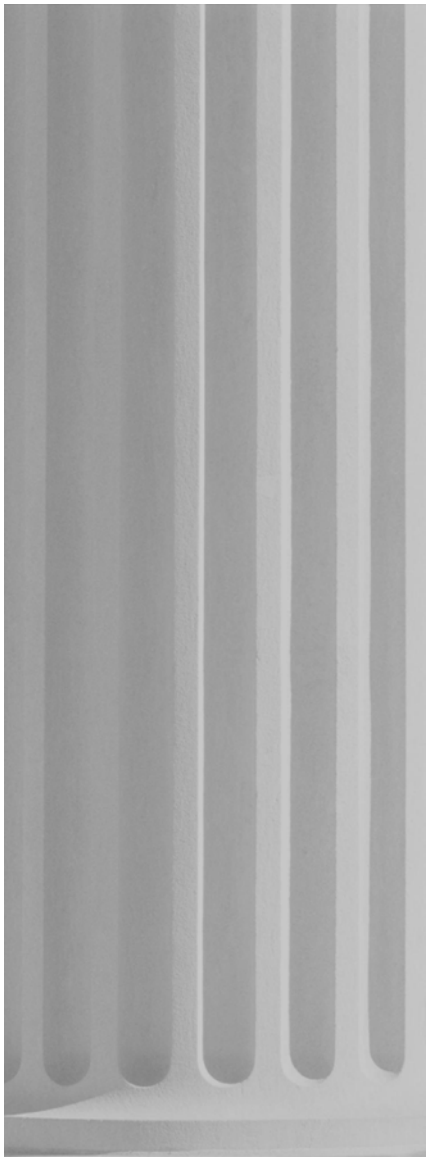
CHAPTER 10

# Quasi-Experimental Design

## KEY TERMS

construct validity
control group
double-pretest design
internal validity
nonequivalent dependent variables (NEDV) design
nonequivalent-groups design (NEGD)
null case
pattern-matching
pattern-matching NEDV design
proxy-pretest design
quantitative
quasi-experimental design
random assignment
regression-discontinuity (RD)
regression line
regression point displacement (RPD)
regression to the mean
selection bias
selection-history threat
selection instrumentation
selection-maturation threat
selection mortality
separate pre-post samples
selection regression
selection testing
selection threat
statistics
switching-replications design
threats to internal validity



## OUTLINE



**quasi-experimental design**
Research designs that have several of the key features of randomized experimental designs, such as pre-post measurement and treatment-control group comparisons, but lack random assignment to a treatment group.

**internal validity**
The approximate truth of inferences regarding cause-effect or causal relationships.

A **quasi-experimental design** is one that looks a bit like an experimental design but lacks the key ingredient—random assignment. My mentor, Don Campbell, often referred to these designs as "queasy" experiments because they give experimental purists a queasy feeling. With respect to **internal validity**, they often appear to be inferior to randomized experiments. However, there is something compelling about these designs; taken as a group, they are more frequently implemented than their randomized cousins.

I'm not going to try to cover the quasi-experimental designs comprehensively. Instead, I'll present two of the classic quasi-experimental designs in some detail. Probably the most commonly used quasi-experimental design (and it may be the most commonly used of all designs) is the nonequivalent-groups design (NEGD). In its simplest form, it requires a pretest and posttest for a treated and comparison group. It's identical to the Analysis of Covariance (ANCOVA) randomized experimental design (see Chapter 9), except that the groups are not created through random assignment. You will see that the lack of random assignment and the potential nonequivalence between the groups complicates the statistical analysis of the nonequivalent groups design (as covered in the discussion of analysis in Chapter 12).

The second design I'll focus on is the regression-discontinuity design. I'm not including it just because I did my dissertation on it and wrote a book about it (although those were certainly factors weighing in its favor). I include it because I believe it is an important (and often misunderstood) alternative to randomized experiments because its distinguishing characteristic—assignment to treatment using a cutoff score on a pretreatment variable—allows you to assign to the program those who need or deserve it most. At first glance, the regression-discontinuity design strikes most people as biased because of regression to the mean (discussed in Chapter 7). After all, you're assigning low scorers to one group and high scorers to the other. In the discussion of the statistical analysis of the regression discontinuity design (see Chapter 12), I'll show you why this isn't the case.

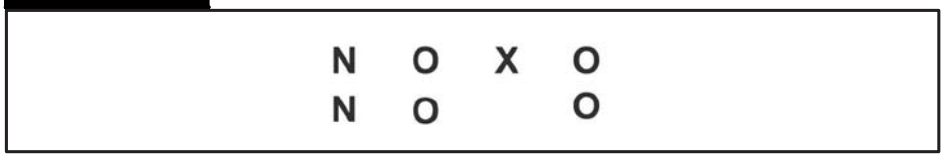
Finally, I'll briefly present an assortment of other quasi-experiments that have specific applicability or noteworthy features, including the proxy-pretest design, double-pretest design, nonequivalent dependent-variables design, pattern-matching design, and the regression point displacement design.

# 10-1 The Nonequivalent-Groups Design

**nonequivalent-groups design (NEGD)**
A pre-post two-group quasi-experimental design structured like a pretest-posttest randomized experiment, but lacking random assignment to group.

The **nonequivalent-groups design (NEGD)** is probably the most frequently used design in social research. Why? Because it is one of the most intuitively sensible designs around. If you want to study the effects of your program, you probably recognize the need to have a group of people receive the program. That's your program group, and, you probably see that it would be sensible to measure that group before and after the program so you can see how much the program improved or

**FIGURE 10–1** **Notation for the nonequivalent-groups design (NEGD)**




changed them. That's the pre-post measurement. Once you understand the basic problem of internal validity (see Chapter 7), you will readily admit that it would be nice to have a comparable group that differs from your program group in only one respect—it doesn't get the program. That's your **control group**. Put all of these elements together and you have the basic NEGD. Although the design is intuitively straightforward, it is not without its difficulties or challenges. The major challenge stems from the term *nonequivalent* in its title. If your comparison group is really similar to the program group in all respects—except for receiving the program—this design is an excellent one. But how do you ensure that the groups are equivalent? And, what do you do if they are not? That's the central challenge for this design and I'll take some time here to address this issue.

**control group**
A group, comparable to the program group, that did not receive the program.

## 10-1a The Basic Design

The NEGD is structured like a pretest-posttest randomized experiment, but it lacks the key feature of the randomized designs—**random assignment**. The design notation for the basic NEGD is shown in Figure 10–1.

**random assignment**
Process of assigning your sample into two or more subgroups by chance. Procedures for random assignment can vary from flipping a coin to using a table of random numbers to using the random number capability built into a computer.

In the NEGD, you most often use intact groups that you think are similar as the treatment and control groups. In education, you might pick two comparable classrooms or schools. In community-based research, you might use similar communities. You try to select groups that are as similar as possible so that you can fairly compare the treated one with the comparison one; but you can never be sure the groups are comparable. Put another way, it's unlikely that the two groups would be as similar as they would if you assigned them through a random lottery. Because it's often likely that the groups are not equivalent, this design was named the *nonequivalent-groups design* to remind us of that. The design notation (see Figure 10–1) uses the letter *N* to indicate that the groups are nonequivalent.

So, what does the term *nonequivalent* mean? In one sense, it means that assignment to group was not random. In other words, the researcher did not control the assignment to groups through the mechanism of random assignment. As a result, the groups may be different prior to the study. That is, the NEGD is especially susceptible to the internal validity threat of selection (see Chapter 7). Any previous differences between the groups may affect the outcome of the study. Under the worst circumstances, this can lead you to conclude that your program didn't make a difference, when in fact it did, or that it did make a difference, when in fact it didn't.

## 10-1b The Bivariate Distribution

Let's begin our exploration of the NEGD by looking at some hypothetical results. Figure 10–2a shows a bivariate distribution in the simple pre-post, two-group study. The *treated cases* are indicated with *X*s and the *comparison cases* are indicated with *O*s. A couple of things should be obvious from the graph. To begin, you don't even need **statistics** to see that there is a whopping treatment effect. (Although statistics would help you estimate the size of that effect more precisely.) The program cases (*X*s) consistently score better on the posttest than the comparison cases (*O*s) do. If positive scores on the posttest are better, you can conclude that the program improved things. Second, in the NEGD the biggest threat to internal validity is

**statistics**
The process of estimating various features from data, often using probability theory.

**FIGURE 10–2a** **Bivariate distribution for a hypothetical example of a nonequivalent-groups design**

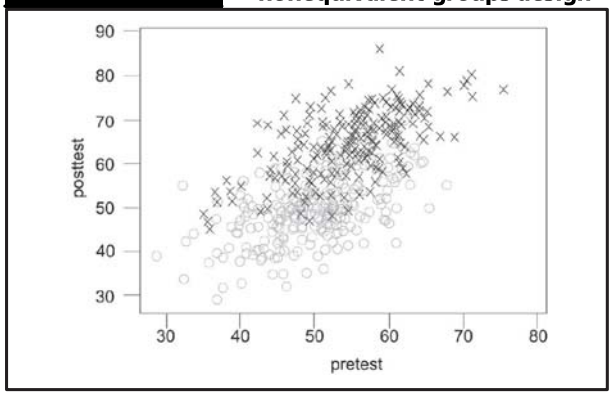
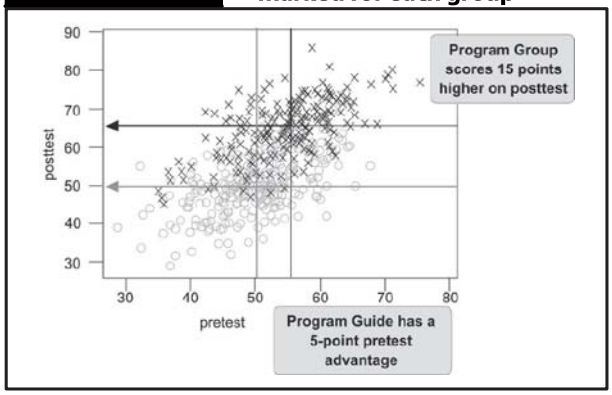



**FIGURE 10–2b** **Nonequivalent-groups design with pretest and posttest averages marked for each group**




selection—that the groups differed before the program. Does that appear to be the case here? Although it may be harder to see, the program group does appear to be a little further to the right on average. This suggests that program group participants did have an initial advantage on the pretest and that the positive results may be due in whole or in part to this initial difference.

