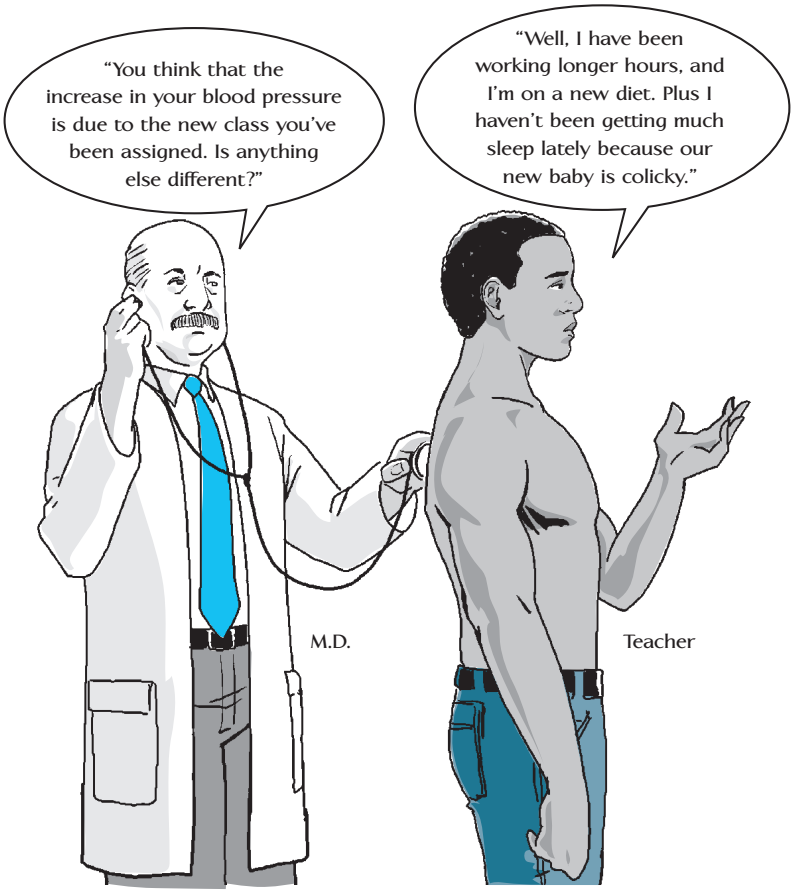
9 Internal Validity



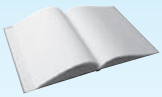
OBJECTIVES **Studying this chapter should enable you to:**

- Explain what is meant by the term "internal validity."
- Explain what is meant by each of the following threats to internal validity and give an example of each:
  - a "subject characteristics" threat
  - a "mortality" threat
  - a "location" threat
  - an "instrumentation" threat
  - a "testing" threat
  - a "history" threat
  - a "maturation" threat
  - a "subject attitude" threat
  - a "regression" threat
  - an "implementation" threat
- Identify various threats to internal validity in published research articles.
- Suggest possible remedies for specific examples of the various threats to internal validity.

INTERACTIVE AND APPLIED LEARNING After, or while, reading this chapter:

Go to the Online Learning Center at www.mhhe.com/fraenkel8e to:

- Learn More About Internal Validity



Go to your online Student Mastery Activities book to do the following activities:

- Activity 9.1: Threats to Internal Validity
- Activity 9.2: What Type of Threat?
- Activity 9.3: Controlling Threats to Internal Validity

Suppose the results of a study show that high school students taught by the inquiry method score higher on a test of critical thinking, on the average, than do students taught by the lecture method. Is this difference in scores due to the difference in methods—to the fact that the two groups have been taught differently? Surely, the researcher who is conducting the study would like to conclude this. Your first inclination may be to think the same. This may not be a legitimate interpretation, however.

What if the students who were taught using the inquiry method were better critical thinkers to begin with? What if some of the students in the inquiry group were also taking a related course during this time at a nearby university? What if the teachers of the inquiry group were simply better teachers? Any of these (or other) factors might explain why the inquiry group scored higher on the critical thinking test. Should this be the case, the researcher may be mistaken in concluding that there is a difference in effectiveness between the two methods, for the obtained difference in results may be due *not* to the difference in methods but to something else.

In any study that either describes or tests relationships, there is always the possibility that the relationship shown in the data is, in fact, due to or explained by something else. If so, the relationship observed is not at all what it seems and it may lose whatever meaning it appears to have. Many alternative hypotheses may exist, in other words, to explain the outcomes of a study. These alternative explanations are often referred to as *threats to internal validity,* and they are what this chapter is about.

What Is Internal Validity?

Perhaps unfortunately, the term *validity* is used in three different ways by researchers. In addition to internal validity, which we discuss in this chapter, you will see reference to instrument (or measurement) validity, as discussed in Chapter 8, and external (or generalization) validity, as discussed in Chapter 6.

When a study has **internal validity**, it means that any relationship observed between two or more variables should be unambiguous as to what it means rather than being due to "something else." The "something else" may, as we suggested above, be any one (or more) of a number of factors, such as the age or ability of the subjects, the conditions under which the study is conducted, or the type of materials used. If these factors are not in some way or another controlled or accounted for, the researcher can never be sure that they are not the reason for any observed results. Stated differently, internal validity means that observed differences on the dependent variable are directly related to the independent variable, and not due to some other unintended variable.

In qualitative research, a study is said to have good internal validity if alternative explanations (the "something else") have been systematically ruled out. Toward that goal, qualitative researchers should have a plan for how they treat discrepant or disconfirming data. Regardless of whether a study is qualitative or quantitative, if these "rival hypotheses" are not controlled or accounted for in some way, the researcher can never be sure that they are not the reason for any observed results.

Consider this example. Suppose a researcher finds a correlation of .80 between height and mathematics test scores for a group of elementary school students (grades 1–5)—that is, the taller students have higher math scores. Such a result is quite misleading. Why? Because it is clearly a by-product of age. Fifth-graders are taller and better in math than first-graders simply because they are older and more developed. To explore this relationship further is pointless; to let it affect school practice would be absurd.

