Validity of Research Results

LEARNING OBJECTIVES

To be able to
- Explain the meaning of confounding variables.
- Explain the meaning of statistical conclusion validity, internal validity, and external validity and their importance in the research process.
- Identify and explain the types of evidence that are needed to reach a causal conclusion.
- Explain the threats to internal validity and be able to identify when they might exist in a research study.
- Explain the threats to external validity and when they might exist in a research study.
- Explain the role of operationalization of constructs in research.
- Identify and explain the types of validity used in qualitative research.
VALIDITY ISSUES IN THE DESIGN OF QUANTITATIVE RESEARCH

In Chapter 1, we stated that description, explanation, explanation, prediction, and influence represent the objectives of research. To accomplish these objectives, we must design research studies that allow us to collect uncontaminated data. This seems to be a straightforward requirement. In most quantitative studies, we want to identify the effect created by some independent variable and to be able to generalize the results beyond the confines of the study. However, in every study, there is the possibility that some variable other than the independent variable influenced the dependent variable or limited the ability to generalize the results. For example, if you are investigating the effects of parent’s involvement in their child’s education (independent variable) on the child’s achievement test scores (dependent variable), you probably want to conclude that greater parent involvement results in high achievement test scores. However, if the parents with the greater involvement also have the brightest children, the higher achievement test scores could be due to the child’s greater intellect. In such an instance, intellect would be an extraneous variable, a variable other than the ones you are specifically studying, which might have confounded the results of the study.

Extraneous variables might or might not introduce a confound. Extraneous variables are confounded when they systematically vary with the independent variable and also influence the dependent variable. It is impossible to draw clear and valid conclusions from the data collected when an uncontrolled confounding variable exists.

To illustrate how extraneous variables can confound the outcome of a study and produce ambiguous results, consider a hypothetical “Pepsi Challenge” study. Assume that Pepsi wants to conduct a study demonstrating that its product is preferred over Coke. In this study, researchers draw samples, and in random order, Pepsi in a cup marked with an M and Coke in a cup marked with a Q. The research participants are to drink the beverage in each cup and then identify the one they like more. Now assume that 80 percent of the participants indicate that they prefer the beverage in the cup marked with an M. Pepsi would take this as an indication that their product was preferred over Coke. However, if people are more likely to choose something with the letter M over the letter Q, this could influence their selection of the beverage of choice. If the letter on the cup does influence choice, the results are ambiguous because it is impossible to tell whether the choice was due to the beverage or to the letter that appeared on the cup. This is the type of subtle extraneous variable that can systematically confound the outcome of a study and lead to ambiguous results.

The key component here is that the extraneous variable has to vary systematically with the independent variable and influence the dependent variable to represent a confound. There are many extraneous variables surrounding a study, but they do not represent a confound because they vary randomly with the independent variable.

To conduct a research study that will provide an answer to your research question, you must develop a plan, outline, or strategy to use in data collection. You naturally want to develop a plan or strategy that will allow you to collect data that will lead to a valid conclusion. To accomplish this goal, you must have knowledge of the factors that will lead to both valid and invalid conclusions. These factors are somewhat different depending on whether you are conducting a quantitative study or a qualitative study.

STATISTICAL CONCLUSION VALIDITY

Statistical conclusion validity refers to the validity with which we can infer that two variables are related and the strength of that relationship. From this definition, you can see that statistical conclusion validity refers to statistical inferences. The first statistical inference is whether a relationship exists between the independent and dependent variables. The second
statistical inference is an estimate of the magnitude of the relationship between the independent and dependent variables. Both of these inferences rely on statistical tests. Making an inference about whether the variables that are investigated in the study are related typically involves null hypothesis significance testing. We will discuss this in more depth in Chapter 16. Right now, all you need to know is that null hypothesis statistical testing involves making use of statistical tests to decide whether the independent and dependent variables are related. Making an inference about the magnitude of the relationship between the variables involves computing effect size estimates. Effect size estimates are obtained by computing a statistical index that gives an estimate of the strength of the relationship between the independent and dependent variables.

On the surface, it seems as though valid inferences should logically follow if the statistical tests are conducted correctly. However, there are a variety of reasons why a researcher might be wrong when making an inference about the existence of a relationship between two or more variables or the size of the relationship between these variables. We are not going to discuss these threats because they focus primarily on statistical issues that are beyond the scope of this textbook. We do want you to realize that the inferences you make from the results of statistical tests might or might not be valid and whether they are valid depends on the existence or nonexistence of a variety of threats. The interested reader can find these threats on the companion website for this book at www.ablongman.com/johnsonchristensen2e, under bonus material.

### REVIEW QUESTIONS

1. Write a research variable and why does a researcher use this in a research study?
2. Define and discuss each of the types of variables used in research. Include the definitions and examples for each.
3. What is the relationship between the size of the relationship and sample size in determining statistical significance?

### INTERNAL VALIDITY

**Internal validity**

Internal validity is a term coined by Campbell and Stanley (1963). Cook and Campbell (1979) later refined the concept to refer to the “approximate validity with which we infer that a relationship between two variables is causal” (p. 37). Although research is conducted for the multiple purposes of description, explanation, prediction, and influence, most research focuses on the goal of attempting to determine whether a causal relationship exists between the variables being investigated.

#### Types of Causal Relationships

Shadish, Cook, and Campbell (2002) have pointed out that there are two different types of causal relationships: causal description and causal explanation. Causal description refers to describing the consequences of manipulating an independent variable. Causal explanation refers to explaining the mechanisms through which and the conditions under which a causal relationship holds. For example, assume that a study was conducted to investigate the benefit derived from incorporating a 10-month intensive internship program into the education of future teachers, as is being done by the Chicago public school district. Assume further that this study demonstrated that teachers who participated in this program were evaluated by their principals as being more effective than teachers who participated in the traditional two- to three-month internship. This study has therefore provided evidence of a causal relationship between teacher effectiveness and the length and intensiveness of the internship. This would be causal description because it would have described the causal relationship that exists between the more intensive internship and later teaching effectiveness.

This study would not, however, explain why this causal relationship exists. The teachers participating in this experimental program might be more effective for any of a number of reasons, such as the program giving them better skills to cope with difficult children, better organizational skills, better skills at presentation of material, more realistic expectations of the demands of the teaching profession, and so on. The only thing that the study would have demonstrated is the causal connection between the more intensive internship program and later teacher effectiveness. A full explanation of why the causal relationship exists, has to “show how the causally efficacious parts of the treatment influence the causally affected parts of the outcome through identified mediating processes” (Shadish, Cook, & Campbell, 2002, p. 9). In other words, causal explanation requires that you identify and show how the processes involved in the intensive internship caused a change in a person's later effectiveness as a teacher. Clearly, many factors are involved in the intensive internship program that influence the student's later teaching effectiveness, such as the student's motivation, the quality of instruction provided by the master teachers, and the support provided by the principal. Once a causal descriptive relationship has been found, most subsequent research is directed at explaining why and how the descriptive relationship exists, and this is causal explanatory research.

The practical importance of causal explanation can be seen if a subsequent study does not replicate the beneficial effect previously demonstrated from the 10-month internship program. If explanatory studies had been conducted, this information could be used to show how to fix the program that did not produce the beneficial results. However, identifying how and why a relationship exists is much more difficult than describing that relationship.

