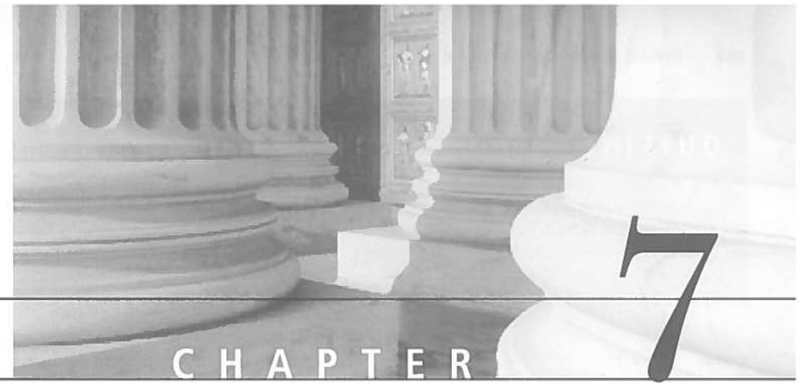


CHAPTER 7

Design

KEY TERMS

causal
causal relationship
compensatory equalization of treatment
compensatory program
compensatory rivalry
construct validity
control group
covariation of the cause and effect
design
diffusion or imitation of treatment
external validity
history threat
instrumentation threat
internal validity
maturation threat
mortality threat
multiple-group threat
null case
plausible alternative explanation
posttest-only nonexperimental design
posttest-only randomized experiment
pre-post nonequivalent groups quasi-experiment
quasi-experimental designs
random selection
regression artifact
regression threat
regression to the mean
resentful demoralization
samples
selection bias
selection threat
selection-history threat
selection-instrumentation
selection-maturation threat
selection-mortality
selection-regression
selection-testing threat
single-group threats
social interaction threats
social threats to internal validity
temporal precedence
testing threat
variable

OUTLINE

design
The design of a study is the specification of how the research question will be answered. A research design should specify how the selection of participants, method of assignment, and choice of measures and time frame work together to accomplish the study objectives.

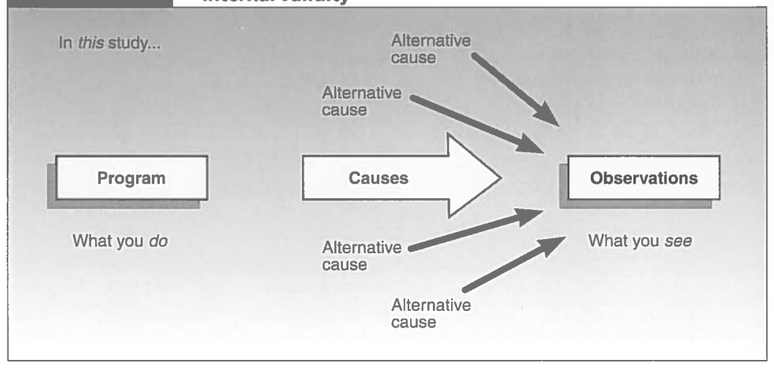
Research **design** provides the glue that holds the research project together. A design is used to structure the research, to show how all of the major parts of the research project—the samples or groups, measures, treatments or programs, and methods of assignment—work together to address the central research questions. In this chapter, after a brief introduction to research design, I'll show you how to classify the major types of designs. You'll see that a major distinction is between the experimental designs that use random assignment to groups or programs and the quasi-experimental designs that don't use random assignment. (People often confuse random selection with the idea of random assignment. You should make sure that you understand the distinction between random selection and random assignment as described in Chapter 9.) Understanding the relationships among designs is important when you need to make design choices, which involves thinking about the strengths and weaknesses of different designs.

7-1 Internal Validity

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

Internal validity is the approximate truth about inferences regarding cause-effect or causal relationships. Thus, internal validity is relevant only in studies that try to establish a causal relationship. It's not relevant in most observational or descriptive studies, for instance. However, for studies that assess the effects of social programs or interventions, internal validity is perhaps the primary consideration. In such contexts, you want to be able to conclude that your program or treatment made a difference—it improved test scores or reduced symptoms, as shown in Figure 7–1. However, there may be reasons, other than your program, that explain why test scores improve or symptoms are reduced. The key question of internal validity is whether observed changes can be attributed to your program or intervention (the cause) and not to other possible causes (sometimes described as alternative explanations for the outcome).

FIGURE 7–1 **A schematic view of the conceptual context for internal validity**



One of the things that's most difficult to grasp about internal validity is that it is relevant only to the specific study in question. That is, you can think of internal validity as a zero-generalizability concern. All that internal validity means is that you have evidence that what you did in the study (for example, the program) caused what you observed (the outcome) to happen. It doesn't tell you whether what you did for the program was what you wanted to do or whether what you observed was what you wanted to observe; those are construct validity concerns (see Chapter 3). It is possible to have internal validity in a study and not have construct validity. For instance, imagine a study in which you are looking at the effects of a new computerized tutoring program on math performance in first-grade students. Imagine that the tutoring is unique in that it has a heavy computer-game component and you think that will really improve math performance. Finally, imagine that you were wrong. (Hard, isn't it?) It turns out that math performance did improve and that it was because of something you did, but it had nothing to do with the computer program. What caused the improvement was the individual attention that the adult tutor gave to the child; the computer program didn't make any difference. This study would have internal validity because something you did affected something that you observed. (You did cause *something* to happen.) The study would not have construct validity because the label "computer-math program" does not accurately describe the actual cause. A more accurate label might be "personal adult attention."

Since the key issue in internal validity is the **causal** one, I'll begin by discussing the conditions that need to be met to establish a **causal relationship** in a research project. Then I'll discuss the different threats to internal validity—the kinds of criticisms your critics will raise when you try to conclude that your program caused the outcome. For convenience, I divide the threats to validity into three categories. The first involves the **single-group threats**—criticisms that apply when you are studying only a single group that receives your program. The second consists of the **multiple-group threats**—criticisms that are likely to be raised when you have several groups in your study (such as a program and a comparison group). Finally, I'll discuss what I call the **social threats to internal validity**—threats that arise because social research is conducted in real-world human contexts where people will react to not only what affects them but also what is happening to others around them.

causal
Pertaining to a cause-effect relationship.

causal relationship
A cause effect relationship. For example, when you evaluate whether your treatment or program causes an outcome to occur, you are examining a causal relationship.

single-group threats
A threat to internal validity that occurs in a study that uses only a single program or treatment group and no comparison or control.

multiple-group threat
An internal validity threat that occurs in studies that use multiple groups—for instance, a program and a comparison group.

social threats to internal validity
Threats to internal validity that arise because social research is conducted in real-world human contexts where people will react to not only what affects them, but also to what is happening to others around them.

temporal precedence
Establishing that the hypothesized cause occurs earlier in time than the effect.