**selection bias**
Any factor other than the program that leads to posttest differences between groups.

You can see the initial difference, the **selection bias**, when you look at the graph in Figure 10–2b. It shows that the program group scored about 5 points higher than the comparison group on the pretest. The comparison group had a pretest average of about 50, whereas the program group averaged about 55. It also shows that the program group scored about 15 points higher than the comparison group on the posttest. That is, the comparison group posttest score was again about 55, whereas this time the program group scored around 65. These observations suggest that there is a potential **selection threat,** although the initial 5-point difference doesn't explain why you observe a 15-point difference on the posttest. It may be that there is still a legitimate treatment effect here, even given the initial advantage of the program group.

**selection threat**
Any factor other than the program that leads to post test differences between groups

**Possible Outcome 1**[1] Let's take a look at several different possible outcomes from a NEGD to see how they might be interpreted. The important point here is that each of these outcomes has a different storyline. Some are more susceptible to threats to internal validity than others. Before you read each of the descriptions, take a good look at the associated graph and try to figure out how you would explain the results. If you were a critic, what kinds of problems would you be looking for? Then, read the synopsis and see if it agrees with your perception.

Sometimes it's useful to look at the means for the two groups. Figure 10–3 shows the means for the distribution in with the pre-post means of the program group joined with a line beginning and ending with triangles and the pre-post means of the comparison group joined with a line beginning and ending with squares. This first outcome shows the situation in the two bivariate plots. Here, you can see much more clearly both the original pretest difference of 5 points and the larger 15-point posttest difference.

How might you interpret these results? To begin, you need to recall that with the NEGD you are usually most concerned about selection threats. Which selection threats might be operating here? The key to understanding this outcome is that the comparison group did not change between the pretest and the posttest. Therefore, it would be hard to argue that that the outcome is due to a **selection-maturation threat**. Why? Remember that a selection-maturation threat means that the groups

**selection-maturation threat**
A threat to internal validity that arises from any differential rates of normal growth between pretest and posttest for the groups.

[1]The discussion of the five possible outcomes is based on the discussion in Cook, T.D. and Campbell, D.T. (1979). *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Houghton Mifflin, Boston, pp. 103–112.

**FIGURE 10–3** **Plot of pretest and posttest means for possible outcome 1**

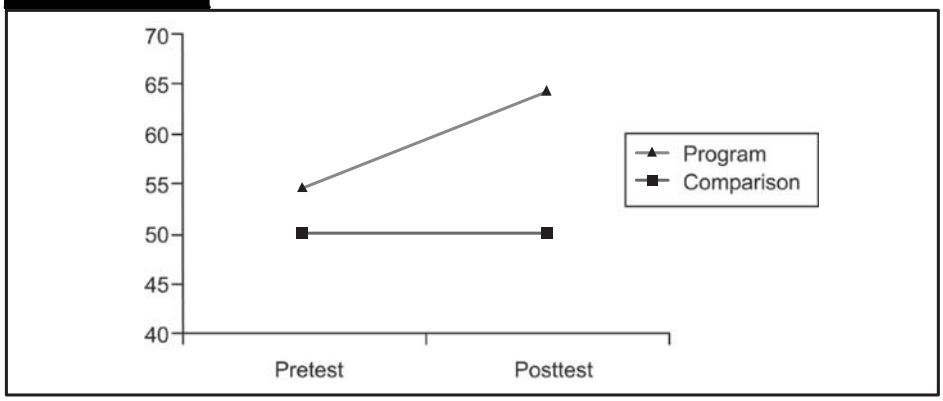
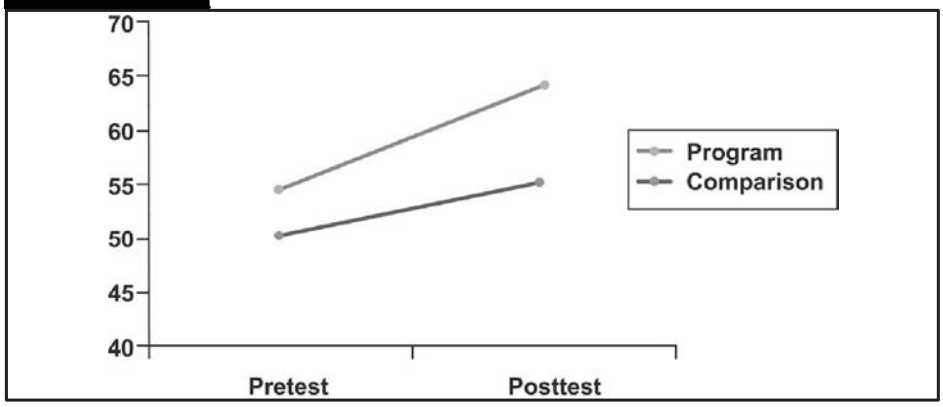



**FIGURE 10–4** **Plot of pretest and posttest means for possible outcome 2**




are maturing at different rates and that this creates the illusion of a program effect when there is not one. However, because the comparison group didn't mature (change) at all, it's hard to argue that differential maturation produced the outcome. What could have produced the outcome? A **selection-history threat** certainly seems plausible. Perhaps some event occurred (other than the program) that the program group reacted to and the comparison group didn't. Maybe a local event occurred for the program group but not for the comparison group. Notice how much more likely it is that outcome pattern 1 is caused by such a history threat than by a maturation difference. What about the possibility of **selection regression?** This one actually works a lot like the selection-maturation threat. If the jump in the program group is due to **regression to the mean**, it would have to be because the program group was below the overall population pretest average and consequently regressed upwards on the posttest. However, if that's true, it should be even more the case for the comparison group who started with an even lower pretest average. The fact that it doesn't appear to regress at all helps rule out the possibility that outcome 1 is the result of regression to the mean.

**selection-history threat**
A threat to internal validity that results from any other event that occurs between pretest and posttest that the groups experience differently.

**selection regression**
A threat to internal validity that occurs when there are different rates of regression to the mean in the two groups.

**regression to the mean**
See regression threat.

**Possible Outcome 2** The second hypothetical outcome (see Figure 10–4) presents a different picture. Here, both the program and comparison groups gain from pre to post, with the program group gaining at a slightly faster rate. This is almost the definition of a selection-maturation threat. The fact that the two groups differed to begin with suggests that they may already be maturing at different rates. The posttest scores don't do anything to help rule out that possibility. This outcome might also arise from a selection-history threat. If the two groups, because of

**FIGURE 10–5** Plot of pretest and posttest means for possible outcome 3

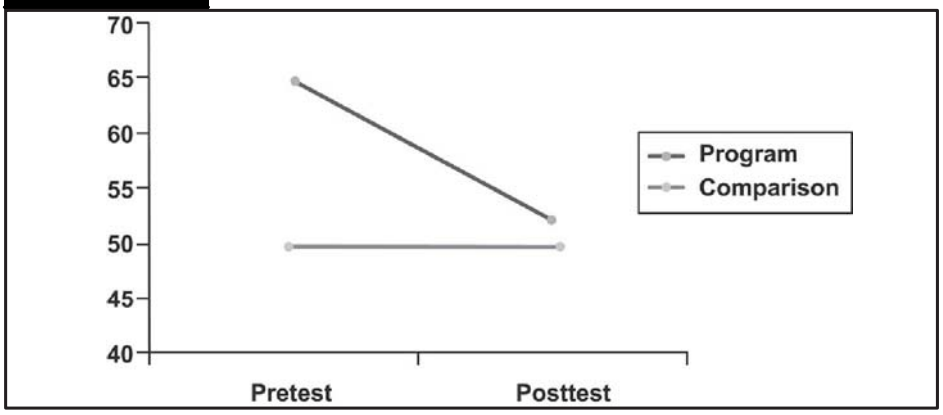
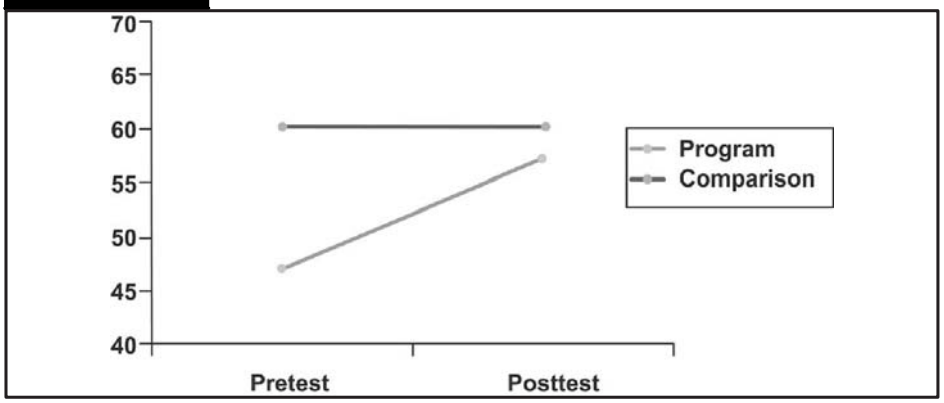