#### Criteria for Inferring Causation

Three types of evidence are needed to reach a causal conclusion. First, you need some evidence that the independent and dependent variables are related. In other words, do changes in the independent variable correspond to changes in the dependent variable? For example, assume that you want to know whether being absent from school, the independent variable, has any effect on the grades students make, the dependent variable. If there is no relationship between these two variables, then one obviously cannot affect the other. However, if there is some relationship between the independent and dependent variables, it is possible that they are causally related. Note that we said that it is possible that the two variables are causally related. We used the word possible because evidence of covariation or correlation...
The second source of evidence needed to infer causation is the temporal ordering of the variables being investigated, because a cause must precede an effect. This means that you need some knowledge of the time sequence of the events. For example, if you are studying the causal relationship between grades and number of absences from school, you might want to determine whether grades cause absences from school or whether absence from school has a causal influence on grades. At first glance, you might think that the direction of causality is from absence from school to grades, with fewer absences resulting in higher grades and more absences resulting in lower grades. However, it is also possible that the direction of causality is from grades to school absences. It might be that students with poor grades become frustrated and therefore miss school, whereas students with good grades enjoy school, so they demonstrate better attendance than do the students with poor grades. As you can see, the temporal order of the relationship has to be identified to reach a causal conclusion because the cause must precede the effect. It is logically impossible for an effect to occur before the existence of its cause.

The third type of evidence needed to reach a causal conclusion is that the variables being investigated are the ones that are causally related rather than being caused by some extraneous variable. In other words, we must look for variables other than the independent variable that might explain the change observed on the dependent variable. In the "Pepsi Challenge" experiment, the letter on the cup represents a logical explanation for participants' preference selection. In the example of student grades and attendance, it is possible that the grades students get and their attendance at school are caused by parents monitoring their children. Children whose parents do not monitor their children's behavior might have poorer grades and lower school attendance, whereas children who are monitored by their parents might get better grades and have fewer absences. In this instance, there is still a relationship between grades and school attendance, but the cause of this relationship is the third variable: parent monitoring. Third variable is simply another term to refer to a confounding extraneous variable.

This third-variable issue means that two variables may be correlated not because they are causally related, but because they are both related to some third variable. Consider the research that has attempted to relate coffee drinking to heart attacks (e.g., Brody, 1975). Some of these studies have found a positive correlation between the number of cups of coffee drunk and the incidence of heart attacks. From this positive relationship, it is tempting to make a causal inference that coffee consumption is contributing to the risk of having a heart attack. This would be a tenuous assumption because cigarette smoking is related to both heart attacks and coffee drinking. Cigarette smoking therefore could be the third variable that is related to and possibly causing both coffee drinking and heart attacks. Nonsmokers not only do not consume much coffee, but also have few heart attacks, whereas smokers consume large amounts of coffee and have a considerable number of heart attacks. Therefore, although there is a relationship between coffee consumption and heart attacks, this relationship could be caused by an underlying third variable of smoking.

As you can see from this example, a researcher cannot automatically assume causality just because two or more variables are related. Before such a causal connection can be made, the temporal ordering of the variables must be established, and some assurance must be provided that an extraneous variable is not causing the observed relationship. Establishing the temporal ordering of variables is easily accomplished in experiments because the experimenter actively manipulates and has control over the presentation of the independent variable, the causal variable, and observes the effect of this manipulation on the dependent variable, the effect. Because the independent variable is presented first and the dependent variable is measured after the occurrence of the independent variable, the time sequence of the events is established in experimental research.

Research studies other than experimental studies also frequently attempt to infer causality. In these studies, the direction of causality is more difficult to establish because of the difficulty in establishing the temporal sequencing of events and ruling out the influence of confounding variables.

### Threats to Internal Validity

To infer one variable was the cause of an effect observed in another variable, we must eliminate or control for all other possible causes. These other possible causes are the threats to internal validity because these threats represent rival and competing explanations for the results obtained. When these rival explanations exist, it is impossible to reach a causal explanation with any degree of certainty, leading to highly suspect results that cannot and should not be taken seriously. This is why it is necessary to control for and eliminate the systematic influence of these threats. These threats to the internal validity of a research study are greatest in a one-group pretest-posttest design. As Figure 8.1 illustrates, this is a research design in which one group of participants is pretested on some dependent variable and then administered a treatment condition. Some time after the treatment condition is administered, the group of participants is posttested on the dependent variable. For example, assume that you want to test the effects of a new drug on controlling the adverse behavior of children with ADHD. You first identify a group of children with ADHD and pretest them on some performance measure such as ability to perform a series of simple arithmetic computations in a fifteen-minute time period. You then give these children the treatment consisting of the new drug. After the children have taken the drug for two months, you posttest them on the arithmetic computations task to see whether their posttest performance is superior to their pretest performance. Although it might be tempting to interpret any improvement in performance from pretesting to posttesting as being due to the new drug, there are a number of

![Figure 8.1 One-group pretest-posttest design](image-url)
other variables that could exist during the interval between pretesting and posttesting that could also affect the posttest performance. These other variables represent threats to the internal validity of the study.

There are also a number of variables that can threaten the internal validity of a study that uses a multigroup research design. The most basic multigroup research design is the two-group design. As illustrated in Figure 8.2, this design has two groups of participants, one of which receives a treatment condition while the other does not. Both groups are posttested on the dependent variable following administration of the treatment condition, and the posttest results are compared to see whether the group that received the treatment condition responds differently on the posttest than does the group that did not receive the treatment condition. This design could, for example, be used to test the effect of a drug on controlling the behavior problems of children with ADHD. One group of children with ADHD is given the drug, and the other group is given a placebo (a pill that looks like the drug but does not contain the drug). Following drug administration, both groups of children with ADHD are posttested on measures such as the amount of time they spend out of their seat. The two groups are compared to determine whether the children who receive the drug stay in their seats for a longer period of time than do the children that do not receive the drug. There are a number of variables that can operate in this design and can threaten its internal validity, making it a very weak design. We now discuss some of the more obvious and common threats to the internal validity. We will begin our discussion with a threat that exists in some nonexperimental studies and then discuss threats that can exist in the one-group pretest-posttest design and the multi-group design.

Ambiguous Temporal Precedence The ambiguous temporal precedence threat refers to the inability to specify which variable preceded which other variable. Remember that a cause must precede the effect, so to identify the causal variable, you have to know that it came first and the effect followed. In some nonexperimental studies, especially those that investigate the degree of relationship that exists between two variables, it is frequently unclear as to whether variable A preceded variable B or vice versa. For example, if we measured the degree of relationship that existed between the frequency of criminal behavior and frequency of incarceration, we would probably find that these two variables were correlated. On the surface, this might suggest that the causal direction would be from criminal behavior to incarceration. However, many individuals learn techniques for engaging in criminal behavior from association with other individuals while incarcerated, so being incarcerated might lead to more criminal behavior. In this case, it is difficult to identify which variable is the cause and which is the effect because it is difficult to identify which variable came first. This is why you frequently find the statement “correlation does not indicate causation,” because a correlation between two variables only indicates their degree of relationship. This does not mean that a correlation between two variables cannot indicate causation, because sometimes one direction of causality is implausible, as might exist with the relationship between clogged arteries and heart attacks. It is implausible to assume that a heart attack caused a person’s artery to clog up with plaque. So you can appropriately infer that the clogged artery came first. However, just because A precedes B does not justify claiming that A caused B. To make that statement, you must eliminate other possible causes for the observed relationship. This is why experiments were created. Only in experiments do we deliberately manipulate one variable before the other is measured and attempts to create an environment that rules out other confounding variables so that we can state that the variable we manipulated was the causal variable.

