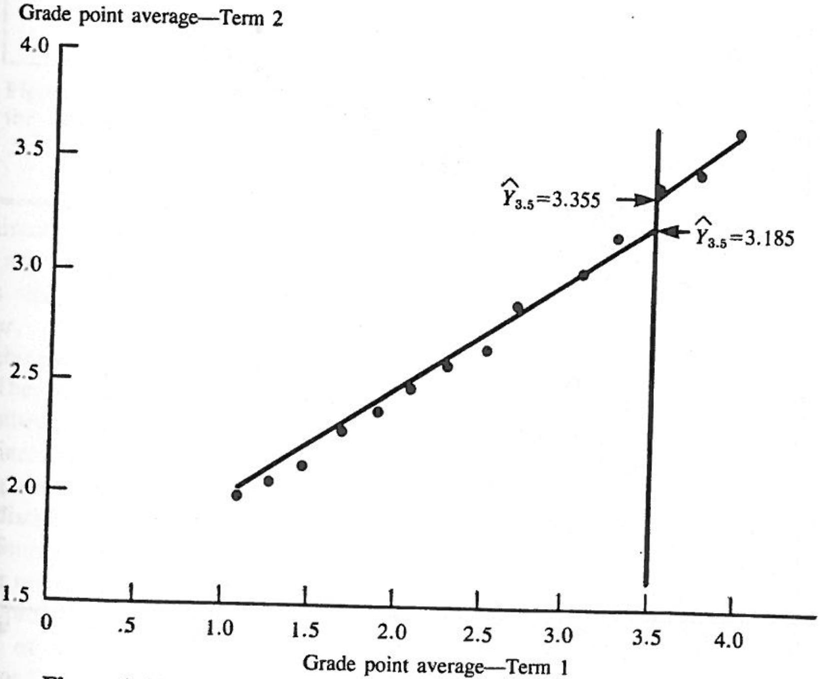


Figure 3.11. Regression of grade point average Term 2 on grade point average Term 1 for the non-Dean's list and Dean's list groups.

distinction. We would expect students who made the list to do better than those who did not for a variety of reasons. The issue with regression-discontinuity analysis is whether the rewarded students perform better than others by a higher factor than would have been the case without the reward.

Seaver kindly provided Joyce Sween and the authors of this book with his data, which Sween has replotted. When plotted as individual scores instead of the array means in Figure 3.11, it looks as though grades are curvilinearly related to each other across quarters. Since the plot of 1,002 scores is complex, we have replotted the data in Figure 3.12 using finer array means than Seaver and Quarton, and we have fitted both curvilinear and linear regressions to the data. Obviously, a strong case can be made for curvilinearity in the data.

What are the consequences of the underlying relationship being curvilinear? As Figure 3.12 shows, fitting a curvilinear trend produces no evidence of a dis-