**FIGURE 10–6** Plot of pretest and posttest means for possible outcome 4




their initial differences, react differently to some historical event, you might obtain the outcome pattern shown. Both **selection testing** and **selection instrumentation** are also possibilities, depending on the nature of the measures used. This pattern could indicate a **selection-mortality** problem if there are more low-scoring program cases that drop out between testings. What about selection-regression? It doesn't seem likely, for much the same reasoning as for outcome 1. If there were an upward regression to the mean from pre to post, you would expect that regression to be greater for the comparison group because it has the lower pretest score.

**selection testing**
A threat to internal validity that occurs when a differential effect of taking the pretest exists between groups on the posttest.

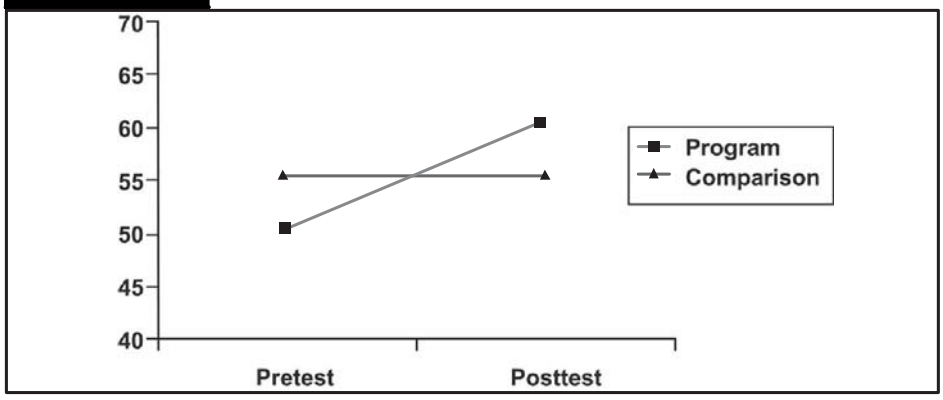
**selection instrumentation**
A threat to internal validity that results from differential changes in the test used for each group from pretest to posttest.

**selection mortality**
A threat to internal validity that arises when there is differential nonrandom dropout between groups during the test.

**Possible Outcome 3** This third possible outcome (see Figure 10–5) cries out selection-regression! Or, at least it would if it could cry out. The regression scenario is that the program group was selected so that it was extremely high (relative to the population) on the pretest. The fact that the group scored lower, approaching the comparison group on the posttest, may simply be due to its regressing toward the population mean. You might observe an outcome like this when you study the effects of giving a scholarship or an award for academic performance. You give the award because students did well (in this case, on the pretest). When you observe the group's posttest performance, relative to an average group of students, it appears to perform worse. Pure regression! Notice how this outcome doesn't suggest a selection-maturation threat. What kind of maturation process would have to occur for the highly advantaged program group to decline while a comparison group evidences no change?

**Possible Outcome 4** The fourth possible outcome also suggests a selection-regression threat (Figure 10–6). Here, the program group is disadvantaged to

**FIGURE 10–7** **Plot of pretest and posttest means for possible outcome 5**




begin with. The fact that it appears to pull closer to the comparison group on the posttest may be due to regression. This outcome pattern may be suspected in studies of compensatory programs—programs designed to help address some problem or deficiency. For instance, compensatory education programs are designed to help children who are doing poorly in some subject. They are likely to have lower pretest performance than more average comparison children. Consequently, they are likely to regress to the mean in a pattern similar to the one shown in outcome 4.

**Possible Outcome 5** This last hypothetical outcome (Figure 10–7) is sometimes referred to as a *crossover pattern*. Here, the comparison group doesn't appear to change from pre to post; but the program group does, starting out lower than the comparison group and ending up above it. This is the clearest pattern of evidence for the effectiveness of the program of all five of the hypothetical outcomes. It's hard to come up with a threat to internal validity that would be plausible here. Certainly, there is no evidence for selection maturation here unless you postulate that the two groups are involved in maturational processes that tend to start and stop and just coincidentally you caught the program group maturing while the comparison group had gone dormant. However, if that were the case, why did the program group actually cross over the comparison group? Why didn't it approach the comparison group and stop maturing? How likely is this outcome as a description of normal maturation? Not very. Similarly, this isn't a selection-regression result. Regression might explain why a low-scoring program group approaches the comparison group posttest score (as in outcome 4), but it doesn't explain why it crosses over.

Although this fifth outcome is the strongest evidence for a program effect, you can't very well construct your study expecting to find this kind of pattern. It would be a little bit like giving your program to the toughest cases and seeing whether you can improve them so much that they not only become like average cases but actually outperform them. That's an awfully big expectation with which to saddle any program. Typically, you wouldn't want to subject your program to that kind of expectation. If you do happen to find that kind of result, you really have a program effect that beats the odds.

# 10-2 The Regression-Discontinuity Design

What a terrible name! In everyday language, both parts of the term *regression-discontinuity* have primarily negative connotations. To most people *regression* implies a reversion backward or a return to some earlier, more primitive state, whereas *discontinuity* suggests an unnatural jump or shift in what might otherwise be a smoother, more continuous process. To a research methodologist, however, the

**regression-discontinuity (RD)**
A pretest-posttest program-comparison group quasi-experimental design in which a cutoff criterion on the preprogram measure is the method of assignment to group.

term *regression-discontinuity* carries no such negative meaning. Instead, the **regression-discontinuity (RD)** design is seen as a useful method for determining whether a program or treatment is effective.

The label *RD design* actually refers to a set of design variations. In its simplest most traditional form, the RD design is a pretest-posttest program-comparison group strategy. The unique characteristic that sets RD designs apart from other pre-post group designs is the method by which research participants are assigned to conditions. In RD designs, participants are assigned to program or comparison groups solely on the basis of a cutoff score on a pre-program measure. Thus, the RD design is distinguished from randomized experiments (or randomized clinical trials) and from other quasi-experimental strategies by its unique method of assignment. This cutoff criterion implies the major advantage of RD designs; they are appropriate when you want to target a program or treatment to those who most need or deserve it. Thus, unlike its randomized or quasi-experimental alternatives, the RD design does not require you to assign potentially needy individuals to a no-program comparison group to evaluate the effectiveness of a program.

The RD design has not been used frequently in social research. The most common implementation has been in compensatory education evaluation where school children who obtain scores that fall below some predetermined cutoff value on an achievement test are assigned to remedial training designed to improve their performance. The low frequency of use may be attributable to several factors. Certainly, the design is a relative latecomer. Its first major field tests did not occur until the mid-1970s, when it was incorporated into the nationwide evaluation system for compensatory education programs funded under Title I of the Elementary and Secondary Education Act (ESEA) of 1965. In many situations, the design has not been used because one or more key criteria were absent. For instance, RD designs force administrators to assign participants to conditions solely on the basis of quantitative indicators, thereby often unpalatably restricting the degree to which judgment, discretion, or favoritism can be used. Perhaps the most telling reason for the lack of wider adoption of the RD design is that at first glance the design doesn't seem to make sense. In most research, you want to have comparison groups that are equivalent to program groups on pre-program indicators so that post-program differences can be attributed to the program itself. However, because of the cutoff criterion in RD designs, program and comparison groups are deliberately and maximally different on pre-program characteristics, an apparently insensible anomaly. An understanding of how the design actually works depends on at least a conceptual familiarity with regression analysis, thereby making the strategy a difficult one to convey to nonstatistical audiences.

Despite its lack of use, the RD design has great potential for evaluation and social research. From a methodological point of view, inferences drawn from a well-implemented RD design are comparable in internal validity to conclusions from randomized experiments. Thus, the RD design is a strong competitor to randomized designs when causal hypotheses are being investigated. From an ethical perspective, RD designs are compatible with the goal of getting the program to those most in need. It is not necessary to deny the program from potentially deserving recipients simply for the sake of a scientific test. From an administrative viewpoint, the RD design is often directly usable with existing measurement efforts, such as the regularly collected statistical information typical of most management-information systems. The advantages of the RD design warrant greater educational efforts on the part of the methodological community to encourage its use where appropriate.