History A history effect may occur in a one-group pretest-posttest research design in which the pretest and posttest measures of the dependent variable are separated by a rather lengthy time interval. History refers to the specific events, other than any planned treatment event, that occur between the first and second measurement of the dependent variable, as illustrated in Figure 8.3. These events, in addition to any treatment effect, can influence the postmeasurement of the dependent variable; therefore, these events are confounded with the treatment effect and become rival explanations concerning the change that occurred between the pretest and posttest measurements.

Consider a study investigating the effect of a peer tutoring procedure on spelling performance. This is a procedure in which one student serves as a tutor and the other as a tutee. Tutors dictate words to a tutee and provide feedback as to whether the tutee correctly spells the word and the correct spelling if the word is spelled wrong. After a given number of words, the students reverse roles and continue the tutoring procedure. One approach to investigating the efficacy of such a tutoring procedure is to test the students on the speed with which they can learn to correctly spell a list of words before the tutoring procedure is implemented. Then implement the tutoring procedure. After the students have had an opportunity to practice and become familiar with this procedure, test them again on the speed with which they learn to correctly spell a list of words equivalent to the list that they had previously been asked to learn. If they require less time to learn to correctly spell the list of words after the tutoring procedure is implemented than before it is implemented, this should indicate that the peer tutoring system is a more effective method of spelling instruction.

The difficulty with making this assumption is that a time interval elapsed between the pretest and posttest measurements. It is possible that some event other than just the tutoring system has an effect on the participants during this time and this event influences their
performance on the spelling test. For example, to implement the peer tutoring system, the teacher has to provide instruction to the students and constantly monitor their performance to ensure that they are conducting the peer tutoring correctly. This monitoring by the teacher might increase the students' motivation to learn to spell the list of words and affect their spelling performance. If the monitoring does influence the students' motivation and therefore their spelling performance, it represents a variable that is valid and functions as a rival explanation for the validity of studies because they represent plausible rival explanations for the outcome of the study.

Maturational Changes
Any physical or mental changes that occur over time that affect the performance on the dependent variable.

Consider a study in which Snowling, Goadland, and Defty (1986) followed a group of children with dyslexia over a two-year period to track the development of their literacy skills. Over this two-year time period, the children demonstrated an improvement in reading, spelling, and vocabulary. This improvement in the literacy skills of the children with dyslexia may represent a maturational effect that could threaten the internal validity of a study. Snowling et al. included two control groups to rule out maturational effects. The inclusion of these groups is not mentioned here to illustrate the maturational threat to internal validity.

For example, assume that Snowling et al. wanted to assess the effect of a special instructional program on the development of literacy skills of children with dyslexia. To test the effect of this program, they tested a group of children with dyslexia on reading, spelling, and vocabulary before they entered the special instructional program and then tested them a second time after they had been in the program for two years. In comparing the pretest with the posttest scores, the investigators found that the children with dyslexia made significant advancements in literacy skills. Although it might be tempting to attribute the improvement in literacy skills to the special instructional program, all or part of the improvement could have been due to a maturational effect, or the improvement in learning that would take place, is without the special instructional program that this natural or normal improvement could result from a maturational effect rather than the special explanation for the student's performance in literacy skills for the children with dyslexia and represents a threat to the internal validity of the study. To conduct an internally valid study, such maturational threats must be controlled.

Testing
Any change in scores obtained on the second administration of a test as a result of having previously taken the test.

Testing refers to changes that may occur in participants' scores obtained on the second administration of a test as a result of previously having taken the test. In other words, the experience of having taken a pretest may alter the results obtained on the posttest independent of any treatment effect or experimental manipulation intervening between the pretest and the posttest. Taking the test does a number of things that can alter a person's performance in a subsequent administration of the same test. Taking a test familiarizes you with the content of the test. After taking a test, you might think about errors you made that could be corrected if the test were taken over. When the test is administered a second time, you are already familiar with it and might remember some of your prior responses. This can lead to an enhanced performance that is entirely due to the initial or posttest administration. Any alteration in performance as a result of a testing effect threatens the internal validity of a one-group study because it serves as a rival hypothesis to some treatment effect or experimental manipulation intervening between the pretest and posttest. Whenever the same test is administered on multiple occasions, some control needs to be implemented for testing rival hypotheses.

Snowling et al. (1996), for example, administered a number of reading, spelling, and vocabulary tests to children with dyslexia at the beginning of their study and two years later. Some of these tests were a usual analog, such as the Rhyme Sensitivity Test, which presented children with a string of four words (e.g., cot, hot, fox, pot) for which they were to identify the odd word in the rhyme segment (fox in this example). The unusual nature of this test suggests that it might be subject to a testing effect because it would seem as though, after participating in this test, children would become more familiar with it and would be able to perform better on a subsequent administration of the test. If this familiarization effect did exist, it would account for some of the improvement in performance demonstrated by the dyslexic children. It would therefore serve as a rival explanation for any improvement demonstrated over the two years and preclude a clear explanation of the observed improvement. Snowling et al. (1996) did attempt to control for such a pretesting effect by administering two practice sessions in which the children could become familiar with the test and minimize a pretesting effect. Such practice sessions do not necessarily eliminate a pretesting effect. However, including control groups that also experienced any testing effect would have controlled for this potential threat to internal validity. (If you read the Snowling et al. study, you will see that control groups were included. We did not mention them until now so that we could illustrate the testing threat to internal validity.)

Instrumentation
Instrumentation refers to any change that occurs in the measuring instrument. There are two primary ways in which an instrumentation threat may occur: An instrumentation threat can occur when the measurement instrument that is used during pretesting is different from that used during posttesting. If the tests used during pretesting and posttesting are not completely equivalent, there may be a difference between the two performance measures that is due to the difference in the way the two tests are assessing performance. For example, assume that children with dyslexia are tested at time 1 with one test of reading sensitivity and are tested two years later with a different test of reading sensitivity. If a comparison is made of reading sensitivity from time 1 and two years later at time 2, any difference that is observed could be due to the children's enhanced development of reading sensitivity. However, it could also be due to the differences in the way the two tests measure reading sensitivity, which would be an instrumentation effect that would represent a rival explanation for any change observed in reading sensitivity. Instrumentation effects therefore represent threats to the internal validity of any study.

A second way in which an instrumentation effect could creep into a study is if the data were collected through observation. Many educational researchers use human observers to collect data. Human observers such as teachers are, unfortunately, subject to such influences as fatigue, boredom, and learning processes. In administering intelligent tests, the tests typically gain facility and skill over time and collect more reliable and valid data as additional tests are given. Observers and interviewers are also used to assess the effects of various experimental treatments. For example, Schafer and Smith (1996) had teachers and children view videotapes of children engaged in playful and real fights to make judgments as to whether the
Regression artifacts
The tendency of very high scores to become lower and very low scores to become higher on posttesting of other or the original measure.

Regressional artifacts.

The concept of regression artifact refers to the fact that extreme scores will tend to regress or move toward the mean of a distribution on a second testing or assessment. Many educational research studies are designed in such a way that the research participants are tested before and after some experimental treatment condition is administered for the purpose of assessing change. Additionally, many of these studies investigate special groups of individuals such as children with learning disabilities or people with a specific deficiency such as poor reading or mathematical ability. These special groups of research participants are typically identified by having extreme scores such as low reading comprehension scores. After the research participants are selected, they are given some experimental treatment condition to improve this deficiency or ameliorate the special condition. Any positive change from pretesting to posttesting is frequently taken as evidence of the efficacy of the treatment program. However, the internal validity of a study such as this could be threatened because high-scoring research participants might score lower on posttesting or low-scoring research participants might score higher on posttesting not because of any experimental treatment effect but because of a regression artifact.