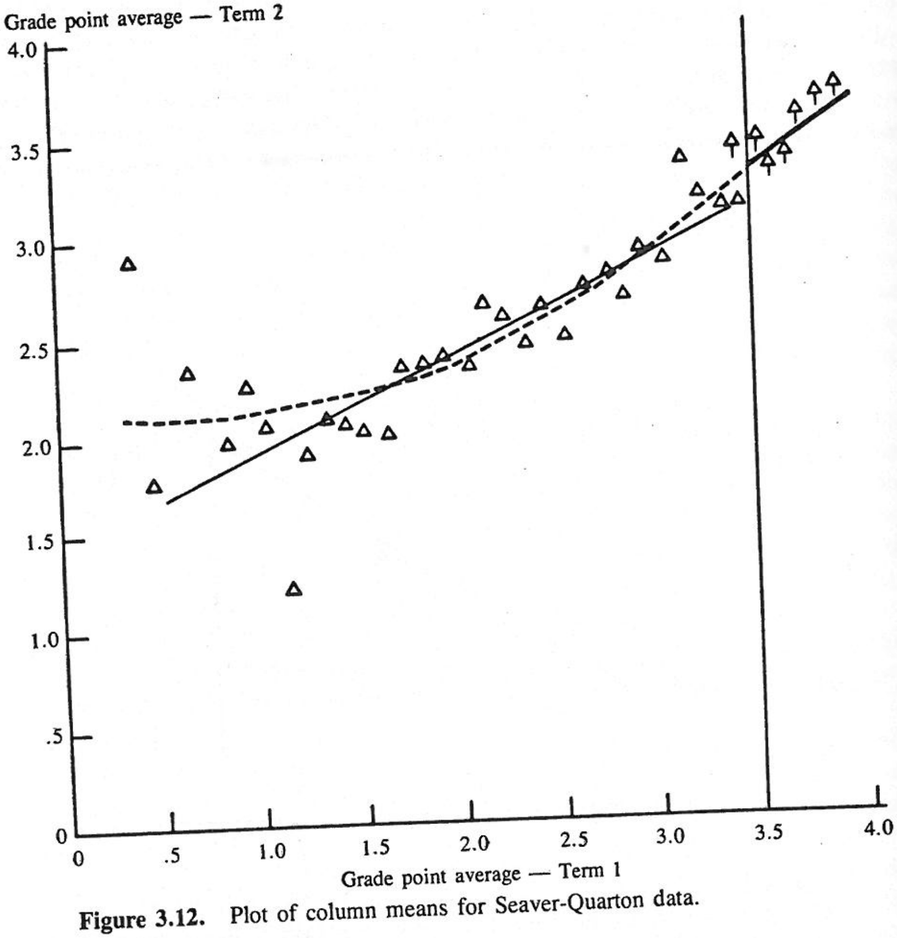


Figure 3.12. Plot of column means for Seaver-Quarton data.

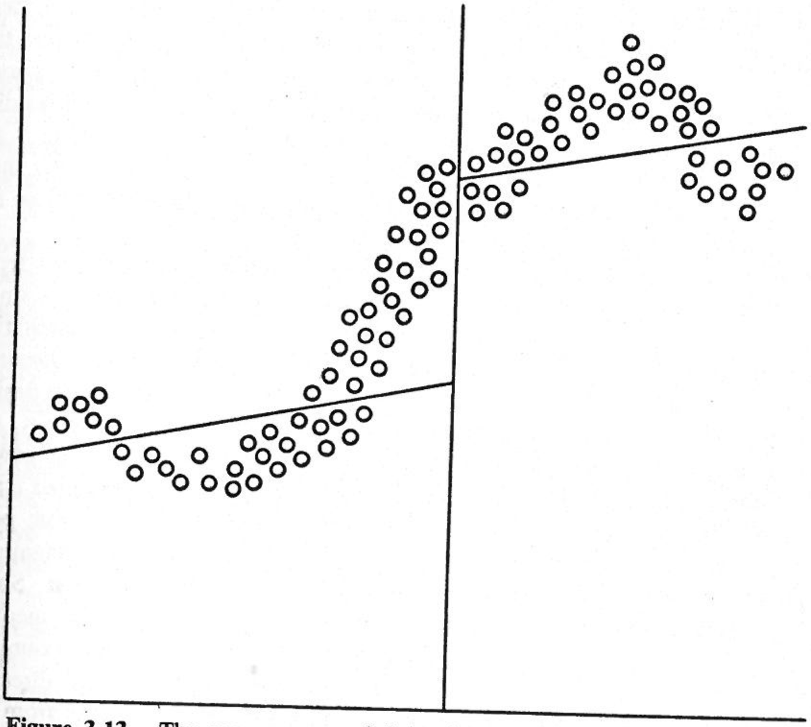


Figure 3.13. The consequences of fitting inappropriate linear regressions when the underlying regression is not linear.

continuity at the cutting point and so no evidence that making the Dean's list had any effect. In other words, when linear regressions are fitted to the data on each side of a cutting point, *but one or more of the regressions are not in fact linear*, spurious intercept differences can emerge. So, too, can spurious slope differences.

The point we are making can perhaps best be understood by considering the hypothetical example in Figure 3.13 where the scatterplot portrays an underlying nonlinear relationship. Fitting linear relationships gives a discontinuity at the cutting point; fitting a nonlinear regression does not. Imagine, now, that the underlying distribution was of a J-form and that the tail at the high end of the pretest distribution in Figure 3.13 did not exist. Then, it would not be appropriate to fit a linear regression to the data at the lower end of the pretest continuum, but it would be appropriate at the higher end. However, if linear regressions were fitted on both sides of the cutting point, a discontinuity in both level and slope would be obtained. The major threat to internal validity with the regression-discontinuity design is of selection-maturation—the possibility that change between the pretest and posttest does not follow a simple linear pattern across the whole of the pretest distribution.