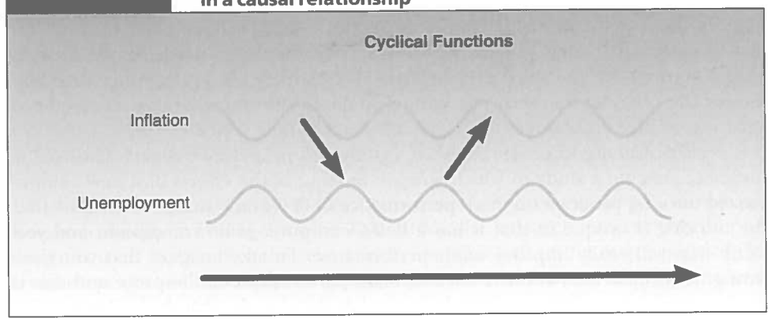
7-1a Establishing Cause and Effect

How do you establish a cause-effect (causal) relationship? What criteria do you have to meet? Generally, you must meet three criteria before you can say that you have evidence for a causal relationship:

- Temporal precedence
- Covariation of the cause and effect
- No plausible alternative explanations

Temporal Precedence To establish **temporal precedence**, you have to show that your cause happened *before* your effect. Sounds easy, huh? Of course, my cause has to happen before the effect. Did you ever hear of an effect happening before its cause? Before you get lost in the logic here, consider a classic example from economics: does inflation cause unemployment? It certainly seems plausible that as inflation increases, more employers find that to meet costs they have to lay off employees. So it seems that inflation could, at least partially, be a cause for unemployment. However, both inflation and employment rates are occurring together on an ongoing basis. Is it possible that fluctuations in employment can affect inflation? If employment in the workforce increases (lower unemployment), there is likely to be more demand for goods, which would tend to drive up the prices (that

FIGURE 7–2 The difficulty in establishing temporal precedence in a causal relationship



is, inflate them), at least until supply can catch up. So which is the cause and which the effect, inflation or unemployment? It turns out that this kind of cyclical situation involves ongoing processes that interact and that both may cause and, in turn, be affected by the other (Figure 7–2). It is hard to establish a causal relationship in this situation.

covariation of the cause and effect
A criterion for establishing a causal relationship that holds that the cause and effect must be related or co-vary.

Covariation of the Cause and Effect What does this mean? Before you can show that you have a causal relationship you have to show that you have some type of relationship. For instance, consider the syllogism:

If *X* then *Y*
If not *X* then not *Y*.

If you observe that whenever *X* is present, *Y* is also present, and whenever *X* is absent, *Y* is too, you have demonstrated that there is a relationship between *X* and *Y*. I don't know about you, but sometimes I find it's not easy to think about *X*'s and *Y*'s. Let's put this same syllogism in program evaluation terms:

If program then outcome
If not program then not outcome.

Or, in colloquial terms: whenever you give the program, you observe the outcome, but when you don't give the program, you don't observe the outcome. This provides evidence that the program and outcome are related. Notice, however, that this syllogism doesn't provide evidence that the program caused the outcome; perhaps some other factor present with the program caused the outcome rather than the program. The relationships described so far are simple binary relationships. Sometimes you want to know whether different amounts of the program lead to different amounts of the outcome—a continuous relationship:

If more of the program then more of the outcome
If less of the program then less of the outcome.

third-variable or missing variable problem
In a two-variable cause-effect relationship, when the effect can be explained by a third variable other than the cause.

plausible alternative explanation
Any other cause that can bring about an effect that is different from your hypothesized or manipulated cause.

No Plausible Alternative Explanations Just because you show there's a relationship doesn't mean it's a causal one. It's possible that some other variable or factor is causing the outcome. This is sometimes referred to as the **third-variable or missing-variable problem**, and it's at the heart of the internal-validity issue. What are some of the possible **plausible alternative explanations?** Later in this chapter, when I discuss the threats to internal validity (see Sections 7-1b through 7-1d), you'll see that each threat describes a type of alternative explanation.

To argue that you have demonstrated internal validity—that you have shown there's a causal relationship—you have to rule out the plausible alternative explanations. How do you do that? One of the major ways is with your research design.

Let's consider a simple single-group threat to internal validity, a history threat which we'll define in Section 7-1b. Let's assume you measure your program group before you begin the program (to establish a baseline), you give the group the program, and then you measure the member's performance afterward in a posttest. You see a marked improvement in the group's performance, which you would like to infer is caused by your program. One of the plausible alternative explanations is that you have a history threat; it's not your program that caused the gain but some other specific historical event. For instance, your antismoking campaign did not cause the reduction in smoking; but rather the Surgeon General's latest report was issued between the time you gave your pretest and posttest. How do you rule this out with your research design? One of the simplest ways would be to incorporate the use of a **control group**—a group, comparable to your program group, that didn't receive the program. However, the group did experience the Surgeon General's latest report. If you find that it didn't show a reduction in smoking even though it experienced the same Surgeon General's report, you have effectively ruled out the Surgeon General's report as a plausible alternative explanation, in this example a history threat.

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

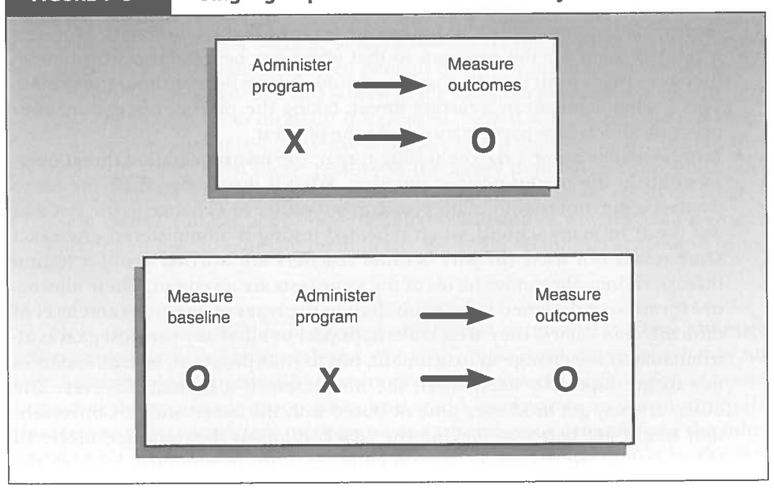
R X$_\phi$ O$_1$
N X$_\phi$ O$_1$

In most applied social research that involves evaluating programs, temporal precedence is not a difficult criterion to meet because you administer the program before you measure effects. Establishing covariation is relatively simple because you have some control over the program and can set things up so you have some people who get it and some who don't (if *X* and if not *X*). Typically, the most difficult criterion to meet is the third—ruling out alternative explanations for the observed effect. That is why research design is such an important issue and why it is intimately linked to the idea of internal validity.