Or consider a study in which the researcher hypothesizes that, in classes for learning-disabled students, teacher expectation of student failure is related to amount of disruptive behavior. Suppose the researcher finds a high correlation between these two variables. Should he or she conclude that this is a meaningful relationship? Perhaps. But the correlation might also be explained by another variable, such as the ability level of the class (classes low in ability might be expected to have more disruptive behavior *and* higher teacher expectation of failure).*

In our experience, a systematic consideration of possible **threats to internal validity** receives the least attention of all the aspects of planning a study. Often, the possibility of such threats is not discussed at all. Probably this is because their consideration is not seen as an essential step in carrying out a study. Researchers cannot avoid deciding on what variables to study, or how the sample will be obtained, or how the data will be collected and analyzed. They can, however, ignore or simply not think about possible alternative explanations for the outcomes of a study until after the study is completed—at which point it is almost always too late to do anything about them. Identifying possible threats during the planning stage of a study, on the other hand, can often lead researchers to design ways of eliminating or at least minimizing these threats.

In recent years, many useful categories of possible threats to internal validity have been identified. Although most of these categories were originally designed for application to experimental studies, some apply to other types of methodologies as well. We discuss the most important of these possible threats in this chapter.

Various ways of controlling for these threats have also been identified. We discuss some of these in the remainder of this chapter and others in subsequent chapters.

Threats to Internal Validity

SUBJECT CHARACTERISTICS

The selection of people for a study may result in the individuals (or groups) differing from one another in unintended ways that are related to the variables to be studied. This is sometimes referred to as *selection bias,* or a **subject characteristics threat.** In our example of teacher expectations and class disruptive behavior, the ability level of the class fits this category. In studies that compare groups, subjects in the groups may differ on such variables as age, gender, ability, socioeconomic background, and the like. If not controlled, these variables may explain away whatever differences between groups are found. The list of such subject characteristics is virtually unlimited, but some examples that might affect the results of a study include:

- Age
- Strength
- Maturity
- Gender
- Ethnicity
- Coordination
- Speed
- Intelligence
- Vocabulary
- Attitude
- Reading ability
- Fluency
- Manual dexterity
- Socioeconomic status
- Religious beliefs
- Political beliefs

In a particular study, the researcher must decide, based on previous research or experience, which variables are most likely to create problems, and do his or her best to prevent or minimize their effects. In studies comparing groups, there are several methods of equating groups, which we discuss in Chapters 13 and 16. In correlational studies, there are certain statistical techniques that can be used to control such variables, provided information on each variable is obtained. We discuss these techniques in Chapter 15.

LOSS OF SUBJECTS (MORTALITY)

No matter how carefully the subjects of a study are selected, it is common to "lose" some as the study progresses (Figure 9.1). This is known as a **mortality threat.** For one reason or another (for example, illness, family relocation, or the requirements of other activities), some individuals may drop out of the study. This is especially true in most intervention studies, since they take place over time.

Subjects may be absent during the collection of data or fail to complete tests, questionnaires, or other instruments. Failure to complete instruments is especially a problem in questionnaire studies. In such studies, it is not uncommon to find that 20 percent or more of the subjects involved do not return their forms. Remember, the actual sample in a study is not the total of those selected but only those from whom data are obtained.

Loss of subjects, of course, not only limits generalizability but also can introduce bias—*if* those subjects who are lost would have responded differently from those

*Can you suggest any other variables that would explain a high correlation (should it be found) between a teacher's expectation of failure and the amount of disruptive behavior that occurs in class?

Figure 9.1 *A Mortality Threat to Internal Validity*

from whom data were obtained. Many times this is quite likely, since those who do not respond or who are absent probably act this way for a reason. In the example we presented earlier in which the researcher was studying the possible relationship between amount of disruptive behavior by students in class and teacher expectations of student failure, it is likely that those teachers who failed to describe their expectations to the researcher (and who would therefore be "lost" for the purposes of the study) would differ from those who did provide this information in ways affecting disruptive behavior.

In studies comparing groups, loss of subjects probably will not be a problem if the loss is about the same in all groups. But if there are sizable differences between groups in terms of the numbers who drop out, this is certainly a conceivable alternative explanation for whatever findings appear. In comparing students taught by different methods (lecture versus discussion, for example), one might expect the poorer students in each group to be more likely to drop out. If more of the poorer students drop out of either group, the other method may appear more effective than it actually is.

Of all the threats to internal validity, mortality is perhaps the most difficult to control. A common misconception is that the threat is eliminated simply by replacing the lost subjects. No matter how this is done—even if they are replaced by new subjects selected randomly—researchers can never be sure that the replacement subjects will respond as those who dropped out would have. It is more likely, in fact, that they will *not*. Can you see why?*

It is sometimes possible for a researcher to argue that the loss of subjects in a study is not a problem. This is done by exploring the reasons for such loss and then offering an argument as to why these reasons are not relevant to the particular study at hand. Absence from class on the day of testing, for example, probably would not in most cases favor a particular group, since it would be incidental rather than intentional—unless the day and time of the testing was announced beforehand.

Another attempt to eliminate the problem of mortality is to provide evidence that the subjects lost were similar to those remaining on pertinent characteristics such as age, gender, ethnicity, pretest scores, or other variables that presumably might be related to the study outcomes. While desirable, such evidence can never demonstrate conclusively that those subjects who were lost would not have responded differently from those who remained. When all is said and done, the best solution to the problem of mortality is to do one's best to prevent or minimize the loss of subjects.

Some examples of a mortality threat include the following:

- A high school teacher decides to teach his two English classes differently. His one o'clock class spends a large amount of time writing analyses of plays, whereas his two o'clock class spends much time acting out and discussing portions of the same plays. Halfway through the semester, several students in the two o'clock class are excused to participate in the annual school play—thus they are "lost" from the study. If they, as a group, are better students than the rest of their class, their loss will lower the performance of the two o'clock class.
- A researcher wishes to study the effects of a new diet on building endurance in long-distance runners. She receives a grant to study, over a two-year period, a group of such runners who are on the track team at several nearby high schools in a large urban school district. The study is designed to compare runners who are given the new diet with similar runners in the district who are not given the diet. About 5 percent of the runners who receive the diet and about 20 percent of those who do not receive the diet, however, are

*Since those who drop out have done so for a reason, their replacements will be different at least in this respect; thus, they may see things differently or feel differently, and their responses may accordingly be different.

seniors, and they graduate at the end of the first year of the study. Because seniors are probably better runners, this loss will cause the remaining no-diet group to appear weaker than the diet group.