Regression artifacts occur because the first and second measurements of performance are not perfectly correlated. This lack of perfect correlation occurs because a person's test-taking performance is influenced by many variables in addition to ability. Think about the variables that could influence the score you receive on a reading comprehension test. There will undoubtedly be some questions that you cannot answer correctly. However, you guess; sometimes you guess correctly, and at other times you guess incorrectly. In addition to chance factors such as guessing, your test-taking performance could be influenced by other stresses in your life, such as not sleeping well the night before the test or having a fight with your roommate or significant other. You could also have misread questions, which might have led you to answer incorrectly. As you can see, there are many variables in addition to a person's ability that might influence the score obtained on a test.

To illustrate the regression effect, assume you want to test a technique that is supposed to increase the reading comprehension of young children. To test this technique, you give a reading comprehension test to a group of six- to ten-year-old children and select for your study all those children who received the lowest 10 percent of the scores on this test. Naturally, some of these individuals received low scores because they had very poor reading comprehension ability. However, others probably received low reading comprehension scores because they did not try very hard, were tired because they stayed up late the night before, or were especially stressed because of something like moving to another school or their parents’ getting a divorce. These individuals would have artificially low scores because of these extraneous factors. On retesting, these children would be expected to do better because it is unlikely that these extraneous factors would again operate to the same extent to depress the reading comprehension scores. Consequently, the posttest scores would be higher. However, these higher scores would be the result of a regression phenomenon and not because of the experimental treatment meant to improve reading comprehension. In this case, regression would threaten the internal validity of the study.

Differential Selection
Differential Selection is a threat to the internal validity of a study when a difference exists, at the outset of the research study, between the characteristics of the participants forming the various comparison groups. Participants in different comparison groups can differ in many ways, as illustrated in Table 8.1. One way in which this difference can occur is if you, as the researcher, have to use groups of participants that are already formed. For example, assume you want to test a procedure for enhancing young children's motivation to learn. To test this procedure, you want to administer it to one group of fourth-grade children and compare their motivation to learn, after this procedure has been implemented, with a group of fourth-grade children who have not experienced this procedure. In conducting this study, you obtain permission from the local school district. However, you have to administer the experimental procedure to one fourth-grade class and compare its performance with that of another fourth-grade class. This might not seem to be a problem because they both represent fourth-grade students. However, there is no guarantee that the students in these two classes have the same motivation to learn before the study is conducted. If the class that receives the experimental procedure has a greater motivation to learn before conducting the study, they naturally show up as having a greater motivation to learn after the experimental procedure is implemented. Any difference in motivation to learn between the two fourth-grade classes could therefore be due entirely to a selection bias.

Additive and Interactive Effects
Additive and interactive effects refers to the fact that the threats to internal validity can combine to produce complex biases. For example, a selection-history effect would occur if two comparison groups experienced a different
History event and the history effects they experienced resulted in their responding in different ways to the dependent variable. For example, assume that a joke was cracked at some time during the study in one of the comparison groups. In the other comparison group, one of the participants got very angry and frustrated at having to do the experimental task and voiced his disapproval aloud. Each of these two experiences represents a history effect, but the history effect occurring in each group would have been different. The different history effects, the joke or the expression of disapproval, would probably have affected the participants in each group very differently and might also have had a differential effect on their response to the dependent variable. If it did, this differential response would look like a treatment effect but would be a selection by history interaction effect.

To illustrate a selection-maturation effect, suppose you want to teach the concepts of good and bad to five-year-old children with and without hearing difficulties. In doing so, you find that the normal children learn these concepts much faster than do the children with hearing difficulties. From this study, you might conclude that the ability of children who have hearing difficulties to learn these concepts is somehow impaired. However, as Figure 8.4 reveals, Kusche and Greenberg (1983) revealed that the ability of children with hearing difficulties to gain an understanding of the concepts of good and bad develops or evolves more slowly than it does in children who can hear normally. Five-year-old children who cannot hear have not matured to the point at which they can understand these concepts as well as children who can hear normally. This difference in the evolution of the ability to understand these concepts is a maturational difference between the two groups of children and not an impairment in the ability of children who cannot hear to acquire these concepts. This indicates that a maturational-selection effect exists or that a maturational difference exists between these two groups of children in their ability to understand these two concepts. This maturational difference represents the most logical explanation for the difference observed in a study attempting to teach five-year-old children with and without hearing difficulties the concepts of good and bad and not a difference in ability to learn these concepts.

Differential Attrition Attrition refers to the fact that some individuals do not complete the outcome measures. This can occur for any of a variety of reasons, such as failure to show up at the scheduled time and place or not participating in all phases of the study. Differential attrition refers to a bias that occurs when the people who do not complete the outcome measures are different in the various comparison groups. In conducting a research study, the typical procedure is to identify the individuals who agree to participate and schedule them to participate at a specific time and place. However, people fail to show up for a study for a variety of reasons ranging from just forgetting about the study to deciding that they don't want to participate.

The loss of research participants from the research study does not, in and of itself, produce a bias. The bias results because the loss, when using a multigroup design, may produce differences between the comparison groups that cannot be attributed to the experimental treatment condition. To attribute differences between comparison groups to the experimental treatment, the comparison groups have to be the same on all variables except the independent variable. The loss of research participants from the various comparison groups can create differences on variables other than the independent variable, producing a mortality threat to the internal validity of the study.

For example, assume that you want to test the efficacy of a new technique for teaching mathematics to students. To test this new technique, you obtain permission from a school system to conduct the study. In conducting the study, one teacher teaches mathematics to one group of students by the traditional method and to another group of students by the new technique for one selected week of the school year. You know that research has revealed that older students perform better than younger students on mathematics, so you try to control for the age influence by having equal numbers of older and younger students in the two experimental groups. However, when you conduct the study, you find that half of the younger students assigned to the new technique group are sick and do not show up for class. When comparing the performance of the mathematical skills of the students in the two groups after the one week of instruction, you find that the new technique is superior to the traditional method. Can you conclude that the new technique is the superior teaching technique and that it should be used instead of the traditional method? This inference is incorrect because more older students were in the new technique group, and past research has indicated that older students do better in mathematics. This age difference and not the method of instruction may have produced the difference in the mathematical performance of the students in the two groups. If it did, then a differential attrition bias existed and represented a rival explanation for the performance difference observed in the two groups of students.

From this example, you can see that differential attrition is really a type of selection bias because attrition can result in the various comparison groups being composed of people that differ on variables other than the independent variable.

![Figure 8.4](image_url)

**FIGURE 8.4** The evolution of good and bad concepts as a function of age and hearing status

EXTERNAL VALIDITY

Extermal validity is a term coined by Campbell and Stanley (1963) and extended by Shadish, Cook, and Campbell (2002) to refer to the extent to which the results of a study can be generalized to and across populations of persons, settings, times, outcomes, and treatment variations. In Chapter 1, we state that one of the basic assumptions of science is that there are regularities in human behavior and these regularities can be discovered through systematic research. Whenever we conduct a research study, we are attempting to discover these regularities. However, each research study is conducted on a specific sample of individuals in a specific setting with a specific independent variable, with specific outcomes, and at a specific point in time.

To generalize the results of a study, you must identify a target group of individuals, settings, times, outcomes, and treatment variations and then randomly select from these populations so that you will have a sample representative of these populations. Most studies cannot randomly sample from the populations of individuals, settings, times, outcomes, and treatment variations because of the expense, time, and effort involved as well as the fact that the populations of outcomes and treatment variations are probably not known and cannot, therefore, be adequately sampled. Therefore, all studies contain characteristics that threaten their external validity. We will discuss each of these threats so that you can be aware of some of the factors that limit the generalizability of a study.