7-1b Single-Group Threats

What is meant by a *single-group threat?* Let's consider two single-group designs and then consider the threats that are most relevant with respect to internal validity. The top design in Figure 7–3 shows a posttest-only single-group design. Here, a group of people receives your program and afterward is given a posttest. In the bottom part of the figure, you see a pretest-posttest, single-group design. In this case,

FIGURE 7–3 Single-group threats to internal validity



the participants receive a pretest or baseline measure, the program or treatment, and then a posttest.

To help make this a bit more concrete, let's imagine that you are studying the effects of a compensatory education program in mathematics for first-grade students on a measure of math performance, such as a standardized math-achievement test. In the post-only design, you would give the first graders the program and then give a math-achievement posttest. You might choose not to give them a baseline measure because you have reason to believe they have no prior knowledge of the math skills you are teaching. It wouldn't make sense to pretest them if you expect them all to get a score of zero. In the pre-post design, you are not willing to assume that your group members have no prior knowledge. You measure the baseline to determine where the students start out in math achievement. You might hypothesize that the change or gain from pretest to posttest is due to your special math-tutoring program. This is a **compensatory program** because it is given only to students who are identified as potentially low in math ability on the basis of some screening mechanism.

compensatory program
A program given to only those who need it on the basis of some screening mechanism.

With either of these scenarios in mind, consider what would happen if you observe a certain level of posttest math achievement or a change or gain from pretest to posttest. You want to conclude that the observed outcome is due to your math program. How could you be wrong? Here are some of the threats to internal validity that your critics might raise, some of the plausible alternative explanations for your observed effect:

- *History threat.* It's not your math program that caused the outcome; it's something else, some historical event that occurred. For instance, lots of first graders watch the public TV program *Sesame Street*, and every *Sesame Street* show presents some elementary math concepts. Perhaps these shows caused the outcome and not your math program. That's a **history threat**.
- *Maturation threat.* The children would have had the exact same outcome even if they had never had your special math-training program. All you are doing is measuring normal maturation or growth in the understanding that occurs as part of growing up; your math program has no effect. How is this maturation explanation different from a history threat? In general, if a specific event or chain of events could cause the outcome, it is a history threat, whereas a **maturation threat** consists of all the events that naturally occur in your life that could cause the outcome (without being specific as to which ones are the active causal agents).
- *Testing threat.* This threat occurs only in the pre-post design. What if taking the pretest made some of the children more aware of that kind of math problem; it primed them for the program so that when you began the math training, they were ready for it in a way that they wouldn't have been without the pretest. This is what is meant by a **testing threat**; taking the pretest, not getting your program affects how participants do on the posttest.
- *Instrumentation threat.* Like the testing threat, the **instrumentation threat** operates only in the pretest-posttest situation. What if the change from pretest to posttest is due not to your math program but rather to a change in the test that was used? In many schools, when repeated testing is administered, the exact same test is not used (in part because teachers are worried about a testing threat); rather, alternative forms of the same tests are given out. These alternative forms were designed to be equivalent in the types of questions and level of difficulty, but what if they aren't? Perhaps part or all of any pre-post gain is attributable to the change in instrument, not to your program. Instrumentation threats are especially likely when the instrument is a human observer. The observers may get tired over time or bored with the observations. Conversely, they might get better at making the observations as they practice more. In either event, the change in instrumentation, not the program, leads to the outcome.

history threat
A threat to internal validity that occurs when some historical event affects your study outcome. hypothesis A specific statement of prediction.

maturation threat
A threat to validity that is a result of natural maturation that occurs between pre- and postmeasurement.

testing threat
A threat to internal validity that occurs when taking the pretest affects how participants do on the posttest.

instrumentation threat
A threat to internal validity that arises when the instruments (or observers) used on the posttest and the pretest differ.

- *Mortality threat.* Mortality doesn't mean that people in your study are dying (although if they are, it would be considered a **mortality threat**). Mortality is used metaphorically here. It means that people are dying with respect to your study. Usually, it means that they are dropping out of the study. What's wrong with that? Let's assume that in your compensatory math-tutoring program you have a nontrivial drop-out rate between pretest and posttest. Assume also that the kids who are dropping out had the low pretest math-achievement test scores. If you look at the average gain from pretest to posttest using all of the scores available to you on each occasion, you would include these low-pretest subsequent dropouts in the pretest and not in the posttest. You'd be dropping out the potential low scorers from the posttest, or you'd be artificially inflating the posttest average over what it would have been if no students had dropped out. You won't necessarily solve this problem by comparing pre-post averages for only those kids who stayed in the study. This subsample would certainly not be representative even of the original entire sample. Furthermore, you know that because of regression threats (see the following section) these students may appear to actually do worse on the posttest, simply as an artifact of the nonrandom dropout or mortality in your study. When mortality is a threat, the researcher can often gauge the degree of the threat by comparing the drop-out group against the non–drop-out group on pretest measures. If there are no major differences, it may be more reasonable to assume that mortality was happening across the entire sample and is not biasing results greatly. However, if the pretest differences are large, you must be concerned about the potential biasing effects of mortality.
- *Regression threat.* A **regression threat**, also known as a **regression artifact** or **regression to the mean**, is a statistical phenomenon that occurs whenever you have a nonrandom sample from a population and two measures that are imperfectly correlated. Okay, I know that's gibberish. Let me try again. Assume that your two measures are a pretest and posttest. You can certainly bet these aren't perfectly correlated with each other. Furthermore, assume that your sample consists of low pretest scorers. The regression threat means that the pretest average for the group in your study will appear to increase or improve (relative to the overall population) even if you don't do anything to them—even if you never give them a treatment. Regression is a confusing threat to understand at first. I like to think about it as the *you can only go up (or down) from here* phenomenon. If you include in your program only the kids who constituted the lowest ten percent of the class on the pretest, what are the chances that they would constitute exactly the lowest ten percent on the posttest? Not likely. Most of them would score low on the posttest, but they aren't likely to be exactly the lowest ten percent twice. For instance, maybe a few kids made a few lucky guesses and scored at the eleventh percentile on the pretest; they might not be so lucky on the posttest. Now if you choose the lowest ten percent on the pretest, they can't get any lower than being the lowest; they can only go up from there, relative to the larger population from which they were selected. This purely statistical phenomenon is what we mean by a regression threat. You can see a more detailed discussion of why regression threats occur and how to estimate them in the following section on regression to the mean.

mortality threat
A threat to validity that occurs because a significant number of participants drop out.