LOCATION

The particular locations in which data are collected, or in which an intervention is carried out, may create alternative explanations for results. This is called a **location threat**. For example, classrooms in which students are taught by, say, the inquiry method may have more resources (texts and other supplies, equipment, parent support, and so on) available to them than classrooms in which students are taught by the lecture method. The classrooms themselves may be larger, have better lighting, or contain better-equipped workstations. Such variables may account for higher performance by students. In our disruptive behavior versus teacher expectations example, the availability of support (resources, aides, and parent assistance) might explain the correlation between the major variables of interest. Classes with fewer resources might be expected to have more disruptive behavior and higher teacher expectations of failure.

The location in which tests, interviews, or other instruments are administered may affect responses (Figure 9.2). Parent assessments of their children at home may be different from assessments of their children at school. Student performance on tests may be lower if tests are given in noisy or poorly lighted rooms. Observations of student interaction may be affected by the physical arrangement of certain classrooms. Such differences might provide defensible alternative explanations for the results in a particular study.

The best method of control for a location threat is to hold location constant—that is, keep it the same for all participants. When this is not feasible, the researcher should try to ensure that different locations do not systematically favor or jeopardize the hypothesis. This may require the collection of additional descriptions of the various locations.

Here are some examples of a location threat:

- A researcher designs a study to compare the effects of team versus individual teaching of U.S. history on student attitudes toward history. The classrooms in which students are taught by a single teacher have fewer books and materials than the ones in which students are taught by a team of three teachers.
- A researcher decides to interview counseling and special education majors to compare their attitudes toward their respective master's degree programs. Over a three-week period, he manages to interview all of the students enrolled in the two programs. Although he is able to interview most of the students in one of the university classrooms, scheduling conflicts prevent this classroom from being available for him to interview the remainder. As a result, he interviews 20 of the counseling students in the coffee shop of the student union.

INSTRUMENTATION

The way in which instruments are used may also constitute a threat to the internal validity of a study. As discussed in Chapter 8, scores from the instruments used in a study can lack evidence of validity. Lack of this

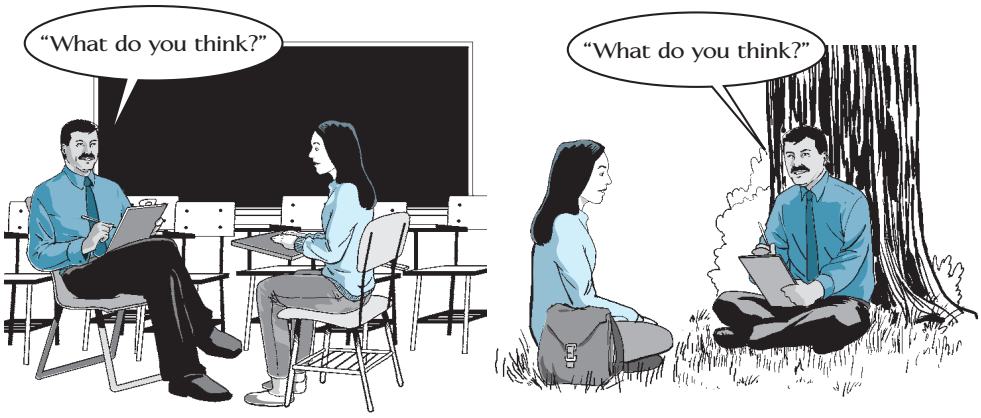


Figure 9.2 *Location Might Make a Difference*

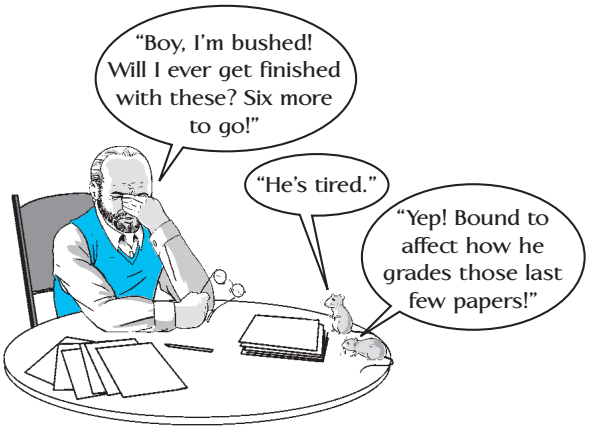


Figure 9.3 *An Example of Instrument Decay*

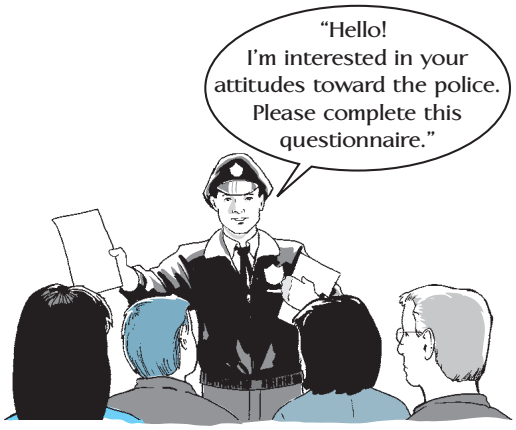


Figure 9.4 *A Data Collector Characteristics Threat*

kind of validity does not necessarily present a threat to *internal* validity—but it may.*

Instrument Decay. Instrumentation can create problems if the nature of the instrument (including the scoring procedure) is *changed* in some way or another. This is usually referred to as **instrument decay**. This is often the case when the instrument permits different interpretations of results (as in essay tests) or is especially long or difficult to score, thereby resulting in fatigue of the scorer (Figure 9.3). Fatigue often happens when a researcher scores a number of tests one after the other; he or she becomes tired and scores the tests differently (for example, more rigorously at first, more generously later). The principal way to control instrument decay is to schedule data collection and/or scoring so as to minimize changes in any of the instruments or scoring procedures.

Here are some examples of instrument decay:

- A professor grades 100 essay-type final examinations over a five-hour period without taking a break. Each essay encompasses between 10 and 12 pages. He grades the papers of each class in turn and then compares the results.
- The administration of a large school district changes its method of reporting absences. Only students who are considered truant (absence is unexcused) are reported as absent; students who have a written excuse (from parents or school officials) are not reported. The district reports a 55 percent decrease in absences since the new reporting system has been instituted.

Data Collector Characteristics. The characteristics of the data gatherers—an inevitable part of most instrumentation—can also affect results. Gender, age, ethnicity, language patterns, or other characteristics of the individuals who collect the data in a study may affect the nature of the data they obtain (Figure 9.4). If these characteristics are related to the variables being investigated, they may offer an alternative explanation for whatever findings appear. Suppose both male and female data gatherers were used in the prior example of a researcher wishing to study the relationship between disruptive behavior and teacher expectations. It might be that the female data collectors would elicit more confessions of an expectation of student failure on the part of teachers and generate more incidents of disruptive behavior on the part of students during classroom observations than would the males. If so, any correlation between teacher expectations of failure and the amount of disruptive behavior by students might be explained (at least partly) as an artifact of who collected the data.