## 10-2a The Basic RD Design

The basic RD design is a pretest-posttest two-group design. The term *pretest-posttest* implies that the same measure (or perhaps alternative forms of the same measure)

**FIGURE 10–8** **Notation for the regression-discontinuity (RD) design**


is administered before and after some program or treatment. (In fact, the RD design does not require that the pre and post measures be the same.) The term *pretest* implies that the same measure is given twice, whereas the term *pre-program* measure implies more broadly that before and after measures may be the same or different. It is assumed that a cutoff value on the pretest or pre-program measure is being used to assign persons or other units to the program. Two-group versions of the RD design might imply either that some treatment or program is being contrasted with a no-program condition or that two alternative programs are being compared. The description of the basic design as a two-group design implies that a single pretest-cutoff score is used to assign participants to either the program, or comparison group. The term *participants* refers to the units assigned. In many cases, participants are individuals, but they could be any definable units such as hospital wards, hospitals, counties, and so on. The term *program* is used in this discussion of the RD design to refer to any program, treatment, or manipulation whose effects you want to examine. In notational form, the basic RD design might be depicted as shown Figure 10–8:

- *C* indicates that groups are assigned by means of a cutoff score.
- An *O* stands for the administration of a measure to a group.
- An *X* depicts the implementation of a program.
- Each group is described on a single line (for example, program group on top and control group on the bottom).

To make this initial presentation more concrete, imagine a hypothetical study examining the effect of a new treatment protocol for inpatients with a particular diagnosis. For simplicity, assume that you want to try the new protocol on patients who are considered most ill and that for each patient you have a continuous quantitative indicator of health that is a composite rating that takes values from 1 to 100, where high scores indicate greater health. Furthermore, assume that a pretest cutoff score of 50 was (more or less arbitrarily) chosen as the assignment criterion or that all who score lower than 50 on the pretest will be given the new treatment protocol, whereas those who score 50 or higher will be given the standard treatment.

It is useful to begin by considering what the data might look like if you did not administer the treatment protocol but instead measured all participants only at two points in time. Figure 10–9a shows the hypothetical bivariate distribution for this situation. Each item on the figure indicates a single person's pretest and posttest scores. The blue *X*s to the left of the cutoff show the program cases. They are more severely ill on both the pretest and posttest. The green circles show the comparison group that is comparatively healthy on both measures. The vertical line at the pretest score of 50 indicates the cutoff point. (In Figure 10–9a, the assumption is that no treatment has been given.) The solid line through the bivariate distribution is the linear **regression line**. The distribution depicts a strong positive relationship between the pretest and posttest; in general, the more healthy a person is at the pretest, the more healthy he or she is on the posttest, and the more severely ill a person is at the pretest, the more ill that person is on the posttest.

**regression line**
A line that describes the relationship between two or more variables.

Consider what the outcome might look like if the new treatment protocol is administered and has a positive effect (Figure 10–9b). For simplicity, assume that

**FIGURE 10–9a** Pre-post distribution for a RD design with no treatment effect

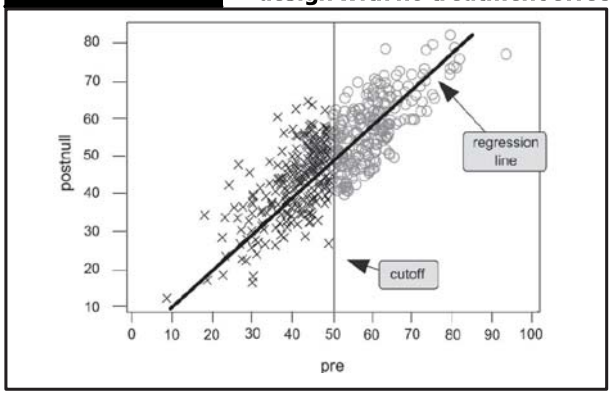



**FIGURE 10–9b** The RD design with 10-point treatment effect

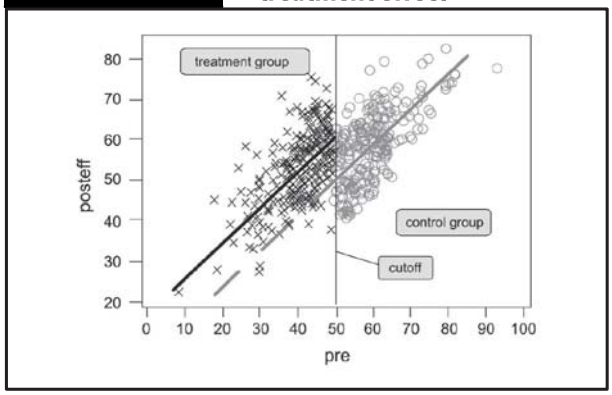



**FIGURE 10–9c** Regression lines for the data shown in Figure 10-9b

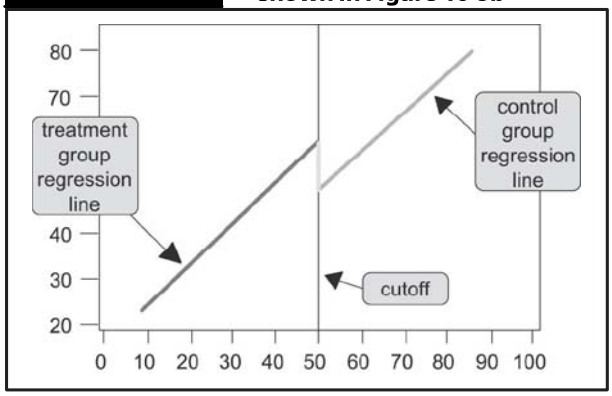
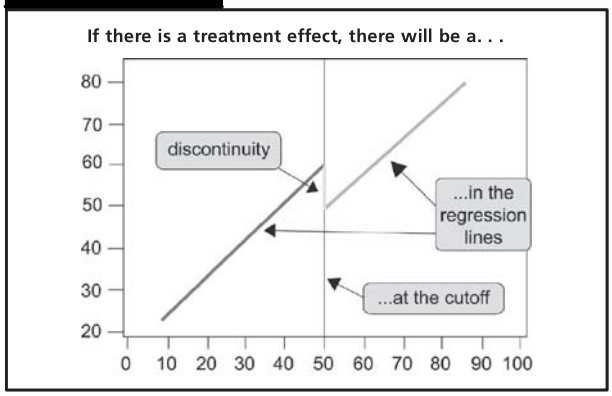



**FIGURE 10–9d** How the RD design got its name




the treatment had a constant effect that raised each treated person's health score by 10 points.

Figure 10–9b is identical to Figure 10–9a except that all points to the left of the cutoff (that is, the treatment group) have been raised by 10 points on the posttest. The dashed line in Figure 10–9b shows what you would expect the treated group's regression line to look like if the program had no effect (as was the case in Figure 10–9a).

It is sometimes difficult to see the forest for the trees in these types of bivariate plots. So, let's remove the individual data points and look only at the regression lines. The plot of regression lines for the treatment effect case of Figure 10–9b is shown in Figure 10–9c.

On the basis of Figure 10–9c, you can now see how the RD design got its name; a program effect is suggested when you observe a *jump* or *discontinuity* in the regression lines at the cutoff point. This is illustrated in Figure 10–9d.

**The Logic of the RD Design** The previous discussion indicates what the key feature of the RD design is: *assignment based on a cutoff value on a pre-program measure.* The cutoff rule for the simple two-group case is essentially as follows:

- All persons on one side of the cutoff are assigned to one group.
- All persons on the other side of the cutoff are assigned to the other group.

The choice of cutoff value is usually based on one of two factors. It can be made solely on the basis of the program resources that are available. For instance, if a

program can handle only 25 persons and 70 people apply, you can choose a cutoff point that distinguishes the 25 most needy persons from the rest. Alternatively, you can choose the cutoff on substantive grounds. If the pre-program assignment measure is an indication of severity of illness measured on a 1 to 7 scale and physicians or other experts believe that all patients scoring 5 or more are critical and fit well the criteria defined for program participants, you might use a cutoff value of 5.

To interpret the results of an RD design, you must know the nature of the assignment variable and the outcome measure, as well as who received the program. Without this information, no distinct outcome pattern directly indicates whether an effect is positive or negative.

To illustrate this, consider a new hypothetical example of an RD design. Assume that a hospital administrator would like to improve the quality of patient care through the institution of an intensive quality-of-care (QOC) training program for staff. Because of financial constraints, the program is too costly to implement for all employees and so instead it will be administered to the entire staff from specifically targeted units or wards that seem most in need of improving quality of care. Two general measures of quality of care are available. The first is an aggregate rating of quality of care based on observation and rating by an administrative staff member and will be labeled here the QOC rating. The second is the ratio of the number of recorded patient complaints relative to the number of patients in the unit over a fixed period of time and will be termed here the *Complaint Ratio.* In this scenario, the administrator could use either the QOC rating or Complaint Ratio as the basis for assigning units to receive the training. Similarly, the effects of the training could be measured on either variable. Figures 10–10a through 10–10d show four outcomes of alternative RD implementations possible under this scenario.

Only the regression lines are shown in Figures 10–10a through 10–10d. It is worth noting that even though all four outcomes have the same pattern of regression lines, they imply very different results. In Figure 10–10a and Figure 10–10b, hospital units were assigned to training because they scored below some cutoff score on the QOC rating. In Figure 10–10c and Figure 10–10d, units received training because they scored above the cutoff score value on the Complaint Ratio measure. In each graph, the dashed line indicates the regression line you would expect to find for the training group if the training had no effect. This dashed line represents the no-discontinuity projection of the comparison-group regression line into the region of the program group pretest scores.