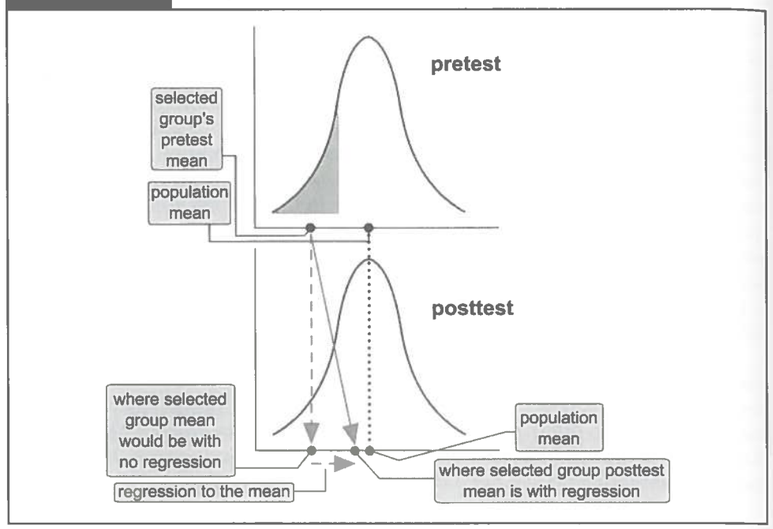
regression threat
A statistical phenomenon that causes a group's average performance on one measure to regress toward or appear closer to the mean of that measure than anticipated or predicted. Regression occurs whenever you have a nonrandom sample from a population and two measures that are imperfectly correlated. A regression threat will bias your estimate of the group's posttest performance and can lead to incorrect causal inferences.

regression artifact
See regression threat.

regression to the mean
See regression threat.

How do you deal with these single-group threats to internal validity? Although you can rule out threats in several ways, one of the most common approaches to ruling those discussed previously is through your research design. For instance, instead of doing a single-group study, you could incorporate a control group. In this scenario, you would have two groups: one receives your program and the other one doesn't. In fact, the only difference between these groups should be the program. If that's true, the control group would experience all the same history and

FIGURE 7–4 **Regression to the mean**



maturation threats, have the same testing and instrumentation issues, and have similar rates of mortality and regression to the mean. In other words, a good control group is one of the most effective ways to rule out the single-group threats to internal validity. Of course, when you add a control group, you no longer have a single-group design, and you still have to deal with two major types of threats to internal validity: the multiple-group threats to internal validity and the social threats to internal validity (see the respective sections later in this chapter).

Regression to the Mean A regression threat, also known as a *regression artifact* or *regression to the mean*, is a statistical phenomenon that occurs whenever you have a nonrandom sample from a population and two measures that are imperfectly correlated.

Figure 7–4 shows the regression to the mean phenomenon. The top part of the figure shows the pretest distribution for a population. Pretest scores are distributed normally; the frequency distribution looks like a bell-shaped curve. Assume that the sample for your study was selected exclusively from the low pretest scorers. You can see on the top part of the figure where their pretest mean is; clearly, it is considerably below the population average. What would you predict the posttest to look like? First, assume that your program or treatment doesn't work at all (the **null case**). A naive assumption would be that your sample would score as badly on the posttest as on the pretest; but they don't! The bottom of the figure shows where the sample's posttest mean would have been without regression and where it actually is. In actuality, the sample's posttest mean wound up closer to the posttest population mean than the pretest mean was to the pretest population mean. In other words, the sample's mean appears to *regress toward the mean* of the population from pretest to posttest.

null case
The case where the null hypothesis appears to be correct. In a two group design, for example, the null case is the finding that there is no difference between the two groups.

Why Does Regression to the Mean Happen? To see why regression to the mean happens, consider a concrete case. In your study, you select the lowest 10 percent of the population based on pretest scores. What are the chances that on the posttest that exact group will once again constitute the lowest 10 percent of the population? Slim to none. Most of them will probably be in the lowest ten percent on the posttest, but if even only a few are not, the group's mean will have to be

proportionally closer to the population's posttest than it was to the pretest. The same thing is true on the other end. If you select as your sample the highest 10 percent pretest scorers, they aren't likely to be the highest ten percent on the posttest (even though most of them may be in the top 10 percent). If even a few score below the top 10 percent on the posttest, the group's posttest mean will have to be proportionally closer to the population posttest mean than to its pretest mean.

Regression to the mean can be very hard to grasp. It even causes experienced researchers difficulty in its more advanced variations. To help you understand what regression to the mean is, and how it can be described, I've listed a few statements you should memorize about the regression to the mean phenomenon (and I provide a short explanation for each):

- *Regression to the mean is a statistical phenomenon.* Regression to the mean occurs for two reasons. First, it results because you asymmetrically sampled from the population. If you randomly sample from the population, you would observe (subject to random error) that the population and your sample have the same pretest average. Because a random sample is already at the population mean on the pretest, it is impossible for it to regress toward the mean of the population any more.
- *Regression to the mean is a group phenomena.* You cannot tell which way an individual's score will move based on the regression to the mean phenomenon. Even though the group's average will move toward the population's, some individuals in the group are likely to move in the other direction.
- *Regression to the mean happens between any two variables.* Here's a common research mistake. You run a program and don't find any overall group effect. So, you decide to look at those who did best on the posttest (your success stories) and see how much they gained over the pretest. You are selecting a group that is extremely high on the posttest. The group members are unlikely to be the best on the pretest as well (although many of them will be). So, the group's pretest mean *has* to be closer to the population mean than its posttest one. You describe this nice gain and are almost ready to write up your results when someone suggests you look at your failure cases (the people who scored worst on your posttest). When you check on how they scored on the pretest, you find that they weren't the worst scorers there. If they had been the worst scorers both times, you would have simply said that your program didn't have any effect on them. But now it looks worse than that; it looks like your program actually made them worse relative to the population! What will you do? How will you ever get your grant renewed? Or your paper published? Or, heaven help you, how will you ever get tenured?

  What you have to realize is that the pattern of results I just described happens every time you measure two measures. It happens forward in time (from pretest to posttest). It happens backward in time (from posttest to pretest). It happens across measures collected at the same time (height and weight)! It will happen even if you don't give your program or treatment.
- *Regression to the mean is a relative phenomenon.* Regression to the mean has nothing to do with overall maturational trends. Notice in Figure 7–4 that I didn't bother labeling the *x*-axis in either the pretest or posttest distribution. It could be that everyone in the population gains 20 points (on average) between the pretest and the posttest. But regression to the mean would still be operating, even in that case. That is, the low scorers would, on average, gain more than the population gain of 20 points (and thus their mean would be closer to the population's).
- *You can have regression up or down.* If your sample consists of below-population-mean scorers, the regression to the mean will make it appear that they move *up* on the other measure. However, if your sample consists of high scorers, the mean will appear to move *down* relative to the population. (Note that even if

the mean increases, the group could lose ground to the population. So, if a high-pretest-scoring sample gains 5 points on the posttest while the overall sample gains 15, you could suspect regression to the mean as an alternative explanation [to our program] for that relatively low change.)