The primary ways to control this threat include using the same data collector(s) throughout, analyzing data separately for each collector, and (in comparison-group studies) ensuring that each collector is used equally with all groups.

Data Collector Bias. There is also the possibility that the data collector(s) and/or scorer(s) may unconsciously distort the data in such a way as to make certain outcomes (such as support for the hypothesis)

*In general, we expect lack of validity of scores to make it *less* likely that any relationships will be found. There are times, however, when "poor" instrumentation can *increase* the chances of "phony" or "spurious" relationships emerging.

more likely. Examples include some classes being allowed more time on tests than other classes; interviewers asking "leading" questions of some interviewees; observer knowledge of teacher expectations affecting quantity and type of observed behaviors of a class; and judges of student essays favoring (unconsciously) one instructional method over another.

The two principal techniques for handling **data collector bias** are to standardize all procedures, which usually requires some sort of training of the data collectors, and to ensure that the data collectors lack the information they would need to distort results—also known as *planned ignorance.* Data collectors should be either unaware of the hypothesis or unable to identify the particular characteristics of the individuals or groups from whom the data are being collected. Data collectors do not need to be told which method group they are observing or testing or how the individuals they are testing performed on other tests.

Some examples of data collector bias are as follows:

- All teachers in a large school district are interviewed regarding their future goals and their views on faculty organizations. The hypothesis is that those planning a career in administration will be more negative in their views on faculty organizations than those planning to continue teaching. Interviews are conducted by the vice principal in each school. Teachers are likely to be influenced by the fact that the person interviewing them is the vice principal, and this may account for the hypothesis being supported.
- An interviewer unconsciously smiles at certain answers to certain questions during an interview.
- An observer with a preference for inquiry methods observes more "attending behavior" in inquiry-identified than noninquiry-identified classes.
- A researcher is aware, when scoring the end-of-study examinations, which students were exposed to which treatment in an intervention study.

TESTING

In intervention studies, where data are collected over a period of time, it is common to test subjects at the beginning of the intervention(s). By *testing,* we mean the use of any form of instrumentation, not just "tests." If substantial improvement is found in posttest (compared to pretest) scores, the researcher may conclude that this improvement is due to the intervention. An alternative explanation, however, may be that the improvement is due to the use of the pretest. Why is this? Let's look at the reasons.

Suppose the intervention in a particular study involves the use of a new textbook. The researcher wants to see if students score higher on an achievement test if they are taught the subject using this new text than did students who have used the regular text in the past. The researcher pretests the students before the new textbook is introduced and then posttests them at the end of a six-week period. The students may be "alerted" to what is being studied by the questions in the pretest, however, and accordingly make a greater effort to learn the material. This increased effort on the part of the students (rather than the new textbook) could account for the improvement. It may also be that "practice" on the pretest by itself is responsible for the improvement. This is known as a **testing threat** (Figure 9.5).

Consider another example. Suppose a counselor in a large high school is interested in finding out whether student attitudes toward mental health are affected by a special unit on the subject. He decides to administer an attitude questionnaire to the students before the unit is introduced and then administer it again after the unit

Figure 9.5 *A Testing Threat to Internal Validity*

is completed. Any change in attitude scores may be due to the students thinking about and discussing their opinions as a result of the pretest rather than as a result of the intervention.

Notice that it is not always the administration of a pretest per se that creates a possible testing effect, but rather the "interaction" that occurs between taking the test and the intervention. A pretest sometimes can make students more alert to or aware of what may be about to take place, making them more sensitive to and responsive toward the treatment that subsequently occurs. In some studies, the possible effects of pretesting are considered so serious that such testing is eliminated.

A similar problem is created if the instrumentation process permits subjects to figure out the nature of the study. This is most likely to happen in single-group (correlational) studies of attitudes, opinions, or similar variables other than ability. Students might be asked their opinions, for example, about teachers and also about different subjects to test the hypothesis that student attitude toward teachers is related to student attitude toward the subjects taught. They may see a connection between the two sets of questions, especially if they are both included on the same form, and answer accordingly.

Some examples of testing threats are as follows:

- A researcher uses exactly the same set of problems to measure change over time in student ability to solve mathematics word problems. The first administration of the test is given at the beginning of a unit of instruction; the second administration is given at the end of the unit of instruction, three weeks later. If improvement in scores occurs, it may be due to sensitization to the problems produced by the first test and the practice effect rather than to any increase in problem-solving ability.
- A researcher incorporates items designed to measure self-esteem and achievement motivation in the same questionnaire. The respondents may figure out what the researcher is after and react accordingly.
- A researcher uses pre- and posttests of anxiety level to compare students given relaxation training with students in a control group. Lower scores for the relaxation group on the posttest may be due to the training, but they also may be due to sensitivity (created by the pretest) to the training.

HISTORY

On occasion, one or more unanticipated, and unplanned for, events may occur during the course of a study that can affect the responses of subjects (Figure 9.6). Such an event is referred to in educational research as a **history threat**. In the study we suggested of students being taught by the inquiry versus the lecture method, for example, a boring visitor who dropped in on and spoke to the lecture class just before an upcoming examination would be an example. If the visitor's remarks in some way discouraged or turned off students in the lecture class, they might have done less well on the examination than if the visitor had not appeared. Another example involves a personal experience of one of the authors of

Figure 9.6 *A History Threat to Internal Validity*

this text. He remembers clearly the day that President John F. Kennedy died, since he had scheduled an examination for that very day. The author's students at that time, stunned into shock by the announcement of the president's death, were unable to take the examination. Any comparison of examination results taken on this day with the examination results of other classes taken on other days would have been meaningless.

Researchers can never be certain that one group has not had experiences that differ from those of other groups. As a result, they should continually be alert to any such influences that may occur (in schools, for example) during the course of a study. As you will see in Chapter 13, some research designs handle this threat better than do others.

Two examples of a history threat follow.