Population Validity

Population validity refers to the ability to generalize from the sample of individuals on which a study was conducted to the larger target population of individuals and across different subpopulations within the larger target population. The target population is the larger population, such as all children with a learning disability, to whom the research study results are to be generalized. Within this larger target population, there are many subpopulations, such as male and female children with a learning disability. Population validity, therefore, has the two components of generalizing from a sample to a target population and generalizing from a sample across types of persons in the target population.
The second step in the generalization process involves inferring from the accessible population to the target population. This is the generalization you want to make but the one that you can seldom make with any degree of confidence because the accessible population is seldom representative of the target population. For example, if the study that you conducted demonstrated that you had developed a method for improving the reading skills of children with a reading disability, you would ideally want the results of your study to generalize to all children with a reading disability. To be able to make such a statement, the sample of children participating in your study would have had to have been randomly selected from the target population, which is rarely possible. Therefore, you probably have to settle for randomly selecting from an accessible population such as a specific school or a specific school district. One school, or even a school district, is seldom representative of the target population. For example, the school at which you conduct your study may consist primarily of children from an impoverished area of the city in which you work. Although this school may have a large percentage of children with learning disabilities, the children are certainly not representative of the target population consisting of all children in the United States with learning disabilities. Yet it is to the target population that you want to generalize. As you can see, generalizing the results of a study to the target population is frequently a tenuous process because the sample of participants used in most studies are not randomly selected from the target population.

Most of this discussion of external validity focuses on generalizing to a specific target population. However, we should not forget that external validity also focuses on the goal of generalizing across populations. In any target population, there are many subpopulations, as we mentioned earlier. When we talk about generalizing across populations, we are really asking the question of whether the results hold for each of the subpopulations within the larger target population. Assume that we conducted a study investigating a specific treatment enhancing the reading ability of children with dyslexia. Assume further that we did randomly select 500 children from the target population of children with dyslexia in the United States and found that the treatment was effective. Because we had randomly selected our sample (something that seldom occurs in this type of study), we could generalize back to the target population and conclude that children with dyslexia would, on the average, benefit from the treatment program.

The results would not, however, say anything about the effectiveness of the treatment for the many subpopulations within the larger target population. Can the results be generalized to both male and female children with dyslexia, to children with dyslexia coming from various socioeconomic groups, age groups, intellectual levels, and so forth? This is the issue of generalizing across populations. In fact, many studies that are conducted to test the generalization of a specific treatment across populations, are attempting to identify the specific subpopulations to which a treatment can and cannot be generalized.

It is very important to understand the distinction between generalizing to and across populations. Survey research is the type of research that most often focuses on generalizing to target populations. Experimental and quasi-experimental research focuses more on generalizing across populations or identifying the populations to which a finding can and cannot be generalized.

**Ecological Validity**

*Ecological validity* refers to the ability to generalize the results of a study across settings. For example, one study might be conducted in a rural school in the South, using old computers that are slow and antiquated. If the results obtained from this study can be generalized to other settings, such as an urban school well equipped with state-of-the-art technology, then the study possesses ecological validity. Ecological validity therefore exists to the extent that the study results are independent of the setting in which the study was conducted.

One of the subtle setting factors that can affect the ability to generalize the results of a study is participant's knowledge of the fact that they are participating in a study. This is known as a reactivity effect. *Reactivity* refers to the alteration in performance that can occur as a result of being aware that one is participating in a research study. It is similar to the effect many people have of being on television: Once you know the camera is on you, you shift to your "television" behavior. A similar phenomenon can occur in research studies. Once you know you are in a research study, you might change your behavior. A reactivity effect can therefore threaten both the internal and external validity of a study.

**Temporal Validity**

*Temporal validity* refers to the extent to which the results of a study can be generalized across time. Temporal validity is an issue because most educational research studies are conducted during one time period. For example, Thokildsen, Nolen, and Poumier (1994) assessed children's views of several practices teachers use to influence motivation to learn. The data for this study were collected by interviewing 7- to 12-year-old children at one point in time. Although the data are valid for the time period in which they were collected, there is no assurance that the same results would hold true 10 years later. Frequently, it is assumed that the results of studies are invariant across time. Although this might exist for the results of some studies, it almost certainly does not exist for the results of others. Failure to consider the time variable can threaten the external validity of the study.

**Treatment Variation Validity**

*Treatment variation validity* refers to the ability to generalize the results across variations of the treatment. Treatment variation validity is an issue because the administration of a treatment can vary from one time to the next. For example, many studies have been conducted demonstrating that behavior therapy is effective in treating depression. However,
These studies have been typically conducted in the context of a research study that has provided maximum assurance that therapists were competent and delivered the therapy in the prescribed manner. Therapists who administer behavior therapy to the general public, however, vary considerably in their competency and the extent to which they deliver the therapy in the prescribed manner. This means that there is considerable variation in the way behavior therapy is administered. If behavior therapy produces a beneficial effect for the treatment of depression across these different variations in the way it is delivered, treatment generalization exists. If behavior therapy is beneficial only when administered exactly as prescribed by a competently trained therapist, then treatment generalization does not exist.

**Outcome Validity**

Outcome validity refers to the ability to generalize the results across different but related dependent variables. Many studies investigate the effect of some independent variable on one or more dependent variables. Outcome validity refers to the extent to which the independent variable influences a number of related outcome measures. For example, a job-training program is expected to increase the likelihood of getting a job after graduation. This is probably the primary outcome measure of interest. However, an equally important issue is maintaining the job. This means that the person must arrive on time, not miss work, and follow orders as well as demonstrate an acceptable level of performance. The effectiveness of the job-training program might increase the probability of getting a job but have no effect on job retention because it has little impact on these other essential adaptive job skills. Sometimes one outcome measure demonstrates that the treatment was effective. However, other outcome measures show no effect and maybe even a negative effect. Using several outcome measures is always desirable because this gives a more complete picture of the overall effect of the treatment. Fortunately, this is one of the easier design features to implement. You just need to include several related dependent variables in your study to answer questions about generalizability over outcomes.

**Construct Representation**

Up to this point in the chapter, we have discussed issues, such as internal validity, that are related to the validity of the design of an educational research study. Any educational research study involves the investigation of a set of variables such as televised instruction, the education of culturally diverse students, or the effect of stress on academic achievement. Additionally, we frequently want to conduct a study on a specific population of individuals such as children with attention deficit hyperactive disorder or dyslexia. Conducting a research study on variables or special populations such as these requires that they be assessed or measured. This creates some difficulty because many of the variables or characteristics of the special populations of interest represent abstract constructs. The educational researcher is faced with the task of identifying or devising some way of representing the constructs being investigated. This is a problem of construct validity. **Construct validity** refers to the extent to which a higher-order construct, such as help seeking, teacher stress, or dyslexia, is accurately represented in the particular study.

So how do we achieve construct validity? Construct validity is fostered by having a good definition and explanation of the meaning of the construct of interest. However, every construct, such as violence, has multiple features, and this creates a difficulty in identifying the prototypical features of a construct. For example, just hurting someone does not qualify as aggression or violence. There must be intent to harm. This problem is exacerbated in educational research because the explanation of the prototypical features of many of the constructs we investigate has not occurred. This difficulty is partly because of the abstract nature of the constructs with which we work. Educational research focuses on issues such as intelligence, burnout, at-risk students, violence, abuse of testing in schools, lack of AIDS education, and stress. These are constructs that are hard to define precisely. Because of the abstract nature of many of the constructs we work with and the lack of a clear explanation of their meaning, there is typically an imperfect relationship between the way a construct is represented or measured in a research study and the higher-order construct we want to represent.