Seaver and Quarton were actually sensitive to this threat. But instead of examining the form of their data more closely for the broader range of cases who did not make the Dean's list, they chose another strategy. They reasoned that, if there were spurious selection-maturation, then it should appear in the data when the pretest in Figure 3.11 was used as the posttest and the previous quarter's grades were used as the pretest. When this analysis was conducted, it produced no discontinuities. However, the test is not as accurate as critically examining scatterplots to examine whether nonlinear forms of regression fit the data better than linear forms. If they do, regressions of a higher order than the linear have to be fit; otherwise spurious causal conclusions will result.

Another set of problems with the regression-discontinuity design arises from the fact that in most social settings the treatment is made available only to a small percentage of the persons who score at one of the extremes of the quantitative continuum. This is not, of course, a necessary feature of the regression-discontinuity design, though it is certainly to be expected in many cases where the design is appropriate. Being restricted to the most needy and the most meritorious, the regression-discontinuity design is therefore weak when researchers wish to generalize to other kinds of persons. For instance, in an example to be presented later, it is shown that Medicaid significantly increased the number of visits that poor persons reportedly made to doctors. This effect, while small in average magnitude (though visually dramatic), represents many millions of dollars per year. Some health planners would like to be able to estimate how the demand for medical services will increase once we have some form of a national health insurance program where individuals of all kinds do not pay for medical services directly. How confident would one feel generalizing to the United States at large from the Medicaid experience with the most-poor and least-healthy? (In practice, of course, one would not rely solely on Medicaid for an estimate of the demand for health services—some labor unions have won completely free medical services for their members, and some cities have private medical schemes allowing unlimited free medical services).

The dependence on small numbers of extremely high or low scores raises a more technical problem: It is often difficult to estimate with any reasonable certainty the shape of the distribution of scores on the short side of the cutting point. Yet, as we shall see in chapter 4 where the statistical analysis of data from the regression-discontinuity design is discussed, it is important to be able to estimate this distribution in order to determine if it can be considered simply an extension of the distribution found on the long side of the cutting point.

A difficulty that can often be anticipated with the regression-discontinuity design is that the cutting point will not be as clearcut as our discussion may suggest. In particular, we can anticipate that some persons will be allowed access to the treatment, not because of need or merit, but because they are friends of friends, or are particularly skillful at manipulating the persons responsible for permitting access, or are owed some type of social debt. A lack of clarity about the cutting point is especially likely when the cutting point is widely known, for this may give rise to special pressures to help some persons achieve the cutting point score. For instance, the Irish government publishes the passing score on various national examinations in education. A frequency distribution of the number of children obtaining all possible scores on the physics exam shows a less than expected number of students scoring just below the cutting point and a higher than expected frequency just above it (Greaney, Kellaghan, Takata and Campbell, in preparation). It seems likely, therefore, that examiners gave students scoring just below the cutting point "an extra helping hand." In many social service settings, clients disguise part of their income if they suspect that full disclosure will take them above a cutting point for eligibility for services. Indeed, some professionals may deliberately condone such practices because the cutting points for eligibility seem so arbitrary to them.

A major problem when cutting points are fuzzy is that the systematic deviations around the cutting point can masquerade as treatment effects. Imagine a social service setting where the clients know the income cutoff point for obtaining supplementary social services. Some will report their income to be lower than it actually is. In order to examine whether obtaining supplementary services increased the social mobility aspirations of children, we might plot the income of a wide range of parents against the mobility aspirations of their children. Let us suppose that the overall relationship is linear and positive, indicating that higher incomes are associated with higher aspirations. However, one group of parents whose reported incomes fall below the cutting point will have actual incomes above it. The aspiration scores of their children will be higher than those of parents who have comparable reported income but lower actual income. Combining the scores of all persons with comparable reported incomes will artifactually increase the obtained scores of those children who score just below the cutting point. As a result, we will have a reduced intercept difference at the cutting point and spurious differences in slope.