- A researcher designs a study to investigate the effects of simulation games on ethnocentrism. She plans to select two high schools to participate in an experiment. Students in both schools will be given a pretest designed to measure their attitudes toward minority groups. School A will then be given the simulation games during their social studies classes over a three-day period, and school B will watch travel films. Both schools will then be given the same test to see if their attitude toward minority groups has changed. The researcher conducts the study as planned, but a special documentary on racial prejudice is shown in school A between the pretest and the posttest.
- The achievement scores of five elementary schools whose teachers use a cooperative learning approach are compared with those of five schools whose teachers do not use this approach. During the course of the study, the faculty of one of the schools where cooperative learning is not used is engaged in a disruptive conflict with the school principal.

MATURATION

Often, change during an intervention may be due to factors associated with the passing of time rather than to the intervention itself (Figure 9.7). This is known as a **maturation threat**. Over the course of a semester, for example, very young students, in particular, will change in many ways simply because of aging and experience. Suppose, for example, that a researcher is interested in studying the effect of special grasping exercises on the ability of 2-year-olds to manipulate various objects. She finds that such exercises are associated with marked increases in the manipulative ability of the children over a six-month period. Two-year-olds mature very rapidly, however, and the improvement in their manipulative ability may be due simply to this fact rather than the grasping exercises. Maturation is a serious threat only in studies using pre-post data for the intervention group, or in studies that span a number of years. The best way to control for maturation is to include a well-selected comparison group in the study.

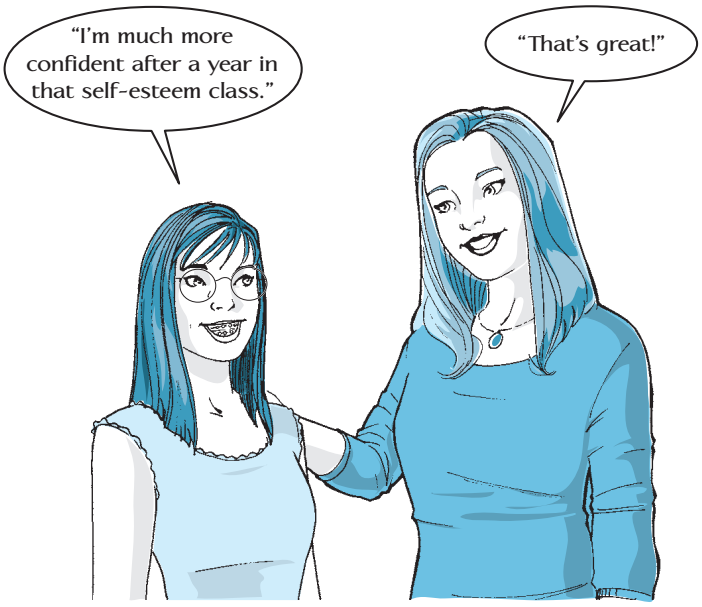


Figure 9.7 *Could Maturation Be at Work Here?*

Examples of a maturation threat are as follows:

- A researcher reports that students in liberal arts colleges become less accepting of authority between their freshman and senior years and attributes this to the many "liberating" experiences they have undergone in college. This may be the reason, but it also may be because they simply have grown older.
- A researcher tests a group of students enrolled in a special class for "students with artistic potential" every year for six years, beginning when they are age 5. She finds that their drawing ability improves markedly over the years.

ATTITUDE OF SUBJECTS

How subjects view a study and participate in it can also threaten internal validity. One example is the well-known **Hawthorne effect**, first observed in the Hawthorne plant of the Western Electric Company some years ago.[1] It was accidentally discovered that productivity increased not only when improvements were made in physical working conditions (such as an increase in the number of coffee breaks and better lighting) but also when such conditions were unintentionally made worse (for instance, the number of coffee breaks was reduced and the lighting was dimmed). The usual explanation for this is that the special attention and recognition received by the workers were responsible; they felt someone cared about them and was trying to help them. This positive effect, resulting from increased attention and recognition of subjects, has subsequently been referred to as the *Hawthorne effect.*

It has also been suggested that recipients of an experimental treatment may perform better because of the novelty of the treatment rather than the specific nature of the treatment. It might be expected, then, that subjects who know they are part of a study may show improvement as a result of a feeling that they are receiving some sort of special treatment—no matter what this treatment may be (Figure 9.8).

An opposite effect can occur whenever, in intervention studies, the members of the control group receive no treatment at all. As a result, they may become demoralized or resentful and hence perform more poorly than the treatment group. It may thus appear that the experimental group is performing better as a result of the treatment, when this is not the case.

One remedy for these **subject attitude threats** is to provide the control or comparison group(s) with a special or novel treatment comparable to that received by the experimental group. While simple in theory, this is not easy to do in most educational settings. Another possibility, in some cases, is to make it easy for students to believe that the treatment is just a regular part

Figure 9.8 *The Attitude of Subjects Can Make a Difference*

MORE ABOUT RESEARCH

Threats to Internal Validity in Everyday Life

Consider the following commonly held beliefs:

- Because failure often precedes suicide, it is therefore the cause of suicide. (probable history and mortality threats)
- Boys are genetically more talented in mathematics than are girls. (probable subject attitude and history threats)
- Girls are genetically more talented in language than are boys. (probable history and subject attitude threats)
- Minority students are less academically able than students from the dominant culture. (probable subject characteristics, subject attitude, location, instrumentation, and history threats)
- People on welfare are lazy. (probable subject characteristics, location, and history threats)
- Schooling makes students rebellious. (probable maturation and history threats)
- A policy of expelling students who don't "behave" improves a school's test scores. (probable mortality threat)
- Indoctrination changes attitude. (probable testing threat)
- So-called miracle drugs cure intellectual retardation. (probable regression threat)
- Smoking marijuana leads eventually to using cocaine and heroin. (probable mortality threat)

of instruction—that is, not part of an experiment. For example, it is sometimes unnecessary to announce that an experiment is being conducted.

Here are examples of a subject attitude threat:

- A researcher decides to investigate the possible reduction of test anxiety by playing classical music during examinations. She randomly selects 10 freshman algebra classes from the five high schools in a large urban school district. In five of these classes, she plays classical music softly in the background during examinations. In the other five (the control group), she plays no music. The students in the control group, however, learn that music is being played in the other classes and express some resentment when their teachers tell them that the music cannot be played in their class. This resentment may actually cause them to be more anxious during exams or intentionally to inflate their anxiety scores.
- A researcher hypothesizes that critical thinking skill is correlated with attention to detail. He administers a somewhat novel test that provides a separate score for each variable (critical thinking and attention to detail) to a sample of eighth-graders. The novelty of the test may confuse some students, while others may think it is silly. In either case, the scores of these students are likely to be lower on *both* variables because of the format of the test, not because of any lack of ability. It may appear, therefore, that the hypothesis is supported. Neither score is a valid indicator of ability for such students, so this particular attitudinal reaction creates a threat to internal validity.