You can clearly see that even though the outcome regression lines are the same in all four groups, you would interpret the four graphs differently. Figure 10–10a depicts a positive effect because training raised the program group's regression line on the QOC rating over what would have been expected. Figure 10–10b, however, shows a negative effect because the program raised training group scores on the Complaint Ratio. indicating increased complaint rates. Figure 10–10c shows a positive effect because the regression line was lowered on the Complaint Ratio relative to what you would have expected. Finally, Figure 10–10d shows a negative effect where the training resulted in lower QOC ratings than you would expect otherwise. The point here is a simple one. A discontinuity in regression lines indicates a program effect in the RD design, but the discontinuity alone is not sufficient to tell you whether the effect is positive or negative. To make this determination, you need to know who received the program and how to interpret the direction of scale values on the outcome measures.

**The Role of the Comparison Group in RD Designs** With this introductory discussion of the design in mind, you can now see what constitutes the benchmark for comparison in the RD design. In experimental or other quasi-experimental designs, you either assume or try to provide evidence that the program and comparison groups are equivalent prior to the program so that post-program differences can be attributed to the manipulation. The RD design involves no such assumption.

**FIGURE 10–10a** Regression lines in hypothetical outcome 1 for an RD design

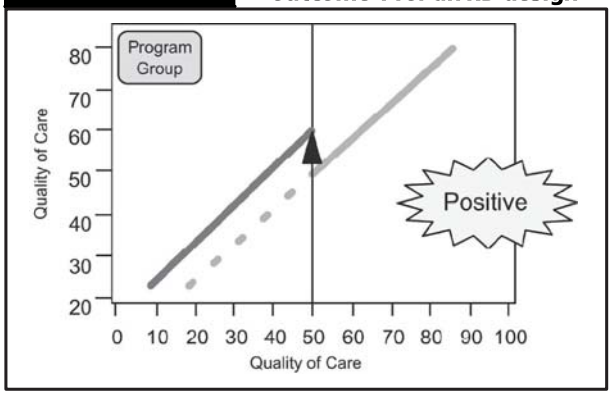



**FIGURE 10–10b** Regression lines in hypothetical outcome 2 for an RD design

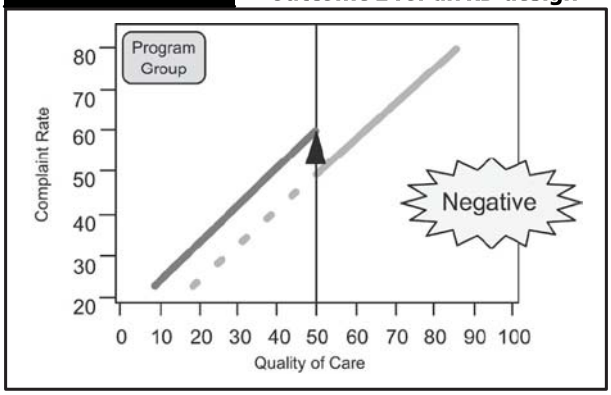



**FIGURE 10–10c** Regression lines in hypothetical outcome 3 for an RD design

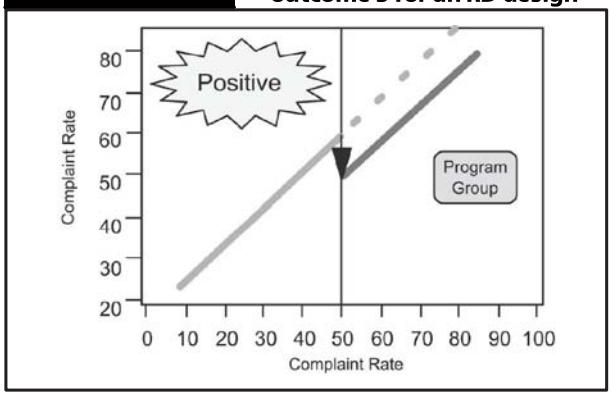
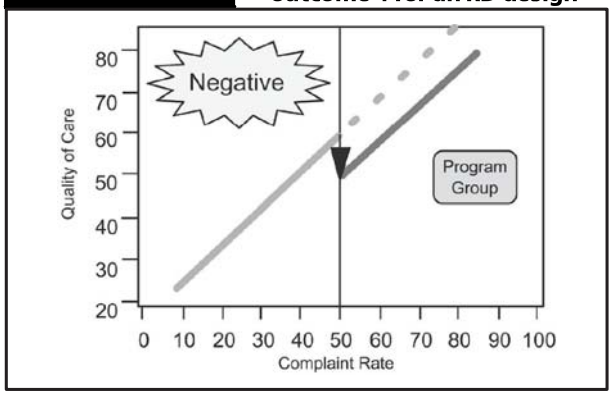



**FIGURE 10–10d** Regression lines in hypothetical outcome 4 for an RD design




Instead, with RD designs you assume that in the absence of the program the pre-post relationship would be equivalent for the two groups. Thus, the strength of the RD design is dependent on two major factors. The first is the assumption that there is no spurious discontinuity in the pre-post relationship that happens to coincide with the cutoff point. The second factor concerns the degree to which you can know and correctly model the pre-post relationship and constitutes the major problem in the statistical analysis of the RD design, which will be discussed in Chapter 12.

## The Internal Validity of the RD Design

*Internal validity* refers to whether one can infer that the treatment or program being investigated caused a change in outcome indicators. Internal validity is not concerned with your ability to generalize but rather focuses on whether a causal relationship can be demonstrated for the immediate research context. Research designs that address causal questions are often compared on their relative ability to yield internally valid results.

In most causal-hypothesis tests, the central inferential question is whether any observed outcome differences between groups are attributable to the program or instead to some other factor. To argue for the internal validity of an inference, the analyst must attempt to demonstrate that the program—and not some plausible alternative explanation—is responsible for the effect. In the literature on internal validity, these plausible alternative explanations or factors are often termed **threats to internal validity** (as described in Chapter 7–1). Many threats can be ruled out by including a control group. Assuming that the control group is equivalent to the program group prior to the study, the control group pre-post gain shows you what

**threats to internal validity**
Any factor that can lead you to reach an incorrect conclusion about whether there is a causal relationship in your study.

would have happened in the program group if it had not had the program. A different rate of gain in the program group provides evidence for the relative effect of the program itself. Thus, randomized-experimental designs are considered strong in internal validity because they give you confidence in the probabilistic, pre-program equivalence between groups that results from random assignment and helps ensure that the control group provides a legitimate reflection of all nonprogram factors that might affect outcomes.

RD designs contain several selection threats to internal validity because of the deliberate pre-program differences between groups. (These are discussed in Chapter 7.) These threats might, at first glance, appear to be a problem. For instance, a selection-maturation threat implies that different rates of maturation between groups explain outcome differences. For the sake of argument, let's consider a pre-post distribution with a linear relationship having a slope equal to two units. This implies that on average, a person with a given pretest score will have a posttest score two times higher. Clearly there is maturation in this situation; that is, people are getting consistently higher scores over time. If a person has a pretest score of 10 units, you would predict a posttest score of 20 for an absolute gain of 10. However, if a person has a pretest score of 50, you would predict a posttest score of 100 for an absolute gain of 50. Thus, the second person naturally gains or matures more in absolute units, although the rate of gain relative to the pretest score is constant. Along these lines, in the RD design, you expect that all participants may mature and that in absolute terms this maturation might be different for the two groups on average. Nevertheless, a program effect in the RD design is not indicated by a difference between the posttest averages of the groups, but rather by a change in the pre-post relationship at the cutoff point. In this example, although you might expect different absolute levels of maturation, a single, continuous regression line with a slope equal to 2 describes these different maturational rates perfectly. More to the point, for selection-maturation to be a threat to internal validity in RD designs, it must induce a discontinuity in the pre-post relationship that happens to coincide with the cutoff point—an unlikely scenario in most studies.

Another selection threat to internal validity that might intuitively seem likely in the RD design concerns the possibility of differential regression to the mean or a selection-regression threat (as described in Chapter 7–1b). The phenomenon of regression to the mean arises when you asymmetrically sample groups from a distribution. On any subsequent measure, the obtained sample group mean will be closer to the population mean for that measure (in standardized units) than the sample mean from the original distribution is to its population mean. In RD designs, you deliberately create asymmetric samples through the cutoff assignment and consequently expect regression towards the mean in both groups. In general, you should expect the low-scoring pretest group to evidence a relative gain on the posttest and the high-scoring pretest group to show a relative loss. As with selection-maturation, even though you expect to see differential regression to the mean, it poses no problem for the internal validity of the RD design. Regression to the mean does not result in a discontinuity in the bivariate relationship coincidental with the cutoff point. In fact, the regression to the mean that occurs is continuous across the range of the pretest scores and is described by the regression line itself. (After all, the term *regression* was originally coined by Galton to refer to the fact that a regression line describes regression to the mean.)

Although the RD design might initially seem susceptible to selection biases, it is not. The previous discussion demonstrates that only factors that would naturally induce a discontinuity in the pre-post relationship could be considered threats to the internal validity of inferences from the RD design. In principle then, the RD design is as strong in internal validity as its randomized experimental alternatives. In practice, however, the validity of the RD design depends directly on how well you can model the true pre-post relationship, certainly a nontrivial statistical problem.