- *The more extreme the sample group, the greater the regression to the mean.* If your sample differs from the population by only a little bit on the first measure, there won't be much regression to the mean because there isn't much room for regression; the group is already near the population mean. So, if you have a sample, even a nonrandom one, that is a good subsample of the population, regression to the mean will be inconsequential (although it will be present). However, if your sample is extreme relative to the population (for example, the lowest or highest 10 percent), the group's mean is further from the population's and has more room to regress.
- *The less correlated the two variables, the greater the regression to the mean.* The other major factor that affects the amount of regression to the mean is the correlation between the two variables. If the two variables are *perfectly* correlated, the highest scorer on one is the highest on the other, next highest on one is next highest on the other, and so on. No regression to the mean occurs. However, this is unlikely to ever happen in practice. Measurement theory demonstrates that there is no such thing as perfect measurement; all measurement is assumed (under the true score model, as discussed in Chapter 3) to have some random error in measurement. It is only when the measure has no random error—is perfectly reliable—that you can expect it to correlate perfectly. Since that doesn't happen in the real world, you have to assume that measures have some degree of unreliability, that relationships between measures will not be perfect, and that there will appear to be regression to the mean between these two measures, given asymmetrically sampled subgroups.

The Formula for the Percent of Regression to the Mean You can estimate exactly the percent of regression to the mean in any given situation with the following formula:

$$P_{rm} = 100(1 - r)$$

where:

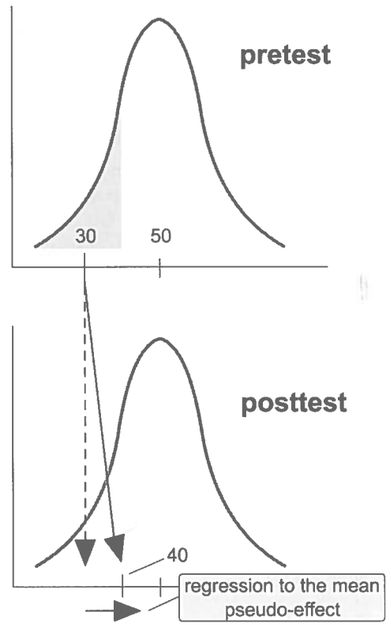
P_{rm} = the percent of regression to the mean
r = the correlation between the two measures

Consider the following four cases:

- If $r = 1$, there is no (0%) regression to the mean.
- If $r = .5$, there is 50% regression to the mean.
- If $r = .2$, there is 80% regression to the mean.
- If $r = 0$, there is 100% regression to the mean.

In the first case, the two variables are perfectly correlated and there is no regression to the mean. With a correlation of .5, the sampled group moves *50 percent* of the distance from the no-regression point to the mean of the population. If the correlation is a small .20, the sample will regress 80 percent of the distance. If no correlation exists between the measures, the sample regresses all the way back to the population mean! It's worth thinking about what this last case means. With zero correlation, knowing a score on one measure gives you absolutely no information about the likely score for that person on the other measure. In that case, your best guess for how any person would perform on the second measure will be the mean of that second measure.

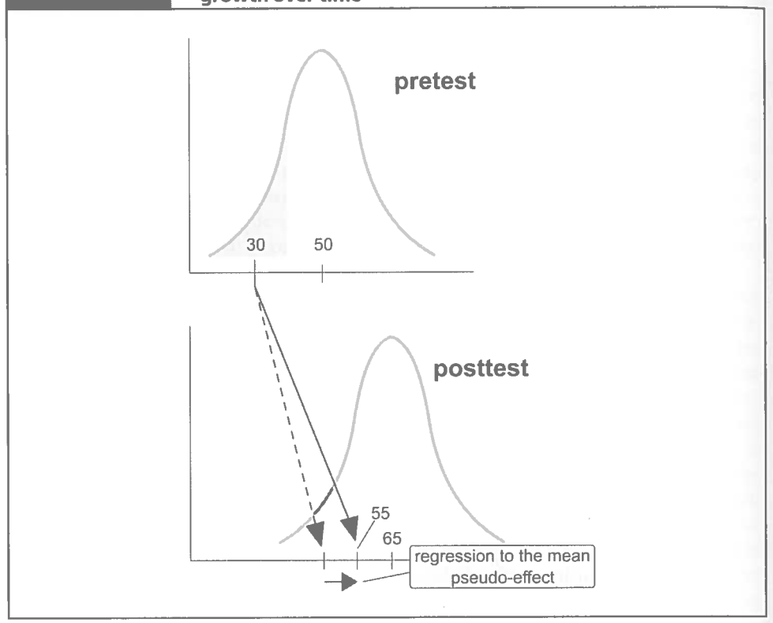
FIGURE 7–5 **An example of regression to the mean when the correlation between measures is .5**



Estimating and Correcting Regression to the Mean Given the percentage formula, for any given situation you can estimate the regression to the mean. All you need to know is the mean of the sample on the first measure, the population mean on both measures, and the correlation between measures. Consider a simple example as shown in Figure 7–5. Here, assume that the pretest population mean is 50 and that you selected a low-pretest scoring sample that has a mean of 30. To begin with, also assume that you do not give any program or treatment (the null case) and that the population is not changing over time on the characteristic being measured (steady-state). Given this, you would predict that the population mean would be 50 and that the sample would get a posttest score of 30 *if there was no regression to the mean.* Now, assume that the correlation is .50 between the pretest and posttest for the population. Given the formula, you would expect the sampled group to regress 50 percent of the distance from the no-regression point to the population mean, or 50 percent of the way from 30 to 50. In this case, you would observe a score of 40 for the sampled group, which would constitute a 10-point pseudo-effect or regression artifact.

Now, relax some of the initial assumptions. For instance, as illustrated in Figure 7–6, assume that between the pretest and posttest the population gained 15 points on average (and that this gain was uniform across the entire distribution; that is, the variance of the population stayed the same across the two measurement occasions). In this case, a sample that had a pretest mean of 30 would be expected to reach a posttest mean of 45 (30 + 15) if there is no regression to the mean ($r = 1$). But here, the correlation between pretest and posttest is .5, so you would expect to see regression to the mean that covers 50 percent of the distance from the mean of 45 to the population posttest mean of 65. That is, you would observe a posttest average of 55 for your sample, again a pseudo-effect of 10 points.

FIGURE 7–6 An example of regression to the mean when there is also growth over time



Regression to the mean is one of the trickiest threats to validity. It is subtle in its effects, and even excellent researchers sometimes fail to catch a potential regression artifact. You might want to learn more about the regression to the mean phenomenon. One good way to do that would be to simulate the phenomenon. If you already understand the basic idea of simulation, you can do a manual (dice rolling) simulation of regression artifacts or a computerized simulation of regression artifacts.

7-1c Multiple-Group Threats

A multiple-group design typically involves at least two groups and before-after measurements. Most often, one group receives the program or treatment while the other does not and constitutes the control or comparison group. However, sometimes one group gets the program and the other gets either the standard program or another program you would like to compare. In this case, you would be comparing two programs for their relative outcomes. Typically, you would construct a multiple-group design so that you could compare the groups directly. In such designs, the key internal validity issue is the degree to which the groups are comparable before the study. If they are comparable and the only difference between them is the program, posttest differences can be attributed to the program; but that's a big *if*. If the groups aren't comparable to begin with, you won't know how much of the outcome to attribute to your program or to the initial differences between groups.