REGRESSION

A **regression threat** may be present whenever change is studied in a group that is extremely low or high in its preintervention performance (Figure 9.9). Studies in special education are particularly vulnerable to this threat, since the students in such studies are frequently selected on the basis of previous low performance. The regression phenomenon can be explained statistically, but for our purposes it simply describes the fact that a group selected because of unusually low (or high) performance will, on the average, score closer to the mean on subsequent testing, regardless of what transpires in the meantime. Thus, a class of students of markedly low ability may be expected to score higher on posttests regardless of the effect of any intervention to which they are exposed. Like maturation, the use of an equivalent control or comparison group handles this threat—and this seems to be understood as reflected in published research.

Some examples of a possible regression threat are as follows:

- An Olympic track coach selects the members of her team from those who have the fastest times during the final trials for various events. She finds that their average time increases the next time they run, however, which she perhaps erroneously attributes to poorer track conditions.
- Those students who score in the lowest 20 percent on a math test are given special help. Six months later their average score on a test involving similar problems has improved, but not necessarily because of the special help.

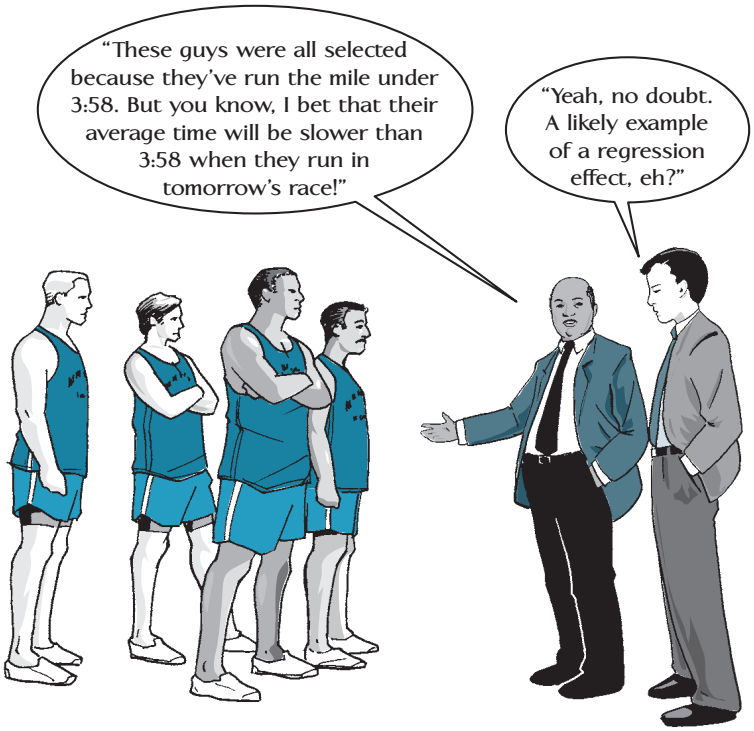


Figure 9.9 *Regression Rears Its Head*

IMPLEMENTATION

The treatment or method in any experimental study must be administered by someone—the researcher, the teachers involved in the study, a counselor, or some other person. This fact raises the possibility that the experimental group may be treated in ways that are unintended and not necessarily part of the method, yet which give them an advantage of one sort or another. This is known as an **implementation threat**. It can happen in either of two ways.

First, an implementation threat can occur when different individuals are assigned to implement different methods, and these individuals differ in ways related to the outcome. Consider our previous example in which two groups of students are taught by either an inquiry or a lecture method. The inquiry teachers may simply be better teachers than the lecture teachers.

There are a number of ways to control for this possibility. The researcher can attempt to evaluate the individuals who implement each method on pertinent characteristics (such as teaching ability) and then try to equate the treatment groups on these dimensions (for example, by assigning teachers of equivalent ability to each group). Clearly, this is a difficult and time-consuming task. Another control is to require that each method be taught by all teachers in the study. Where feasible, this is a preferable solution, though it also is vulnerable to the possibility that some teachers may have different abilities to implement the different methods. Still another control is to use *several* different individuals to implement each method, thereby reducing the chances of an advantage to either method.

Second, an implementation threat can occur when some individuals have a personal bias in favor of one method over the other. Their preference for the method, rather than the method itself, may account for the superior performance of students taught by that method. This is a good reason why a researcher should, if at all possible, *not* be one of the individuals who implements a method in an intervention study. It is sometimes possible to keep individuals who are implementers ignorant of the nature of a study, but it is generally very difficult—in part because teachers or others involved in a study will usually need to be given a rationale for their participation. One solution for this is to allow individuals to choose the method they wish to implement, but this creates the possibility of differences in characteristics discussed above. An alternative is to have all methods used by all implementers, but with their preferences known beforehand. Note that preference for a method as a *result* of using it does not constitute a threat—it is simply one of the by-products of the method itself.

MORE ABOUT RESEARCH

Some Thoughts About Meta-Analysis

As we mentioned in Chapter 3, the main argument in favor of doing a meta-analysis is that the weaknesses in individual studies should balance out or be reduced by combining the results of a series of studies. In short, researchers who do a meta-analysis attempt to remedy the shortcomings of any particular study by statistically combining the results of several (hopefully many) studies that were conducted on the same topic. Thus, the threats to internal validity that we discussed in this chapter should be reduced and generalizability should be enhanced.

How is this done? Essentially by calculating what is called *effect size* (see Chapter 12). Researchers conducting a meta-analysis do their best to locate all of the studies on a particular topic (i.e., all of the studies having the same independent variable). Once located, effect sizes and an overall average effect size for each dependent variable are calculated.* As an example, Vockell and Asher report an average delta (Δ) of .80 on the effectiveness of cooperative learning.†

*This is not always easy to do. Frequently, published reports lack the necessary information, although it can sometimes be deduced from what is reported.

†E. L. Vockell and J. W. Asher (1995). *Educational research,* 2nd ed. Englewood Cliffs, NJ: Prentice Hall, p. 361.