### 10-2b The RD Design and Accountability

It makes sense intuitively that the accountability of a program depends largely on the explicitness of the assignment or allocation of the program to recipients. Lawmakers and administrators need to recognize that programs are more evaluable and accountable when the allocation of the program is more public and verifiable. The three major pre-post designs—the pre-post randomized experiments (Chapter 9), the RD design, and the NEGD design (both discussed earlier in this chapter)—are analogous to the three types of program allocation schemes that legislators or administrators might choose. Randomized experiments are analogous to the use of a lottery for allocating the program. RD designs can be considered explicit, accountable methods for assigning programs based on need or merit. NEGD designs might be considered a type of political allocation because they enable the use of unverifiable, subjective, or politically motivated assignment. Most social programs are politically allocated. Even when programs are allocated primarily based on need or merit, the regulatory agency usually reserves some discretionary capability in deciding who receives the program. Without debating the need for such discretion, it is clear that the methodological community should encourage administrators and legislators who want their programs to be accountable to make explicit their criteria for program eligibility by either using probabilistically based lotteries or by relying on quantitative eligibility ratings and cutoff values as in the RD design. To the extent that legislators and administrators can be convinced to move toward more explicit assignment criteria, both the potential utility of the RD design and the accountability of the programs will be increased.

### 10-2c Statistical Power and the RD Design

The previous discussion argues that the RD design is strong in internal validity, certainly stronger than the NEGD design, and perhaps as strong as the randomized-experiments design, but the RD designs are not as statistically powerful as the randomized experiments (see Chapter 12, for a discussion of statistical power). That is, to achieve the same level of statistical accuracy, an RD design needs as much as 2.75 times the participants as a randomized experiment. For instance, if a randomized experiment needs 100 participants to achieve a certain level of power, the RD design might need as many as 275.

### 10-2d Ethics and the RD Design

So why would you ever use the RD design instead of a randomized one? The real allure of the RD design is that it allows you to assign the treatment or program to those who most need or deserve it. Thus, the real attractiveness of the design is ethical; you don't have to deny the program or treatment to participants who might need it as you do in randomized studies.

## 10-3 Other Quasi-Experimental Designs

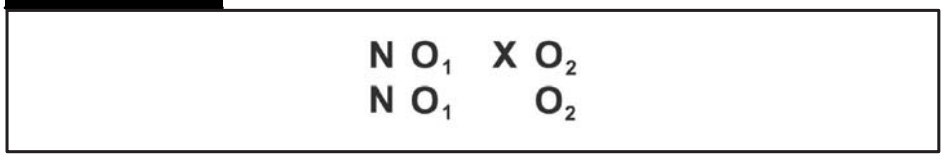
There are many different types of quasi-experimental designs that have a variety of applications in specific contexts. Here, I'll briefly present a number of the more interesting or important quasi-experimental designs. By studying the features of these designs, you can gain a deeper understanding of how to tailor design components to address threats to internal validity in your own research contexts.

**proxy-pretest design**
A post-only design in which, after the fact, a pretest measure is constructed from preexisting data. This is usually done to make up for the fact that the research did not include a true pretest.

### 10-3a The Proxy Pretest Design

The **proxy-pretest design** (Figure 10–11) looks like a standard pre-post design with an important difference. The pretest in this design is collected after the program is

**FIGURE 10–11** **The proxy-pretest design**




given! But how can you call it a pretest if it's collected after the program? Because you use a proxy variable to estimate where the groups would have been on the pretest. There are essentially two variations of this design. In the first, you ask the participants to estimate where their pretest level would have been. This can be called the *recollection proxy-pretest design.* For instance, you might ask participants to complete your measures by estimating how they would have answered the questions 6 months ago. This type of proxy pretest is not good for estimating actual pre-post changes because people may forget where they were at some prior time or they may distort the pretest estimates to make themselves look better. However, at times, you might be interested not so much in where they were on the pretest but rather in where they think they were. The Recollection Proxy Pretest would be a sensible way to assess participants' perceived gain or change.

The other proxy-pretest design uses archived records to stand in for the pretest. This design is called the *archived proxy-pretest design.* For instance, imagine that you are studying the effects of an educational program on the math performance of eighth graders. Unfortunately, you were brought in to do the study after the program had already been started (a too-frequent case, I'm afraid). You are able to construct a posttest that shows math ability after training, but you have no pretest. Under these circumstances, your best bet might be to find a proxy variable that would estimate pretest performance. For instance, you might use the students' grade point average in math from the seventh grade as the proxy pretest.

The proxy-pretest design is not one you should ever select by choice; but if you find yourself in a situation where you have to evaluate a program that has already begun, it may be the best you can do and would almost certainly be better than relying only on a posttest-only design.

## 10-3b The Separate Pre-Post Samples Design

**separate pre-post samples**
A design in which the people who receive the pretest are not the same as the people who take the posttest.

The basic idea in the **separate pre-post samples** design (and its variations) is that the people you use for the pretest are not the same as the people you use for the posttest (Figure 10–12). Take a close look at the design notation for the first variation of this design. There are four groups (indicated by the four lines), but two of the groups come from a single nonequivalent group and the other two also come from a single nonequivalent group (indicated by the subscripts next to N). Imagine that you have two agencies or organizations that you think are similar. You want to implement your study in one agency and use the other as a control. The program you are looking at is an agency-wide one and you expect the outcomes to be most noticeable at the agency level. For instance, let's say the program is designed to improve customer satisfaction. Because customers routinely cycle through your agency, you can't measure the same customers pre-post. Instead, you measure customer satisfaction in each agency at one point in time, implement your program, and then measure customer satisfaction in the agency at another point in time after the program. Notice that the customers will be different within each agency for the pretest and posttest. This design is not a particularly strong one because you cannot match individual participant responses from pre to post; you can only look at the change in average customer satisfaction. Here, you always run the risk that you have

**FIGURE 10–12** The separate pre-post samples design

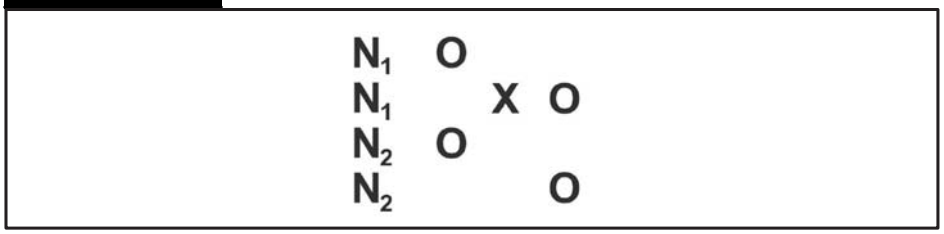
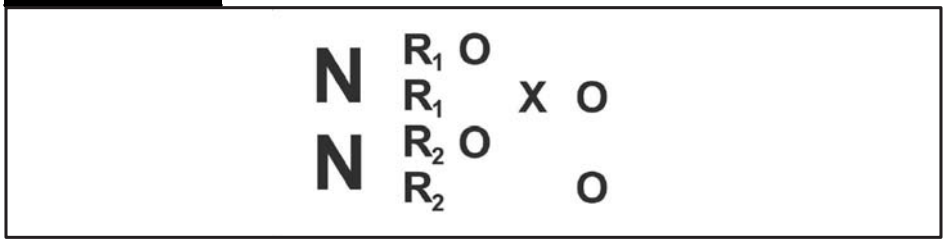



**FIGURE 10–13** The separate pre-post sample design with random sampling




nonequivalence not only between the agencies but within the pre and post groups as well. For instance, if you have different types of clients at different times of the year, this could bias the results. You could also look at this as having a proxy pretest on a different group of people.

The second example of the separate pre-post sample design is shown in design notation in Figure 10–13. Again, four groups are in the study. This time, however, you are taking random samples from your agency or organization at each point in time. This is essentially the same design as the one in Figure 10–12 except for the random sampling. Probably the most sensible use of this design would be in situations where you routinely do sample surveys in an organization or community. For instance, assume that every year, two similar communities do a community-wide survey of residents to ask about satisfaction with city services. Because of costs, you randomly sample each community each year. In one of the communities, you decide to institute a program of community policing and you want to see whether residents feel safer and have changed in their attitudes towards police. You would use the results of last year's survey as the pretest in both communities and this year's results as the posttest. Again, this is not a particularly strong design. Even though you are taking random samples from each community each year, it may still be the case that the community changes fundamentally from one year to the next and that the random samples within a community cannot be considered equivalent.

## 10-3c The Double-Pretest Design

**double-pretest design**
A design that includes two waves of measurement prior to the program.

The **double-pretest design** (Figure 10–14) is a strong quasi-experimental design with respect to internal validity. Why? Recall that the pre-post NEGD is especially susceptible to selection threats to internal validity. In other words, the nonequivalent groups may be different in some way before the program is given and you may incorrectly attribute posttest differences to the program. Although the pretest helps you assess the degree of pre-program similarity, it does not determine whether the groups are changing at similar rates prior to the program. Thus, the NEGD is especially susceptible to selection-maturation threats.