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

selection threat
See selection bias.

There really is only one multiple-group threat to internal validity: that the groups were not comparable before the study. This threat is called a **selection bias** or **selection threat**. A selection threat is *any* factor other than the program that leads to posttest differences between groups. Whenever you suspect that outcomes differ between groups not because of your program but because of prior group differences, you are suspecting a selection bias. Although the term *selection bias* is used as the general category for all prior differences, when you know specifically what

the group difference is, you usually hyphenate it with the selection term. The multiple-group selection threats directly parallel the single-group threats. For instance, whereas history is a single-group threat, selection-history is its multiple-group analogue.

As with the single-group threats to internal validity, I'll provide simple examples involving a new compensatory mathematics-tutoring program for first graders. The design is a pretest-posttest design that divides the first graders into two groups: one receiving the new tutoring program and the other not receiving it. Here are the major multiple-group threats to internal validity for this case:

- *Selection-history threat.* A **selection-history threat** is any other event that occurs between pretest and posttest that the groups experience differently. Because this is a selection threat, the groups differ in some way. Because it's a history threat, the way the groups differ is with respect to their reactions to history events. For example, what if the television-viewing habits of the children in one group differ from those of the children in the other group? Perhaps the program-group children watch *Sesame Street* more frequently than those in the control group do. Since *Sesame Street* is a children's show that presents simple mathematical concepts in interesting ways, it may be that a higher average posttest math score for the program group doesn't indicate the effect of your math tutoring; it's really an effect of the two groups experiencing a relevant event differentially—in this case *Sesame Street*—between the pretest and posttest.
- *Selection-maturation threat.* A **selection-maturation threat** results from differential rates of normal growth between pretest and posttest for the groups. In this case, the two groups are different in their rates of maturation with respect to math concepts. It's important to distinguish between history and maturation threats. In general, *history* refers to a discrete event or series of events, whereas *maturation* implies the normal, ongoing developmental process that takes place. In any case, if the groups are maturing at different rates with respect to the outcome, you cannot assume that posttest differences are due to your program; they may be selection-maturation effects.
- *Selection-testing threat.* A **selection-testing threat** occurs when a *differential* effect of taking the pretest exists between groups on the posttest. Perhaps the test primed the children in each group differently or they may have learned differentially from the pretest. In these cases, an observed posttest difference can't be attributed to the program. It could be the result of selection-testing.
- *Selection-instrumentation threat.* **Selection-instrumentation** refers to any differential change in the test used for each group from pretest to posttest. In other words, the test changes differently for the two groups. Perhaps the test consists of observers, who rate the class performance of the children. What if the program group observers, for example, become better at doing the observations while, over time, the comparison group observers become fatigued and bored. Differences on the posttest could easily be due to this differential instrumentation—selection-instrumentation—and not to the program.
- *Selection-mortality threat.* **Selection-mortality** arises when there is differential nonrandom dropout between pretest and posttest. In our example, different types of children might drop out of each group, or more may drop out of one than the other. Posttest differences might then be due to the different types of dropouts—the selection-mortality—and not to the program.
- *Selection-regression threat.* Finally, **selection-regression** occurs when there are different rates of regression to the mean in the two groups. This might happen if one group is more extreme on the pretest than the other. In the context of our example, it may be that the program group is getting a disproportionate number of children with low math ability because teachers think they need the math tutoring more (and the teachers don't understand the need for comparable program and comparison groups). Because the tutoring group has

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

selection-testing threat
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 pretest and posttest.

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

the lower scorers, its mean regresses a greater distance toward the overall population mean and its group members appear to gain more than their comparison-group counterparts. This is not a real program gain; it's a selection-regression artifact.

When you move from a single group to a multiple group study, what do you gain from the rather significant investment in a second group? If the second group is a control group and is comparable to the program group, you can rule out the single-group threats to internal validity because those threats will all be reflected in the comparison group and cannot explain why posttest group differences would occur. But the key is that the groups must be comparable. How can you possibly hope to create two groups that are truly comparable? The best way to do that is to randomly assign persons in your sample into the two groups—you conduct a randomized or true experiment (see the discussion of experimental designs in Chapter 9).

However, in many applied research settings you can't randomly assign, either because of logistical or ethical factors. In those cases, you typically try to assign two groups nonrandomly so that they are as equivalent as you can make them. You might, for instance, have one classroom of first graders assigned to the math-tutoring program and the other class assigned to the comparison group. In this case, you would hope the two are equivalent, and you may even have reasons to believe that they are. Nonetheless, they may not be equivalent, and because you did not use a procedure like random assignment to at least ensure that they are probabilistically equivalent, you have to take extra care to look for preexisting differences and adjust for them in the analysis. If you measure the groups on a pretest, you can examine whether they appear to be similar on key measures before the study begins and make some judgment about the plausibility that a selection bias exists. There are also ways to adjust statistically for preexisting differences between groups if they are present, although these procedures are notoriously assumption-laden and fairly complex. Research designs that look like randomized or true experiments (they have multiple groups and pre-post measurement) but use nonrandom assignment to choose the groups are called *quasi-experimental designs* (see the discussion of quasi-experimental designs in Chapter 10).

Even if you move to a multiple-group design and have confidence that your groups are comparable, you cannot assume that you have strong internal validity. A number of social threats to internal validity arise from the human interaction in applied social research and you will need to address them.

7-1d Social Interaction Threats

social interaction threats
Threats to internal validity that arise because social research is conducted in real-world human contexts where people react to not only what affects them, but also to what is happening to others around them.

external validity
The degree to which the conclusions in your study would hold for other persons in other places and at other times.

diffusion or imitation of treatment
A social threat to internal validity that occurs because a comparison group learns about the program either directly or indirectly from program group participants.

Applied social research is a human activity. The results of such research are affected by the human interactions involved. The **social interaction threats** to internal validity refer to the social pressures in the research context that can lead to posttest differences not directly caused by the treatment itself. Most of these threats occur because the various groups (for example, program and comparison), or key people involved in carrying out the research (such as managers, administrators, teachers, and principals), are aware of each other's existence and of the role they play in the research project or are in contact with one another. Many of these threats can be minimized by *isolating the two groups from each other*, but this leads to other problems. For example, it's hard to randomly assign and then isolate; this is likely to reduce generalizability or **external validity** (see external validity in Chapter 2). Here are the major social interaction threats to internal validity:

- *Diffusion or imitation of treatment.* **Diffusion or imitation of treatment** occurs when a comparison group learns about the program either directly or indirectly from program group participants. In a school context, children from different groups within the same school might share experiences during lunch hour. Or, comparison group students, seeing what the program group is getting, might set up their own experience to try to imitate that of the program group. In either case, if the diffusion or imitation affects the posttest performance of the comparison group, it can jeopardize your ability to assess whether your program is causing the outcome. Notice that this threat to validity tends to equalize the outcomes between groups, minimizing the chance of seeing a program effect even if there is one.

compensatory rivalry
A social threat to internal validity that occurs when one group knows the program another group is getting and, because of that, develops a competitive attitude with the other group.