As we have mentioned, meta-analysis is a way of quantifying replications of a study. It is important to note, however, that the term *replication* is used rather loosely in this context, since the studies that the researcher(s) has collected may have little in common except that they all have the same independent variable. Our concerns are twofold: Merely obtaining several studies, even if they all have the same independent variable, does not mean that they will necessarily balance out each other's weaknesses—they might all have the *same* weakness. Secondly, in doing a meta-analysis, equal weight is given to both good *and bad* studies—that is, no distinction is made between studies that have been well designed and conducted and those that have not been so well designed and/or conducted. Results of a well-designed study in which the researchers used a large random sample, for example, would count the same as results from a poorly controlled study in which researchers used a convenience or purposive sample.

A partial solution to these problems that we support is to combine meta-analysis with judgmental review. This has been done by judging studies as good or bad and comparing the results; sometimes they agree. If, however, there is a sufficient number of good studies (we would argue for a minimum of seven), we see little to be gained by including poor ones.

Meta-analyses are here to stay, and there is little question that they can provide the research community with valuable information. But we do not think excessive enthusiasm for the technique is warranted. Like many things, it is a tool, not a panacea.

This is also true of other by-products. If teacher skill or parent involvement, for example, improves as a *result* of the method, it would not constitute a threat. Finally, the researcher can observe in an attempt to see that the methods are administered as intended.

Examples of an implementation threat are as follows:

- A researcher is interested in studying the effects of a new diet on the physical agility of young children. After obtaining the permission of the parents of the children to be involved, all of whom are first-graders, he randomly assigns the children to an experimental group and a control group. The experimental group is to try the new diet for three months, and the control group is to stay with its regular diet. The researcher overlooks the fact, however, that the teacher of the experimental group is an accomplished instructor of some five years' experience, while the instructor of the control group is a first-year teacher, newly appointed.
- A group of clients who stutter is given a relatively new method of therapy called *generalization training.* Both client and therapist interact with people in the "real world" as part of the therapy. After six months of receiving therapy, the fluency of these clients is compared with that of a group receiving traditional in-the-office therapy. Speech therapists who use new methods are likely to be more generally competent than those working with the comparison group. If so, greater improvement for the generalization group may be due not to the new method but rather to the skill of the therapist.

Figure 9.10 illustrates, and Table 9.1 briefly summarizes, each of the threats we have discussed.

Figure 9.10 *Illustration of Threats to Internal Validity*

Note: We are not implying that any of these statements are necessarily true; our guess is that some are and some are not.

*This seems unlikely.

†If these teacher characteristics are a *result* of the type of school, then they do not constitute a threat.

FACTORS THAT REDUCE THE LIKELIHOOD OF FINDING A RELATIONSHIP

In many studies, the various factors we have discussed could also serve to *reduce,* or even prevent, the chances of a relationship being found. For example, if the methods (the treatment) in a study are not adequately implemented—that is, adequately tried—the effect of actual differences between them on outcomes may be obscured. Similarly, if the members of a control or comparison group become "aware" of the experimental

TABLE 9.1 *Threats to the Internal Validity of a Study*

Threat	Definition
Subject Characteristics	The selection of people for a study may result in the individuals or groups differing from one another in unintended ways that are related to the variables being studied. Also called "selection bias."
Mortality	The loss of subjects in a study due to attrition, withdrawal, or low participation rates may introduce bias and affect the outcome of a study.
Location	The particular locations in which data are collected, or in which an intervention is carried out, may create alternative explanations for results.
Instrumentation	The ways in which instruments are used may constitute an internal validity threat. Possible instrumentation threats include changes in the instrument and how it is scored, characteristics of the data collector, and/or bias on the part of the data collector.
Testing	The use of a pretest in intervention studies may create a "practice effect" that can affect the results of a study and/or how participants respond to an intervention.
History	A history threat is when an unforeseen or unplanned event occurs during the course of a study.
Maturation	Change during an intervention may be due sometimes to factors associated with the passing of time rather than the intervention.
Subject Attitude	The way subjects view a study and their participation in it can be considered a threat to internal validity; the positive impact of an intervention is known as the "Hawthorne effect."
Regression	A regression threat is possible when change is studied in a group with extreme low or high performances as determined by a pretest. On average, the group will score closer to the mean on subsequent testing regardless of the treatment or intervention.
Implementation	The experimental group may be treated in unintended ways that give them an undue advantage affecting results.

treatment, they may increase their efforts because they feel "left out," thereby reducing real differences in achievement between treatment groups that otherwise would be seen. Sometimes, teachers of a control group may unwittingly give some sort of "compensation" to motivate the members of their group, thereby lessening the impact of the experimental treatment. Finally, the use of instruments that produce unreliable scores and/or the use of small samples may result in a reduced likelihood of a relationship or relationships being observed.

How Can a Researcher Minimize These Threats to Internal Validity?

Throughout this chapter, we have suggested a number of techniques or procedures that researchers can employ to control or minimize the possible effects of threats to internal validity. Essentially, they boil down to four alternatives. A researcher can try to do any or all of the following.

1. Standardize the conditions under which the study occurs—such as the way(s) in which the treatment is implemented (in intervention studies), the way(s) in which the data are collected, and so on. This helps control for location, instrumentation, subject attitude, and implementation threats.
2. Obtain more information on the subjects of the study—that is, on relevant characteristics of the subjects—and use that information in analyzing and interpreting results. This helps control for a subject characteristics threat and (possibly) a mortality threat, as well as maturation and regression threats.
3. Obtain more information on the details of the study—that is, where and when it takes place, extraneous events that occur, and so on. This helps control for location, instrumentation, history, subject attitude, and implementation threats.

4. Choose an appropriate design. The proper design can do much to control these threats to internal validity.

Because control by design applies primarily to experimental and causal-comparative studies, we shall discuss it in detail in Chapters 13 and 16. The four alternatives are summarized in Table 9.2.

TWO POINTS TO EMPHASIZE

We want to end this chapter by emphasizing two things. First, these various threats to internal validity can be greatly reduced by planning. Second, such planning often requires collecting additional information *before* a study begins (or while it is taking place). It is often too late to consider how to control these threats once the data have been collected.

TABLE 9.2 *General Techniques for Controlling Threats to Internal Validity*

	Technique			
Threat	**Standardize Conditions**	**Obtain More Information on Subjects**	**Obtain More Information on Details**	**Choose an Appropriate Design**
Subject characteristics		X		X
Mortality		X		X
Location	X		X	X
Instrumentation	X		X	
Testing				X
History			X	X
Maturation		X		X
Subject attitude	X		X	X
Regression		X		X
Implementation	X		X	X

Go back to the **INTERACTIVE AND APPLIED LEARNING** feature at the beginning of the chapter for a listing of interactive and applied activities. Go to the **Online Learning Center** at **www.mhhe.com/fraenkel8e** to take quizzes, practice with key terms, and review chapter content.