The multiple features of any construct and the lack of a clear explanation of the prototypical features of a construct create difficulty for researchers when they try to represent constructs of interest in a research study. What is the researcher to do? The researcher must make use of the available knowledge and measures of the construct he or she is investigating and identify the specific way in which a construct will be represented in the study he or she is conducting. This is where the concept of operationalism enters and is useful as a communication tool for the researcher. **Operationalism** means that terms or constructs are represented by a specific set of steps or operations. For example, if stress is measured by the Stress in Teaching Questionnaire, then stress is represented by a score on this questionnaire. However, this set of operations might or might not accurately represent the construct of interest in the study because of the presence of any of a number of threats to construct validity. We are not going to discuss these threats because they are too remote beyond the scope of this textbook. We have, however, listed them on the book's companion website at www.ablongman.com/johnsonchristensen2e under bonus material. We do want you to realize that the validity of our inferences about constructs makes up the existence or nonexistence of such threats.

Operationalism is the technique that most researchers use when conducting their studies. They select a specific operation or set of operations as their representation of the construct they are investigating. Manthey and Gilmore (1996) used the Stress in Teaching Questionnaire as their representation of teacher stress. Butler and Neuman's (1993) representation of help-seeking behaviors in second- and sixth-grade children was whether they asked the experimenter for assistance in solving puzzles.
Identifying the specific operations used to represent a construct is very convenient and is essential for communication. For example, Butler and Neuman’s (1995) operationization (i.e., their construct representation) of help seeking left little room for interpretation or question as to the way in which they had conceptualized and interpreted help-seeking. Note, however, that this is not the only way in which a person can seek help. There can also be obtained by asking peers for help or going to the library and looking for reference material that would provide assistance. The important point to remember is that specification of a set of operations is for accuracy in communication. This is the beauty of operationizations. They specify, in a concrete and specific way, how a construct is conceptualized and measured in a given study. This degree of specificity permits exact communication of the construct and allows anyone else to repeat the steps and represent the construct in the same way.

Although the operationalization of constructs is necessary for educational researchers to communicate the way in which a construct is represented in a given research study, remember that seldom, if ever, does a given operationalization completely represent the construct being investigated. Consider, for example, the study by Mantleb and Gilmore (1996) in which they operationally represented teacher stress as the response teachers provided on the Stress in Teaching Questionnaire. Although this questionnaire probably does measure some component of teacher stress, it would be foolish to assume that this single measure provides a completely accurate representation of the construct of teacher stress. Rather, stress of any type probably includes physiological reactions such as changes in heart rate and blood pressure as well as behavioral changes such as being less tolerant of the students, both of which are probably not adequately assessed by a questionnaire. Additionally, Campbell (1985) has pointed out that every observation is affected by factors that bear no relation to the construct that is being measured. For example, the Stress in Teaching Questionnaire probably measures, in part, stressful teacher stress. However, it is also a function of events that are irrelevant to the stress that occurs as a result of teaching, such as the type of questions asked, the interpretation of the questions by the teacher completing the questionnaire, the tolerance a teacher has for stress, and stress factors influencing the teacher that are not related to the profession of teaching.

The important point to remember is that there are many different ways of operationally representing a construct and that each operationalization represents only a portion of the construct. The most accurate representation of a construct would involve measuring it in several different ways. For example, teacher stress could be measured by a questionnaire, by the teacher’s reaction to students, and by having others rate or identify factors influencing teacher stress. As more and more measures of the same construct are included, the probability of obtaining a more complete representation of the construct increases. The use of multiple measures of a construct is called multiple operationalism, which is the recommended approach to use in research studies (Campbell, 1988).

We also must point out that it is not sufficient to specify a set of operations as a representation of a construct and then assume that this is a valid measure of the intended construct or even some component of the construct of interest. Any set of operations does not have any necessary relationship to the construct being investigated. All operations do have some specific or concrete terms the way in which the construct is being represented in a particular study. To drive this point home with a ridiculous example, assume that you want to investigate the effect of intelligence on learning. In this study, you operationally represent intelligence as a person’s income on the assumption that more intelligent people make more money. This is obviously a poor representation of the construct of intelligence. However, it is operationialized in specific and concrete terms. This communicates how intelligence was represented in this study. However, it is a very poor measure and therefore a poor way of representing intelligence.

8.22 What is construct validity, and how is it achieved?
8.23 What is operationalism, and what is its purpose?
8.24 What is multiple operationalism, and why is it used?

RESEARCH VALIDITY IN QUALITATIVE RESEARCH

Discussions of the term validity have traditionally been attached to the quantitative research tradition. Not surprisingly, reactions by qualitative researchers have been mixed regarding whether or not this concept should be applied to qualitative research. At the extreme, some qualitative researchers have suggested that the traditional quantitative criteria of reliability and validity are not relevant to qualitative research (e.g., Smith, 1984). Smith contends that the basic assumptions of quantitative and qualitative research are incompatible and that the concepts of reliability and validity should therefore be abandoned. Most qualitative researchers, however, do not hold this viewpoint, and neither do we. Most qualitative researchers argue that some qualitative research studies are better than others, and they frequently use the term validity to refer to this difference. When qualitative researchers speak of research validity, they usually refer to qualitative research that is plausible, credible, trustworthy, and therefore defensible. We believe it is important to think about the issue of validity in qualitative research and to examine some strategies that have been developed to maximize validity. A list of these strategies is provided in Table 8.2. Keep in mind that most of these strategies can also be used in quantitative research.

One potential threat to validity that researchers must be careful to watch out for is called researcher bias. This problem is summed up in a statement a colleague of ours once made. She said, “The problem with qualitative research is that the researchers find what they want to find, and then they write up their results.” It is true that the problem of researcher bias is frequently an issue in qualitative research because qualitative research tends to be exploratory and is open-ended and less structured than quantitative research. (One would be remiss, however, to think that researcher bias is never a problem in quantitative research.) Researcher bias tends to result from selective observation and selective recording of information and also from allowing one’s personal views and perspectives to affect how data are interpreted and how the research is conducted.

The key strategy that is used to understand researcher bias is called reflexivity, which means that the researcher actively engages in critical self-reflection about his or her potential biases and predispositions (Table 8.2). Through reflexivity, researchers become more self-aware, and they monitor and attempt to control their biases. Many qualitative researchers
## Table 8.2 Strategies Used to Promote Qualitative Research Validity