- *Compensatory rivalry.* In the **compensatory rivalry** case, the comparison group knows what the program group is getting and develops a competitive attitude with the program group. The students in the comparison group might see the special math-tutoring program the other group is getting and feel jealous. This could lead them to compete with the program group "just to show" how well they can do. Sometimes, in contexts like these, the participants are even encouraged by well-meaning teachers or administrators to compete with one another. (Although this might make educational sense as a motivation for the students in both groups to work harder, it works against the ability of researchers to see the effects of their program.) If the rivalry between groups affects posttest performance, it could make it more difficult to detect the effects of the program. As with diffusion and imitation, this threat generally equalizes the posttest performance across groups, increasing the chance that you won't see a program effect, even if the program is effective.

resentful demoralization
A social threat to internal validity that occurs when the comparison group knows what the program group is getting and becomes discouraged or angry and gives up.

- *Resentful demoralization.* **Resentful demoralization** is almost the opposite of compensatory rivalry. Here, students in the comparison group know what the program group is getting and instead of developing a rivalry, the group members become discouraged or angry and give up (sometimes referred to informally as the *screw-you effect*). Or, if the program group is assigned to an especially difficult or uncomfortable condition, they can rebel in the form of resentful demoralization. Unlike the previous two threats, this one is likely to exaggerate posttest differences between groups, making your program look even more effective than it actually is.

compensatory equalization of treatment
A social threat to internal validity that occurs when the control group is given a program or treatment (usually, by a well-meaning third party) designed to make up for or "compensate" for the treatment the program group gets.

- *Compensatory equalization of treatment.* **Compensatory equalization of treatment** is the only threat of the four that primarily involves the people who help manage the research context rather than the participants themselves. When program and comparison-group participants are aware of one another's conditions, they might wish they were in the other group (depending on the perceived desirability of the program, it could work either way). In our education example, they or their parents or teachers might pressure the administrators to have them reassigned to the other group. The administrators may begin to feel that the allocation of goods to the groups is not fair and may compensate one group for the perceived advantage of the other. If the special math-tutoring program were being done with state-of-the-art computers, you can bet that the parents of the children assigned to the traditional noncomputerized comparison group will pressure the principal to equalize the situation. Perhaps the principal will give the comparison group some other good or grant access to the computers for other subjects. If these compensating programs equalize the groups on posttest performance, they will tend to work against your detecting an effective program even when it does work. For instance, a compensatory program might improve the self-esteem of the comparison group and eliminate your chances of discovering whether the math program would cause changes in self-esteem relative to traditional math training.

As long as people engage in applied social research, you have to deal with the realities of human interaction and its effect on the research process. The threats described here can often be minimized by constructing multiple groups that are unaware of each other (for example, a program group from one school and a comparison group from another) or by training administrators in the importance of

preserving group membership and not instituting equalizing programs. However, researchers will never be able to eliminate entirely the possibility that human interactions are making it more difficult to assess cause-effect relationships.

7-2 Introduction to Design

Research design can be thought of as the *structure* of research; the research design tells you how all the elements in a research project fit together. Researchers often use notational systems to describe a design, which enables them to summarize a complex design structure efficiently. A design includes the following elements:

- *Observations or measures.* These are symbolized by an *O* in design notation. An *O* can refer to a single measure (a measure of body weight), a single instrument with multiple items (a ten-item, self-esteem scale), a complex multipart instrument (a survey), or a whole battery of tests or measures given out on one occasion. If you need to distinguish among specific measures, you can use subscripts with the *O*, as in O_1, O_2, and so on.
- *Treatments or programs.* These are symbolized with an *X* in design notation. The *X* can refer to a simple intervention (such as a one-time surgical technique) or to a complex hodgepodge program (such as an employment-training program). Usually, a no-treatment control or comparison group has no symbol for the treatment. (Although some researchers use *X*+ and *X*– to indicate the treatment and control, respectively.) As with observations, you can use subscripts to distinguish different programs or program variations.
- *Groups.* Each group in a design is given its own line in the design structure. For instance, if the design notation has three lines, the design contains three groups.
- *Assignment to group.* Assignment to group is designated by a letter at the beginning of each line (or group) that describes how the group was assigned. The major types of assignment are:

  R = random assignment
  N = nonequivalent groups
  C = assignment by cutoff

  Don't worry at this point if you don't know what some of these are; each of these assignment strategies characterizes a different type of design and will be described later when discussing that design type.

- *Time.* Time moves from left to right. Elements that are listed on the left occur before elements that are listed on the right.

It's always easier to explain design notation through examples than it is to describe it in words. Figure 7–7 shows the design notation for a pretest-posttest (or before-after) treatment versus comparison-group randomized experimental design. Let's go through each of the parts of the design. There are two lines in the notation, so you should realize that the study has two groups. There are four *O*s in the notation: two on each line and two for each group. When the *O*s are stacked vertically on top of each other, it means they are collected at the same time. In the notation, the two *O*s taken before (to the left of) the treatment are the pretest. The two *O*s taken after the treatment is given are the posttest. The *R* at the beginning of each line signifies that the two groups are randomly assigned (making it an experimental design as described in Chapter 9).

The design is a treatment-versus-comparison-group one, because the top line (treatment group) has an *X*, whereas the bottom line (control group) does not. You should be able to see why many of my students call this type of notation the tic-tac-toe method of design notation; there are lots of *X*s and *O*s! Sometimes you have to use more than simply the *O*s or *X*s. Figure 7–8 shows the identical research