The double-pretest design includes two measures prior to the program. Consequently, if the program and comparison group are maturing at different rates, you should detect this as a change from pretest 1 to pretest 2. Therefore, this design

**FIGURE 10–14** The double-pretest design

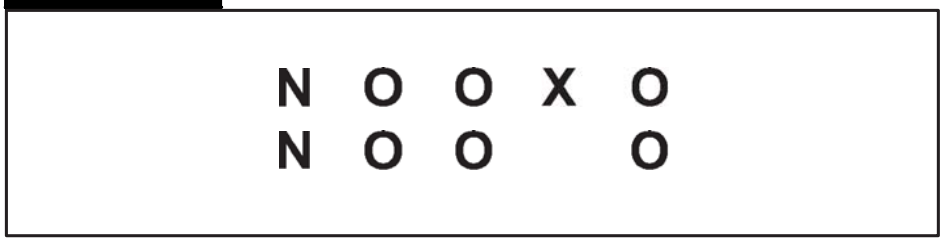
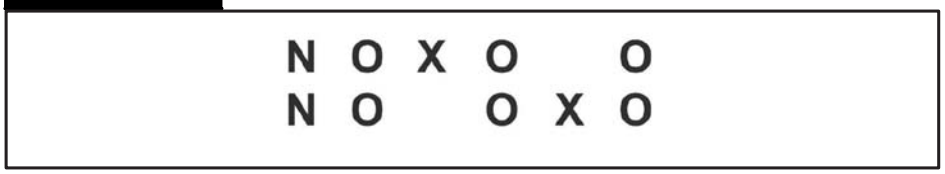



**FIGURE 10–15** The switching-replications design




explicitly controls for selection-maturation threats. The design is also sometimes referred to as a *dry-run, quasi-experimental design* because the double pretests simulate what would happen in the **null case**.

**null case**
A situation in which the treatment has no effect.

## 10-3d The Switching-Replications Design

The switching-replications quasi-experimental design is also strong with respect to internal validity, and because it allows for two independent implementations of the program, it may enhance external validity or generalizability (Figure 10–15). The **switching-replications design** has two groups and three waves of measurement. In the first phase of the design, both groups are given pretests; one is given the program, and both are posttested. In the second phase of the design, the original comparison group is given the program while the original program group serves as the control. This design is identical in structure to its randomized experimental version (described in Chapter 9) but lacks the random assignment to group. It is certainly superior to the simple NEGD. In addition, because it ensures that all participants eventually get the program, it is probably one of the most ethically feasible quasi-experiments.

**switching-replications design**
A two-group design in two phases defined by three waves of measurement. The implementation of the treatment is repeated in both phases. In the repetition of the treatment, the two groups *switch roles*: The original control group in phase 1 becomes the treatment group in phase 2, whereas the original treatment acts as the control. By the end of the study, all participants have received the treatment.

## 10-3e The Nonequivalent Dependent Variables (NEDV) Design

The **nonequivalent dependent variables (NEDV) design** is a deceptive one. In its simple form, it is an extremely weak design with respect to internal validity. However, in its **pattern-matching** variations (covered later in this chapter), it opens the door to an entirely different approach to causal assessment that is extremely powerful. The design notation shown in Figure 10–16 is for the simple two-variable case. Notice that this design has only a single group of participants (indicated by the large N that encompasses both lines in the design). The two lines in the notation indicate separate variables, not separate groups.

The idea in this design is that you have a program designed to change a specific outcome. For instance, assume you are training first-year high-school students in algebra. Your training program is designed to affect algebra scores, but it is not designed to affect geometry scores. You reasonably expect pre-post geometry performance to be affected by other internal validity factors such as history or maturation. In this case, the pre-post geometry performance acts like a control group; it models what would likely have happened to the algebra pre-post scores if the

**nonequivalent dependent variables (NEDV) design**
A single-group pre-post quasi-experimental design with two outcome measures, where only one measure is theoretically predicted to be affected by the treatment and the other is not.

**pattern-matching**
The degree of correspondence between two data items. For instance, you might look at a pattern match of a theoretical expectation pattern with an observed pattern to see if you are getting the outcomes you expect.

**FIGURE 10–16** The NEDV design

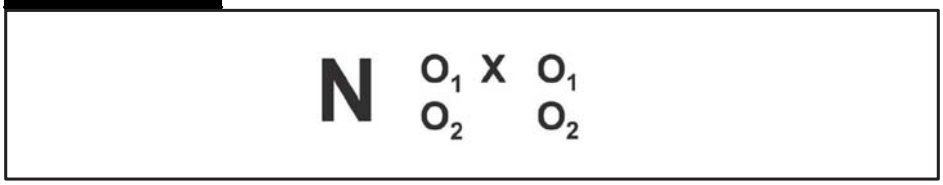
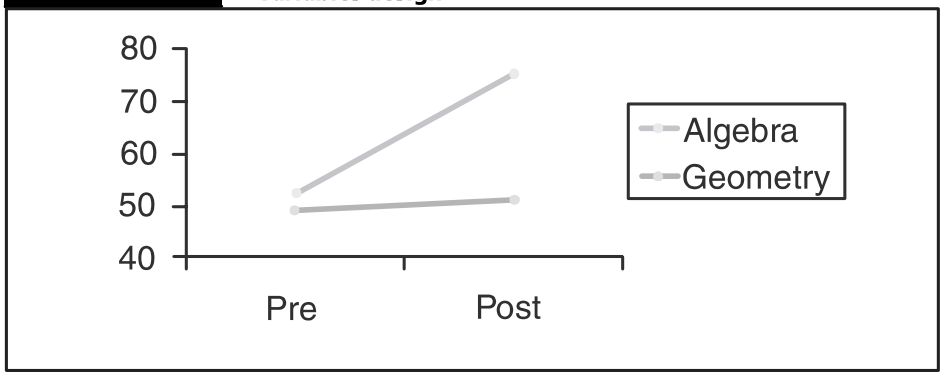



**FIGURE 10–17** Example of a two-variable nonequivalent dependent variables design




program hadn't been given. The key is that the control variable has to be similar enough to the target variable to be affected in the same way by history, maturation, and the other single group internal validity threats, but not so similar that it is affected by the program.

Figure 10–17 shows the results you might get for the two-variable, algebra-geometry example. Note that this design works only if the geometry variable is a reasonable proxy for what would have happened on the algebra scores in the absence of the program. The real allure of this design is the possibility that you don't need a control group; you can give the program to your entire sample. The problem is that in its two-variable simple version, the assumption of the control variable is a difficult one to meet. (Note that a double-pretest version of this design would be considerably stronger.)

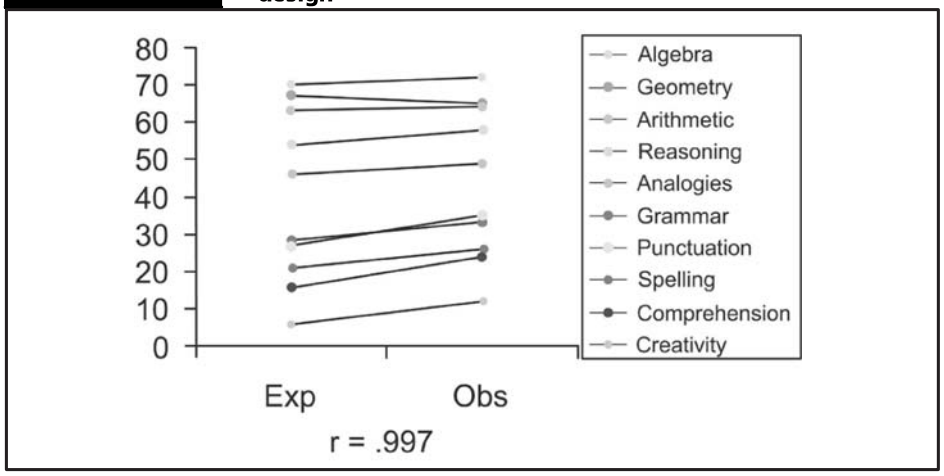
**The Pattern-Matching NEDV Design** Although the two-variable NEDV design is quite weak, you can make it considerably stronger by adding multiple outcome variables. In this variation, you need many outcome variables and a theory that tells *how affected* (from most to least) each variable will be by the program. Let's reconsider the example from the algebra program in the previous discussion. Now, instead of having only an algebra and geometry score, imagine you have ten measures that you collect pre and post. You would expect the algebra measure to be most affected by the program (because that's what the program was most designed to affect). However, in this variation, you recognize that geometry might also be affected because training in algebra might be relevant, at least tangentially, to geometry skills. On the other hand, you might theorize that creativity would be much less affected, even indirectly, by training in algebra and so you predict the creativity measure to be the least affected of the ten measures.

**pattern-matching NEDV design**
A single-group pre-post quasi-experimental design with multiple outcome measures where there is a theoretically specified pattern of expected effects across the measures. To assess the treatment effect, the theoretical pattern of expected outcomes is correlated or matched with the observed pattern of outcomes as measured.

Now, line up your theoretical expectations against your pre-post gains for each variable. You can see in Figure 10–18 that the expected order of outcomes (on the left) is mirrored well in the actual outcomes (on the right).