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Researcher-as-detective</td>
<td>A metaphor characterizing the qualitative researcher as he or she searches for evidence about causes and effects. The researcher develops an understanding of the data through careful consideration of potential causes and effects and by systematically eliminating rival explanations or hypotheses until the final case is made beyond a reasonable doubt. The detective can utilize any of the strategies listed here.</td>
</tr>
<tr>
<td>Extended fieldwork</td>
<td>To provide for both discovery and validation researchers should collect data in the field over an extended time period.</td>
</tr>
<tr>
<td>Low-inference descriptors</td>
<td>The use of description phrased very close to the participants' accounts and researchers' field notes. Verbatim (i.e., direct quotations) are a commonly used type of low-inference descriptors.</td>
</tr>
<tr>
<td>Triangulation</td>
<td>Cross-checking information and conclusions through the use of multiple procedures or sources. When the different procedures or sources are in agreement you have corroboration.</td>
</tr>
<tr>
<td>Data triangulation</td>
<td>The use of multiple data sources to help understand a phenomenon.</td>
</tr>
<tr>
<td>Methods triangulation</td>
<td>The use of multiple research methods to study a phenomenon.</td>
</tr>
<tr>
<td>Investigator triangulation</td>
<td>The use of multiple investigators (i.e., multiple researchers) in collecting, analyzing, and interpreting the data.</td>
</tr>
<tr>
<td>Theory triangulation</td>
<td>The use of multiple theories and perspectives to help interpret and explain the data.</td>
</tr>
<tr>
<td>Participant feedback</td>
<td>The feedback and discussion of the researcher’s interpretations and conclusions with the actual participants and other members of the participant community for verification and insight.</td>
</tr>
<tr>
<td>Peer review</td>
<td>Discussion of the researcher’s interpretations and conclusions with other people. This includes discussion with a disinterested peer (e.g., with another researcher not directly involved). This peer should be skeptical and play the devil’s advocate, challenging the researcher to provide solid evidence for any interpretations or conclusions. Discussion with peers who are familiar with the research can also help provide useful challenges and insights.</td>
</tr>
<tr>
<td>External audit</td>
<td>Using outside experts to assess the study quality.</td>
</tr>
<tr>
<td>Negative-case sampling</td>
<td>Locating and examining cases that disconfirm the researcher’s expectations and tentative explanation.</td>
</tr>
<tr>
<td>Reflexivity</td>
<td>Involves self-awareness and critical self-reflection by the researcher on his or her potential biases and predispositions as these may affect the research process and conclusions.</td>
</tr>
<tr>
<td>Pattern matching</td>
<td>Predicting a series of results that form a distinctive pattern and then determining the degree to which the actual results fit the predicted pattern or &quot;fingerprint.&quot;</td>
</tr>
</tbody>
</table>

**Descriptive Validity**

Descriptive validity refers to the factual accuracy of the account as reported by the researchers. The key questions that are addressed in descriptive validity are: Did what was reported as taking place in the group being studied actually happen? And did the researchers accurately report what they saw and heard? In other words, descriptive validity refers to accuracy in reporting descriptive information (description of events, objects, behaviors, people, settings, times, places, and so forth). This form of validity is important because description is a major objective in nearly all qualitative research.

One effective strategy used to obtain descriptive validity is investigator triangulation (Table 8.2). In the case of descriptive validity, investigator triangulation involves the use of multiple observers to record and describe the research participants’ behavior and the context in which they were located. The use of multiple observers allows cross-checking of observations to make sure the investigators agree about what took place. When corroborated (agreement) of observations across multiple investigators is obtained, it is less likely that outside reviewers of the research will question whether something occurred. As a result, the research will be more credible and defensible.

**Interpretive Validity**

Interpretive validity refers to accurately portraying the meaning attached by participants to what is being studied by the researcher. More specifically, it refers to the degree to which the research participants’ viewpoints, thoughts, feelings, intentions, and experiences are accurately understood by the qualitative researcher and portrayed in the research report. An important part of qualitative research is understanding research participants' "inner worlds" (i.e., their subjective worlds), and interpretive validity refers to the degree of accuracy in presenting these inner worlds. Accurate interpretive validity requires that the researcher get inside the heads of the participants, look through the participants’ eyes, and see and feel what they see and feel. In this way the qualitative researcher can understand things from the participants’ perspectives and thus provide a valid account of these perspectives.
Some strategies for achieving interpretive validity are provided in Table 8.2. Participant feedback (or member checking) is perhaps the most important strategy (see Table 8.2). By sharing your interpretations of participants' viewpoints with the participants and other members of the group, you may clear up areas of miscommunication. Do the people being studied agree with what you have said about them? Although this strategy is not perfect, because some participants may attempt to put on a good face, useful information is frequently obtained and inaccuracies are often identified.

When writing the research report, using many low-inference descriptors is also helpful so that the reader can experience the participants' actual language, dialect, and personal meanings (Table 8.2). In this way, the reader can hear how the participants think and feel about issues and experiences. A verbatim is the lowest inference descriptor of all because the participants' exact words are provided in direct quotations. Here is an example of a verbatim from a high school dropout who was part of an ethnographic study of high school dropouts:

I wouldn't do the work. I didn't like the teacher and I didn't like my Mom and Dad. So, even if I did my work, I wouldn't turn it in. I completed it. I just didn't want to turn it in. I was angry with my Mom and Dad because they were talking about moving out of state at the time. (Okey & Casick, 1995, p. 257)

This verbatim provides some description (i.e., what the participant did), but it also provides some information about the participant's interpretations and personal meanings (which is the topic of interpretive validity). The participant expresses his frustration and anger toward his parents and teacher and shares with us what homework meant to him at this time and why he acted as he did. By reading verbatim like this one, readers of a report can experience for themselves the participants' perspectives. Again, getting into the minds of research participants is a common goal in qualitative research, and Maxwell calls our accuracy in portraying this "inner content" interpretive validity.

Theoretical Validity

Theoretical validity refers to the degree to which a theoretical explanation developed from a research study fits the data and is therefore credible and defensible. As we discuss in Chapter 1, theory usually refers to discussions of how a phenomenon operates and why it operates as it does. Theory is usually more abstract and less concrete than description and interpretation. Theory development moves beyond "just the facts" and provides an explanation of the phenomenon. In the words of Joseph Maxwell (1992):

one could label the student's throwing of the eraser as an act of resistance, and connect this act to the repressive behavior or values of the teacher, the social structure of the school, and class relationships in U.S. society. The identification of the throwing as "resistance" constitutes the application of a theoretical construct... the connection of this to other aspects of the participants, the school, or the community constitutes the postulation of theoretical relationships among these constructs. (p. 291)

In this example, the theoretical construct called resistance is used to explain the student's behavior. Maxwell points out that the construct of resistance may also be related to other theoretical constructs or variables. In fact, theories are often developed by relating theoretical constructs.

A strategy for promoting theoretical validity is extended fieldwork (Table 8.2), which means that you should spend a sufficient amount of time studying your research participants and their setting so that you can have confidence that the patterns of relationships you believe are operating are stable and so you can understand why these relationships occur. As you spend more time in the field collecting data and generating and testing your interpretations, your theoretical explanation might become more detailed and intricate. You may decide to use the strategy called theoretical triangulation (Table 8.2; Denzin, 1989). This means that you would examine how the phenomenon being studied would be explained by different theories and perspectives. The various theories might provide you with insights and help you develop a more cogent explanation. In a related way, you might also use investigator triangulation and consider the ideas and explanations generated by additional researchers studying the research participants.

As you develop your theoretical explanation, you should make some predictions based on the theory and test the accuracy of those predictions. When doing this, you can use the pattern matching strategy (Table 8.2). In pattern matching, the strategy is made to verify predictions at once so that, if all of the predictions occur as predicted (i.e., if the pattern or "fingerprint" is found), you have evidence supporting your explanation. As you develop your theoretical explanation, you should also use the negative-case sampling strategy mentioned earlier (Table 8.2). That is, you must always search for cases or examples that do not fit your explanation so that you do not simply find data that support your developing theory. As a general rule, your final explanation should accurately reflect the majority of the people in your research study. Another useful strategy for promoting theoretical validity is called peer review (Table 8.2), which means that you should try to spend some time discussing your explanation with your colleagues so that they can identify any problems in it. Each problem must then be resolved. In some cases, you will find that you will need to go back to the field and collect additional data. Finally, when developing a theoretical explanation, you must also think about the issues of internal validity and external validity, to which we now turn.