Main Points

THE MEANING OF INTERNAL VALIDITY

- When a study lacks internal validity, one or more alternative hypotheses exist to explain the outcomes. These alternative hypotheses are referred to by researchers as *threats to internal validity.*
- When a study has internal validity, it means that any relationship observed between two or more variables is unambiguous, rather than being due to something else.

THREATS TO INTERNAL VALIDITY

- Some of the more common threats to internal validity are differences in subject characteristics, mortality, location, instrumentation, testing, history, maturation, attitude of subjects, regression, and implementation.

- The selection of people for a study may result in the individuals or groups differing (i.e., the characteristics of the subjects may differ) from one another in unintended ways that are related to the variables to be studied.
- No matter how carefully the subjects of a study (the sample) are selected, it is common to lose some of them as the study progresses. This is known as *mortality.* Such a loss of subjects may affect the outcomes of a study.
- The particular locations in which data are collected, or in which an intervention is carried out, may create alternative explanations for any results that are obtained.
- The way in which instruments are used may also constitute a threat to the internal validity of a study. Possible instrumentation threats include changes in the instrument, characteristics of the data collector(s), and/or bias on the part of the data collectors.
- The use of a pretest in intervention studies sometimes may create a "practice effect" that can affect the results of a study. A pretest can also sometimes affect the way subjects respond to an intervention.
- On occasion, one or more unanticipated and unplanned for events may occur during the course of a study that can affect the responses of subjects. This is known as a *history threat.*
- Sometimes change during an intervention study may be due more to factors associated with the passing of time than to the intervention itself. This is known as a *maturation threat.*
- The attitude of subjects toward a study (and their participation in it) can create a threat to internal validity. This is known as *subject attitude threat*.
- When subjects are given increased attention and recognition because they are participating in a study, their responses may be affected. This is known as the *Hawthorne effect.*
- Whenever a group is selected because of unusually high or low performance on a pretest, it will, on average, score closer to the mean on subsequent testing, regardless of what transpires in the meantime. This is called a *regression threat.*
- Whenever an experimental group is treated in ways that are unintended and not a necessary part of the method being studied, an implementation threat can occur.

CONTROLLING THREATS TO INTERNAL VALIDITY

- Researchers can use a number of techniques or procedures to control or minimize threats to internal validity. Essentially they boil down to four alternatives: (1) standardizing the conditions under which the study occurs, (2) obtaining and using more information on the subjects of the study, (3) obtaining and using more information on the details of the study, and (4) choosing an appropriate design.

Key Terms

data collector bias 171
Hawthorne effect 174
history threat 172
implementation threat 176
instrument decay 170
internal validity 166
location threat 169
maturation threat 173
mortality threat 167
regression threat 175
subject attitude threat 174
subject characteristics threat 167
testing threat 171
threats to internal validity 167

For Discussion

1. Can a researcher prove conclusively that a study has internal validity? Explain.
2. In Chapter 6, we discussed the concept of external validity. In what ways, if any, are internal and external validity related? Can a study have internal validity but not external validity? If so, how? What about the reverse?
3. Students often confuse the concept of internal validity with the idea of instrument validity. How would you explain the difference between the two?
4. What threat (or threats) to internal validity might exist in each of the following?
   a. A researcher decides to try out a new mathematics curriculum in a nearby elementary school and to compare student achievement in math with that of students in another elementary school using the regular curriculum. The researcher is not aware, however, that the students in the new-curriculum school have computers to use in their classrooms.
   b. A researcher wishes to compare two different kinds of textbooks in two high school chemistry classes over a semester. She finds that 20 percent of one group and 10 percent of the other group are absent during the administration of unit tests.
   c. In a study investigating the possible relationship between marital status and perceived social changes during the last five years, men and women interviewers get different reactions from female respondents to the same questions.
   d. Teachers of an experimental English curriculum as well as teachers of the regular curriculum administer both pre- and posttests to their own students.
   e. Eighth-grade students who volunteer to tutor third-graders in reading show greater improvement in their own reading scores than a comparison group that does not participate in tutoring.
   f. A researcher compares the effects of weekly individual and group counseling on the improvement of study habits. Each week the students counseled as a group fill out questionnaires on their progress at the end of their meetings. The students counseled individually, however, fill out the questionnaires at home.
   g. Those students who score in the bottom 10 percent academically in a school in an economically depressed area are selected for a special program of enrichment. The program includes special games, extra and specially colored materials, special snacks, and new books. The students score substantially higher on achievement tests six months after the program is instituted.
   h. A group of elderly people are asked to fill out a questionnaire designed to investigate the possible relationship between activity level and sense of life satisfaction.
5. How could you determine whether the threats you identified in each of the situations in question 4 actually exist?
6. Which threats discussed in this chapter do you think are the most important for a researcher to consider? Why? Which do you think would be the most difficult to control? Explain.

Note

1. F. J. Roethlisberger and W. J. Dickson (1939). *Management and the worker.* Cambridge, MA: Harvard University Press.

Research Exercise 9: Internal Validity

State the question or hypothesis of your study at the top of Problem Sheet 9. In the spaces indicated, place an *X* after each of the threats to internal validity that apply to your study, explain why they are threats, and describe how you intend to control for those most likely to occur (i.e., prevent them from affecting the outcome of your study). Finally, what can you say to convince others that the results of your study are credible and not due merely to coincidence or chance?

Problem Sheet 9

Internal Validity

1. Place an *X* after any of the threats listed below that you think might apply to your study:

   Subject characteristics _____ Instrumentation _____ Maturation _____

   Mortality _____ Testing _____ History _____ Subject attitude _____

   Implementation _____ Location _____ Regression _____ Other

2. Please describe how you will attempt to control for those threats that you have marked above:

   Threat #1: _____

   _____

   Threat #2: _____

   _____

   Threat #3: _____

   _____

   Threat #4: _____

   _____

3. What assurances can you provide (through your design, sampling procedure, etc.) to support the claims that your study findings are valid? In other words, how will you convince the reader that the findings or relationships resulting from the study are not due to or explained by something other than what you claim?

   _____

   _____

   _____

An electronic version of this Problem Sheet that you can fill in and print, save, or e-mail is available on the Online Learning Center at www.mhhe.com/fraenkel8e.

This page intentionally left blank