Depending on the circumstances, the **pattern-matching NEDV design** can be quite strong with respect to internal validity. In general, the design is stronger if

**FIGURE 10–18** **Example of a pattern-matching variation of the NEDV design**




you have a larger set of variables and your expectation pattern matches well with the observed results. What are the threats to internal validity in this design? Only a factor (such as an historical event or maturational pattern) that would yield the same outcome pattern can act as an alternative explanation. Furthermore, the more complex the predicted pattern, the less likely it is that some other factor would yield it. The problem is, the more complex the predicted pattern, the less likely it is that you will find it matches your observed data as well.

The pattern-matching NEDV design is especially attractive for several reasons. It requires you to specify expectations prior to institution of the program. Doing so can be a sobering experience. Researchers often make naive assumptions about how programs or interventions will work. When you're forced to look at the programs in detail, you begin to see that your assumptions may be unrealistic. The design also requires a detailed measurement net—a large set of outcome variables and a detailed sense of how they are related to each other. Developing this level of detail about your measurement constructs is liable to improve the **construct validity** of your study. Increasingly, methodologies can help researchers empirically develop construct networks that describe the expected interrelationships among outcome variables. (See Section 1-4b, Concept Mapping, for more information about how to do this.) Finally, the pattern-matching NEDV is especially intriguing because it suggests that it is possible to assess the effects of programs even if you have only a treated group. Assuming the other conditions for the design are met, control groups are not necessarily needed for causal assessment. Of course, you can also couple the pattern-matching NEDV design with standard experimental or quasi-experimental control group designs for even more enhanced validity. In addition, if your experimental or quasi-experimental design already has many outcome measures as part of the measurement protocol, the design might be considerably enriched by generating variable-level expectations about program outcomes and testing the match statistically.

**construct validity**
The degree to which inferences can legitimately be made from the operationalizations in your study to the theoretical constructs on which those operationalizations are based.

One of my favorite questions to my statistician friends goes to the heart of the potential of the pattern-matching NEDV design. I ask them, "Suppose you have ten outcome variables in a study and that you find that all ten show no statistically significant treatment effects when tested individually (or even when tested as a multivariate set). And suppose, like the desperate graduate students who find in their initial analysis that nothing is significant, that you decide to look at the *direction* of the effects across the ten variables. You line up the variables in terms of which should be most to least affected by your program. And, miracle of miracles, you find that there is a strong and statistically significant correlation between the expected and observed order of effects even though no individual effect was statistically

significant. Is this finding interpretable as a treatment effect?" My answer is "Yes." I think the graduate student's desperation-driven intuition to look at order of effects is a sensible one. I would conclude that the reason you did not find statistical effects on the individual variables is that you didn't have sufficient statistical power. (You can find more about this in the discussion on statistical power in Chapter 12.) Of course, the results will be interpretable as a treatment effect only if you can rule out any other plausible factor that could have caused the ordering of outcomes. Nonetheless, the more detailed the predicted pattern and the stronger the correlation to observed results, the more likely it becomes that the treatment effect is the most plausible explanation. In such cases, the expected pattern of results is like a unique fingerprint, and the observed pattern that matches it can only be due to that unique source pattern.

I believe that the pattern-matching notion implicit in the NEDV design opens the way to an entirely different approach to causal assessment, one that is closely linked to detailed prior explication of the program and to detailed mapping of constructs. It suggests a much richer model for causal assessment than one that relies only on a simplistic dichotomous treatment-control model. In fact, I'm so convinced of the importance of this idea that I've staked a major part of my career on developing pattern-matching models for conducting research!

## 10-3f The Regression Point Displacement (RPD) Design

**regression point displacement (RPD)** A pre-post quasi experimental research design where the treatment is given to only one unit in the sample, with all remaining units acting as controls. This design is particularly useful to study the effects of community-level interventions, where outcome data is routinely collected at the community level.

The **regression point displacement (RPD)** design is a simple quasi-experimental strategy that has important implications, especially for community-based research. The problem with community-level interventions is that it is difficult to do causal assessment to determine whether your program made a difference as opposed to other potential factors. Typically, in community-level interventions, program costs preclude implementation of the program in more than one community. You look at pre-post indicators for the program community and see whether there is a change. If you're relatively enlightened, you seek out another similar community and use it as a comparison. However, because the intervention is at the community level, you have only a single unit of measurement for your program and comparison groups.

The RPD design (Figure 10–19) attempts to enhance the single program unit situation by comparing the performance on that single unit with the performance of a large set of comparison units. In community research, you would compare the pre-post results for the intervention community with a large set of other communities. The advantage of doing this is that you don't rely on a single nonequivalent community; you attempt to use results from a heterogeneous set of nonequivalent communities to model the comparison condition and then compare your single site to this model. For typical community-based research, such an approach may greatly enhance your ability to make causal inferences.

I'll illustrate the RPD design with an example of a community-based AIDS education program to be implemented in one particular community in a state, perhaps a county. Assume that the state routinely publishes annual HIV positive rates by county for the entire state. So, the remaining counties in the state

**FIGURE 10–19** **The regression point displacement (RPD) design**

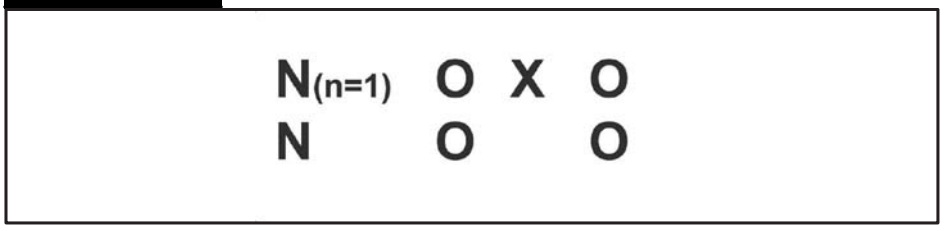


**FIGURE 10–20a** An example of the RPD design

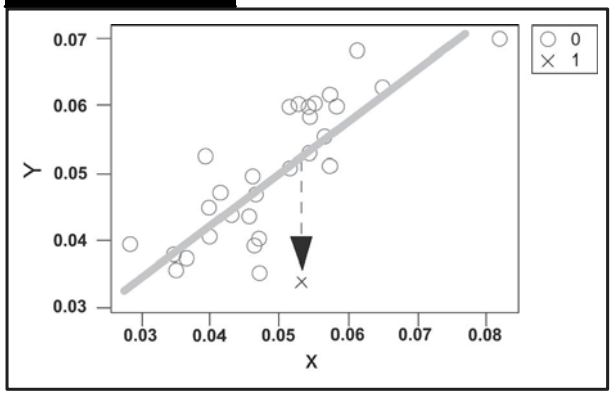
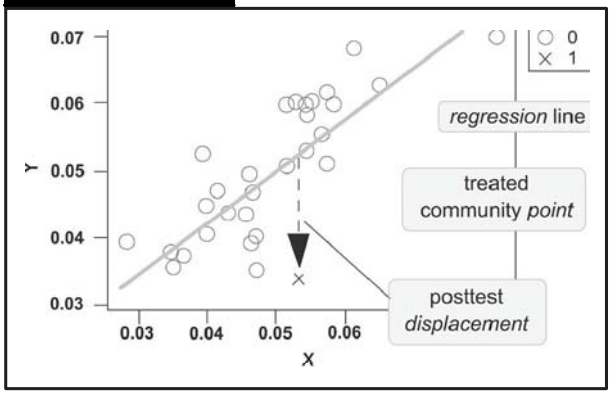



**FIGURE 10–20b** How the RPD design got its name




function as control counties. Instead of averaging all the control counties to obtain a single control score, you use them as separate units in the analysis. Figure 10–20a shows the bivariate pre-post distribution of HIV positive rates per 1000 people for all the counties in the state. The program county—the one that gets the AIDS education program—is shown as an *X* and the remaining control counties are shown as *O*s. You compute a regression line for the control cases (shown in blue on the figure) to model your predicted outcome for a count with any specific pretest rate. To estimate the effect of the program, you test whether the displacement of the program county from the control county regression line is statistically significant.

Figure 10–20b shows why the RPD design was given its name. In this design, you know you have a treatment effect when there is a significant *displacement* of the program *point* from the control group *regression* line.

The RPD design is especially applicable in situations where a treatment or program is applied in a single geographical unit (such as a state, county, city, hospital, or hospital unit) instead of an individual, where many other units are available as control cases, and where there is routine measurement (for example monthly, or annually) of relevant outcome variables.

The analysis of the RPD design turns out to be a variation of the Analysis of Covariance model.

# Summary

This chapter introduced the idea of quasi-experimental designs. These designs look a bit like their randomized or true experimental relatives (described in Chapter 9), but they lack their random assignment to groups. Two major types of quasi-experimental designs were explained in detail. Both are pre-post, two-group designs, and they differ primarily in the manner used to assign the groups. In the NEGD, groups are assigned naturally or are used intact; the researcher does not control the assignment. In RD designs, participants are assigned to groups solely on the basis of a cutoff score on the preprogram measure; the researcher explicitly controls this assignment. Because assignment is explicitly controlled in the RD design and not in NEGD, the former is considered stronger with respect to internal validity, perhaps comparable in strength to randomized experiments. Finally, the versatility and range of quasi-experimental design was illustrated through brief presentation of a number of lesser-known designs that illustrate various combinations of sampling, measurement, or analysis strategies.

Login to the Online Edition of your text at www.atomicdog.com to find additional resources located in the Study Guide at the end of each chapter.