Since achieving the cutting point permits access to scarce and desired resources, we should always anticipate that cutting points will be unclear and that some individuals will gain what, strictly speaking, they should not receive. Where such individuals can be identified, the solution is to drop them from the analysis and proceed as though they were not part of the study. (It is always desirable, of course, to try and understand the pressures that lead to such "biased" [but probably sociologically lawful] assignments.) If it is not possible to identify all the individuals who received treatments for which they were not strictly eligible, then an estimate should be made of the range around the cutting point in which biased assignments are likely to have occurred. This range is then treated as the cutting point, and the analysis proceeds with a hatched area demarcating the cutting range rather than with a straight line at a particular cutting score. The logic of the analysis is the same whether a range or cutting point is used, although statistical sensitivity will be greater with a point.

Quantified Multiple Control Groups, Posttest-Only Design

Using the pretest to classify units on a merit or need basis is merely a special case of a more general principle. The principle states that regression-discontinuity designs are possible wherever units can be ordered along some quantifiable dimension which is systematically related to assignment of treatment. We can illustrate this by referring to a study by Lohr (1972; Wilder, 1972), who was interested in exploring the effects of Medicaid. The program was designed to make medical care available to the very poor (income under $3,000 per family per year) by means of federal government payments. One important question was whether the poor would actually take advantage of the new medical policy.

Lohr's data can be displayed to plot the mean visits to the doctor per family per year as a function of annual family income (each measure was based on interviews done in connection with the Current Population Reports). The relationship of the two variables is portrayed in Figure 3.14, where it can be seen that the number of visits per year systematically decreases as income decreases. The one discontinuity from this trend is for families with an income under $3,000, where the number of medical visits sharply increases and even tends to exceed the number of visits made by the more affluent families. Since these data indicate that Medicaid might have increased medical visits by the poor, we have to ask ourselves the perennial question: Are there any plausible alternative interpretations of the relationship?

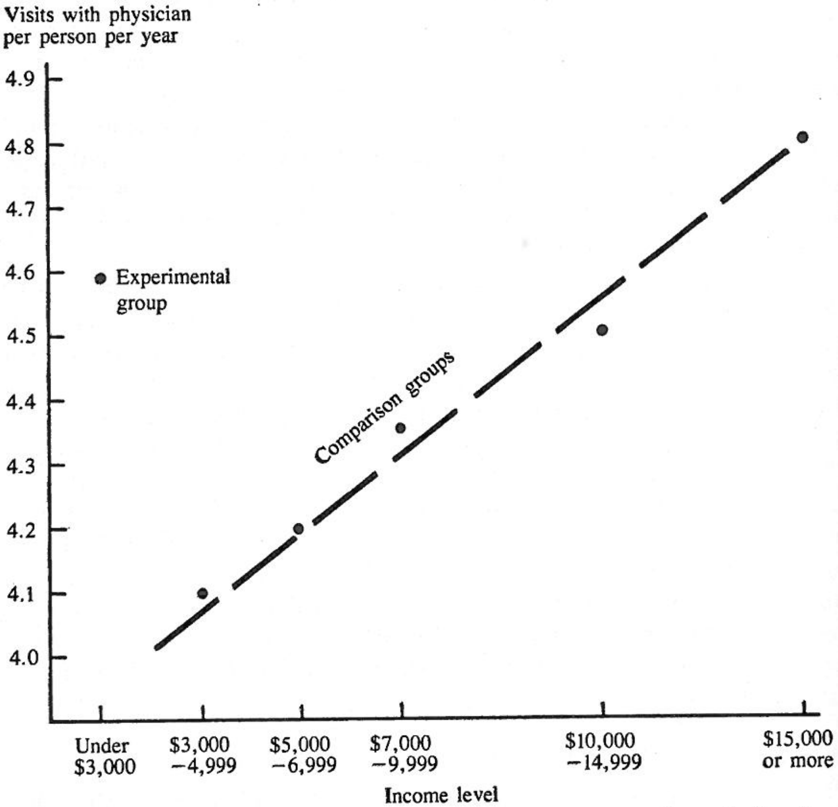


Figure 3.14. Quantified multiple control group posttest-only analysis of the effects of Medicaid. (After Lohr, 1972; Wilder, 1972.)

The chronically sick aside, visits to the doctor are presumably highest among the aged. Income is also lower among the aged. Thus, if the aged fell disproportionately into the lowest-income category, the relationship in Figure 3.14 might reflect a special selection phenomenon. Against this fact, we have to consider that there is no reason why the aged should be so heavily represented in the lowest-income group rather than spread more systematically across each lower-income group. Fortunately, the relationship of age to income is ultimately an empirical issue and national demographic data exist for checking it. It is perhaps more important to note that persons over 65 are eligible for Medicare as well as Medicaid and that there are indications that many older persons use both programs. Hence, an evaluation of Medicaid by itself should be restricted to persons under 65 years of age, though an evaluation of the program in its social context should also include separate analyses of persons over 65. Thus, there is good reason for wanting to see the Figure 3.14 data presented separately for families where no one is eligible for Medicare and for families where someone is eligible.