Internal Validity

You are already familiar with internal validity, which is the fourth type of validity in qualitative research of interest to us. As you know, internal validity refers to the degree to which a researcher is justified in concluding that an observed relationship is causal. Often qualitative researchers are not interested in cause-and-effect relationships. Sometimes, however, qualitative researchers are interested in identifying potential causes and effects. In fact, qualitative research can be very helpful in describing how phenomena operate (i.e., studying process) and in developing and testing preliminary causal hypotheses and theories (Campbell, 1979; Johnson, 1984; LeCompte & Preissle, 1993; Strauss, 1995; Yin, 1994). However, after potential causal relationships are studied using qualitative research, they should be tested and confirmed by using experimental methods when this is feasible. In this way, more conclusive evidence about cause and effect can be obtained. When qualitative researchers identify potential cause-and-effect relationships, they must think about many of the same issues discussed earlier in this chapter when we talked...
about internal validity and about the strategies that are used for obtaining theoretical validity. The qualitative researcher takes on the role of the "detective" searching for the true root cause(s) of a phenomenon, examining each possible "clue," and attempting to rule out each rival explanation generated (see researcher-as-detective in Table 8.2). When trying to identify a causal relationship, the researcher makes mental comparisons. The comparison might be to a hypothetical control group. Although a control group is rarely used in qualitative research, the researcher can think about what would have happened if the causal factor had not occurred. The researcher can sometimes rely on his or her expert opinion, as well as published research studies, when available, in deciding what would have happened. Furthermore, if the event is something that should occur again the researcher can determine whether the causal factor precedes the outcome. In other words, when the causal factor occurs again, does the effect follow?

When a researcher believes that an observed relationship is causal, he or she must also attempt to make sure that the observed change in the dependent variable is due to the independent variable and not to something else (e.g., a confounding extraneous variable). The successful researcher will always make a list of rival explanations or rival hypotheses that are possible or plausible reasons for the relationship other than the originally suspected cause. One way to justify rival explanations is to be a skeptic and think of reasons why the relationship should not be causal. Each rival explanation must be examined after the list has been developed. Sometimes you will be able to check a rival explanation with the data you have already collected through additional data analysis. At other times, you will need to collect additional data. One strategy would be to observe the relationship you believe to be causal under conditions in which the confounding variable is not present and compare this outcome with the original outcome. For example, if you concluded that a teacher effectively maintained classroom discipline on a given day but a critic maintained that it was the result of a parent visiting the classroom on that day, then you should try to observe the teacher again when the parent is not present. If the teacher is still successful, you have some evidence that the original finding was not because of the presence of the parent in the classroom.

All the strategies shown in Table 8.2 are used to improve the internal validity of qualitative research. Now we will explain the only two strategies not yet discussed: methods triangulation and data triangulation. When using methods triangulation (Table 8.2), the researcher uses more than one method of research in a single research study. The word methods is used broadly here, and it refers to different methods of research (ethnography, correlation, experimental, and so forth) as well as different methods of data collection (e.g., interviews, questionnaires, focus groups, observations). You can internally any of these methods (e.g., ethnography and survey research methods, interviews and observations, or experimental research and interviews). The logic is to combine different methods that have nonoverlapping weaknesses and strengths. The weaknesses (and strengths) of one method will tend to be different from those of a different method, which means that when you combine two or more methods, you will have better evidence. In other words, the whole is better than its parts.

Here is an example of methods triangulation. Perhaps you are interested in why students in an elementary classroom stigmatize a certain student named Brian. A stigmatized student is an individual who is not well liked, has a lower status, and is seen as different from the "normal" students. Perhaps Brian has a different haircut from the other students, is dressed differently, or doesn't act like the other students. In this case, you might decide to observe how students treat Brian in various situations. In addition to observing the students, you will probably decide to interview Brian and the other students to understand their beliefs and feelings about Brian. A strength of observational data is that you can actually see students' behavior. A weakness of interviewing is that what the students say and what they actually do may be different. However, using interviews, you can delve into the students' thinking and reasoning, whereas you cannot do this using observational data. Therefore, the whole will likely be better than the parts.

When using data triangulation (Table 8.2), the researcher uses multiple data sources in a single research study. "Data sources" does not mean using different methods. Data triangulation refers to the use of multiple data sources using a single method. For example, the use of multiple interviews would provide multiple data sources while using a single method (i.e., the interview method). Likewise, the use of multiple observations is another example of data triangulation; multiple data sources would be provided while using a single method (i.e., the observational method). Another important part of data triangulation involves collecting data at different times, at different places, and with different people.

Here is an example of data triangulation. Perhaps a researcher is interested in studying why certain students are apathetic. It would make sense to get the perspectives of several different kinds of people. The researcher might interview teachers, interview students identified by the teachers as being apathetic, and interview peers of apathetic students. Then the researcher could check to see whether the information obtained from these different data sources was in agreement. Each data source may provide additional reasons as well as a different perspective on the question of student apathy, resulting in a more complete understanding of the phenomenon. The researcher should also interview apathetic students at different times during the year to see if there are different perspectives on the question of student apathy. A qualitative researcher can develop a better understanding of why students are apathetic if only one data source is used.

External Validity

As you know, external validity is important when you want to generalize from a set of research findings to other people, settings, times, treatments, and outcomes. Typically, generalizability is not the major purpose of qualitative research. There are at least two reasons for this view. First, the people and settings examined in qualitative research are rarely randomly selected, and as you know, random selection is the best way to generalize from a sample to a population. As a result, qualitative research is virtually always weak in the form of population validity focused on "generalizing to" populations.

Second, some qualitative researchers are more interested in documenting "particularistic" findings than "universalistic" findings. In other words, in certain forms of qualitative research, the goal is to describe a certain group of people or a certain event in a specific context, rather than to generate findings that are broadly applicable. At a fundamental level, many qualitative researchers do not believe in the presence of "general laws" or "universal
SOFTWARE

When we conduct a study, we develop a plan, outline, or strategy to use that will allow us to collect data that will lead to a valid conclusion. In any study, there are a number of extraneous variables that could systematically vary with the independent variable and confound the results making it impossible to assess the effect of the independent variable. To eliminate potentially confounding extraneous variables, we must design our study so that we can make valid inferences about the relationship between independent and dependent variables. In quantitative studies, there are four types of validity that are used to evaluate the accuracy of the inferences that can be made from the study results.

Statistical conclusion validity refers to the validity with which we can infer that two variables are related and the strength of that relationship. Internal validity refers to the validity with which we can infer that the relationship between two variables is causal. Causal relationship can be a causal descriptive relationship or a causal explanatory relationship.

To make this causal connection between the independent and dependent variables, we need evidence that they are related, that the direction of effect is from the independent variable (the cause) to the dependent variable (the effect), and that the observed effect on the dependent variable is due to the independent variable and not some extraneous variable. Internal validity is related to the ability to rule out the influence of extraneous variables. A study is internally valid when the effect observed on the dependent variable is due to the independent variable. However, there are many extraneous variables that can creep into a study and confound the results. The influence of these extraneous variables must be controlled or eliminated. A number of the more obvious threats to the internal validity of a study are the following:

1. Ambiguous temporal precedence—the inability to specify which variable preceded which other variable.
2. History—specific events, other than the independent variable, that occur between the first and second measurement of the dependent variable.
3. Maturation—the physical or mental changes that may occur in individuals over time such as aging, learning, boredom, hunger, and fatigue.