FIGURE 7–7 **A detailed example of design notation**

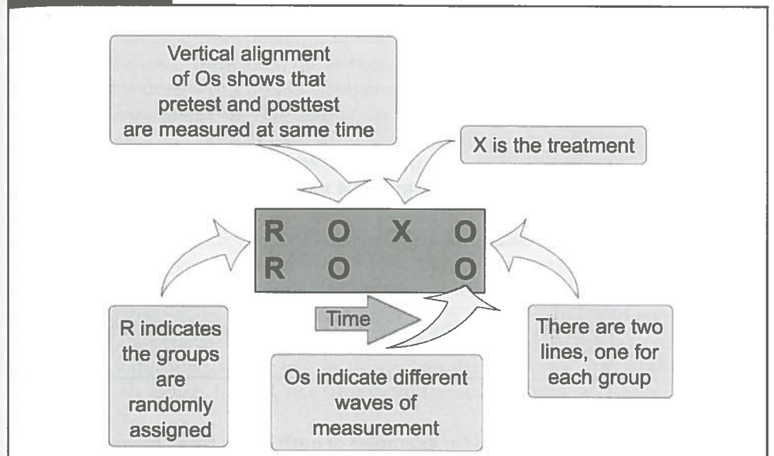
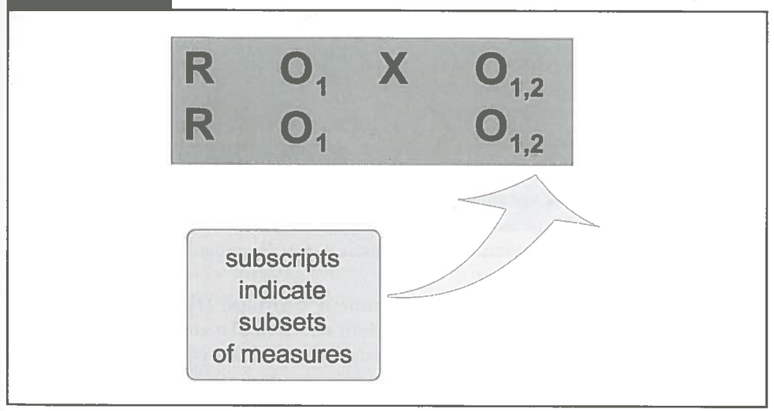


FIGURE 7–8 **An example of a design notation that includes subscripts**



design with some subscripting of the *O*s. What does this mean? Because all of the *O*s have a subscript of 1, some measure or set of measures was collected for both groups on both occasions. But the design also has two *O*s with a subscript of 2, both taken at the posttest. This means that some measure or set of measures was collected *only* at the posttest.

With this simple set of rules for describing a research design in notational form, you can concisely explain even complex design structures. In addition, using a notation helps show common design substructures across different designs that you might not recognize as easily without the notation.

7-3 Types of Designs

What are the different major types of research designs? You can classify designs into a simple threefold classification by asking some key questions as shown in Figure 7–9.

FIGURE 7–9 Basic questions that distinguish the major types of designs

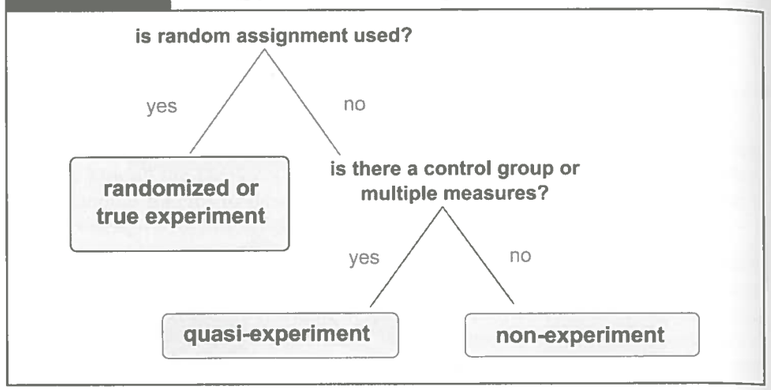


FIGURE 7–10 Notational for examples of each of the three major classes of research design

Posttest Only Randomized Experiment	R X O R O
Pretest-Posttest Nonequivalent Groups Quasi-Experiment	N O X O N O O
Posttest Only Non-Experiment	X O

First, does the design use random assignment to groups? (Don't forget that random *assignment* is not the same thing as random selection of a sample from a population!) If random assignment is used, the design is a randomized experiment or *true* experiment. If random assignment is not used, ask a second question: Does the design use *either* multiple groups or multiple waves of measurement? If the answer is yes, label it a quasi-experimental design. If not, call it a nonexperimental design.

This threefold classification is especially useful for describing the design with respect to internal validity. A randomized experiment generally is the strongest of the three designs when your interest is in establishing a cause-effect relationship. A nonexperiment is generally the weakest in this respect. I have to hasten to add here that I don't mean that a nonexperiment is the weakest of the three designs *overall*, but only with respect to internal validity or causal assessment. In fact, the simplest form of nonexperiment is a one-shot survey design that consists of nothing but a single observation *O*. This is probably one of the most common forms of research and, for some research questions—especially descriptive ones—is clearly a strong design. When I say that the nonexperiment is the weakest with respect to internal validity, all I mean is that it isn't a particularly good method for assessing the cause-effect relationships that you think might exist between a program and its outcomes.

posttest-only randomized experiment
An experiment in which the groups are randomly assigned and receive only a posttest.

To illustrate the different types of designs, consider one of each in design notation as shown in Figure 7–10. The first design is a **posttest-only randomized experiment**. You can tell it's a randomized experiment because it has an *R* at the beginning of each line, indicating random assignment. The second design is a **pre-post nonequivalent groups quasi-experiment**. You know it's not a randomized experiment because random assignment wasn't used. You also know it's not a non-experiment because both multiple groups and multiple waves of measurement exist. That means it must be a quasi-experiment. You add the label *nonequivalent* because in this design you do not explicitly control the assignment and the groups may be nonequivalent or not similar to each other (see nonequivalent group designs, Chapter 11). Finally, you see a **posttest-only nonexperimental design**. You might use this design if you want to study the effects of a natural disaster like a flood or tornado and you want to do so by interviewing survivors. Notice that in this design, you don't have a comparison group (for example, you didn't interview in a town down the road that didn't have the tornado to see what differences the tornado caused) and you don't have multiple waves of measurement (a pre-tornado level of how people in the ravaged town were doing before the disaster). Does it make sense to do the nonexperimental study? Of course! You could gain valuable information by well-conducted post-disaster interviews. However, you may have a hard time establishing which of the things you observed are due to the disaster rather than to other factors like the peculiarities of the town or pre-disaster characteristics.

pre-post nonequivalent groups quasi-experiment
A research design in which groups receive both a pre- and posttest, and group assignment is not randomized, and therefore, the groups may be nonequivalent, making it a quasi-experiment.

posttest-only nonexperimental design
A research design in which only a posttest is given. It is referred to as nonexperimental because no control group exists.

Summary

Research design helps you put together all of the disparate pieces of your research project: the participants or sample, the measures, and the data analysis. This chapter showed that research design is intimately connected with the topic of internal validity because the type of research design you construct determines whether you can address causal questions, such as whether your treatment or program made a difference on outcome measures. There are three major types of problems—threats to validity—that occur when trying to ensure internal validity. Single-group threats occur when you have only a single program group in your study. Researchers typically try to avoid single-group threats by using a comparison group, but this leads to multiple-group threats or selection threats when the groups are not comparable. Since all social research involves human interaction, you must also be concerned about social threats to internal validity that can make your groups perform differently but are unrelated to the treatment or program. Research designs can get somewhat complicated. To keep them straight and describe them succinctly researchers use design notation that describes the design in abstract form.

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.