A different age explanation is based on the possibility that medical visits are most needed by pregnant women and young children. Hence, we have to consider whether the lowest-income group was disproportionately composed of persons prone to pregnancy and large families. If so, they might have had more frequent visits to doctors even before Medicaid, though such visits were presumably to state hospitals on a nonpayment basis. There is every need, therefore, to disaggregate the data even further to examine the relationship of income to medical visits among persons of different family sizes as well as at different age levels.

A different kind of possible selection bias cannot be ruled out merely by disaggregating on the basis of demographic factors that are routinely measured in surveys and that can be easily used in archival studies. Some families in the lowest-income group were eligible for assistance from many programs, some of which mandated (and paid for) medical visits by recipients and their children as a precondition for receiving aid or for continuing to receive it. The issue, therefore, arises: Were the disproportionately frequent visits to doctors by the poor the result of Medicaid meeting a need or a response to the pre-Medicaid requirement of other programs that a doctor be consulted? If the latter, no effect of Medicaid would need to be invoked. This problem would be easy to solve if we knew something about the programs in which family members were enrolled, especially those requiring work-related and welfare-related physical checkups. Only if data on mandated checkups were collected at the time of the survey could disaggregation on this variable take place. Without foresight, however, the required information would probably not be collected. In that case, the best one could do would be to consult the most reliable data available on the number of persons eligible for mandated medical visits and determine (1) if such eligibility related to income in the discontinuous manner suggested by Figure 3.14 and (2) was of a magnitude that could plausibly account for the pattern in that figure.

Another possible alternative interpretation is based on selection-maturation. This interpretation suggests that the demand for medical care was greatest among the poor, that the supply of doctors was increasing year by year, and that the new supply could only find outlets among those sections of the population whose prior demands had not been met and whose physical state required urgent care. Against this, however, is the fact reported to us by Lohr that, though the number of doctors per capita increased between 1960 and 1970, the number *in medical practice* did not. Presumably, some doctors went into medical research or into nonmedical careers.

A final threat to internal validity arises because the direction of causality is not clear from Figure 3.14. Did Medicaid cause an increase in medical visits, or did the desire for medical visits by the sick and hypochondriac lead these persons to underreport their true income both to doctors and to interviewers in order to continue the pretense that they were eligible for such programs? An indirect check of this might be possible by using nonmedical surveys to estimate the proportion of persons in each of the income categories in Figure 3.14. The "opposite direction of causality" explanation would be ruled out if equal proportions fell into each income category in each type of survey. (However, this test is only approximate, for sick persons might generalize their underreporting of true income to all surveys, whether oriented to medical services or not.)

What perhaps should be emphasized regarding most of the threats to internal validity we have listed is that their plausibility can be assessed without undue effort by consulting available archives in order to disaggregate the data from Figure 3.14 or to collect additional data that rule out specific alternative explanations. Thus, our list of threats should not discourage researchers; like other lists for other projects, it should spur into action those persons whose interest lies in strengthening a particular causal inference.

It should be noted that the Lohr-Wilder data actually cover three waves, one coming before Medicaid. As displayed in Riecken et al. (1974, Fig. 4.18), the data for the year prior to Medicaid indicated that doctors devoted the lowest level of attention to the least financially advantaged group, which—as Figure 3.14 shows—was not the case after Medicaid. Such a change invalidates many of the alternative interpretations listed above that were presented here for pedagogical reasons.

Some important problems of construct validity should also be mentioned with respect to Lohr's quasi-experiment. Given the stimulation of demand by Medicaid and the apparent inelasticity of supply, does an increase in the quantity of care for the poor entail a decrease in the quality of care for them and for others? Furthermore, is the dependent variable appropriately labeled as "an increase in physician visits" or as "a *temporary* increase in physician visits"? The frequency of chronic and ill-monitored disease is presumably higher among the poor and might well be decreased by Medicaid, thereby leading to a later decrease in visits as more and more chronic problems are cured or detected before they become worse.