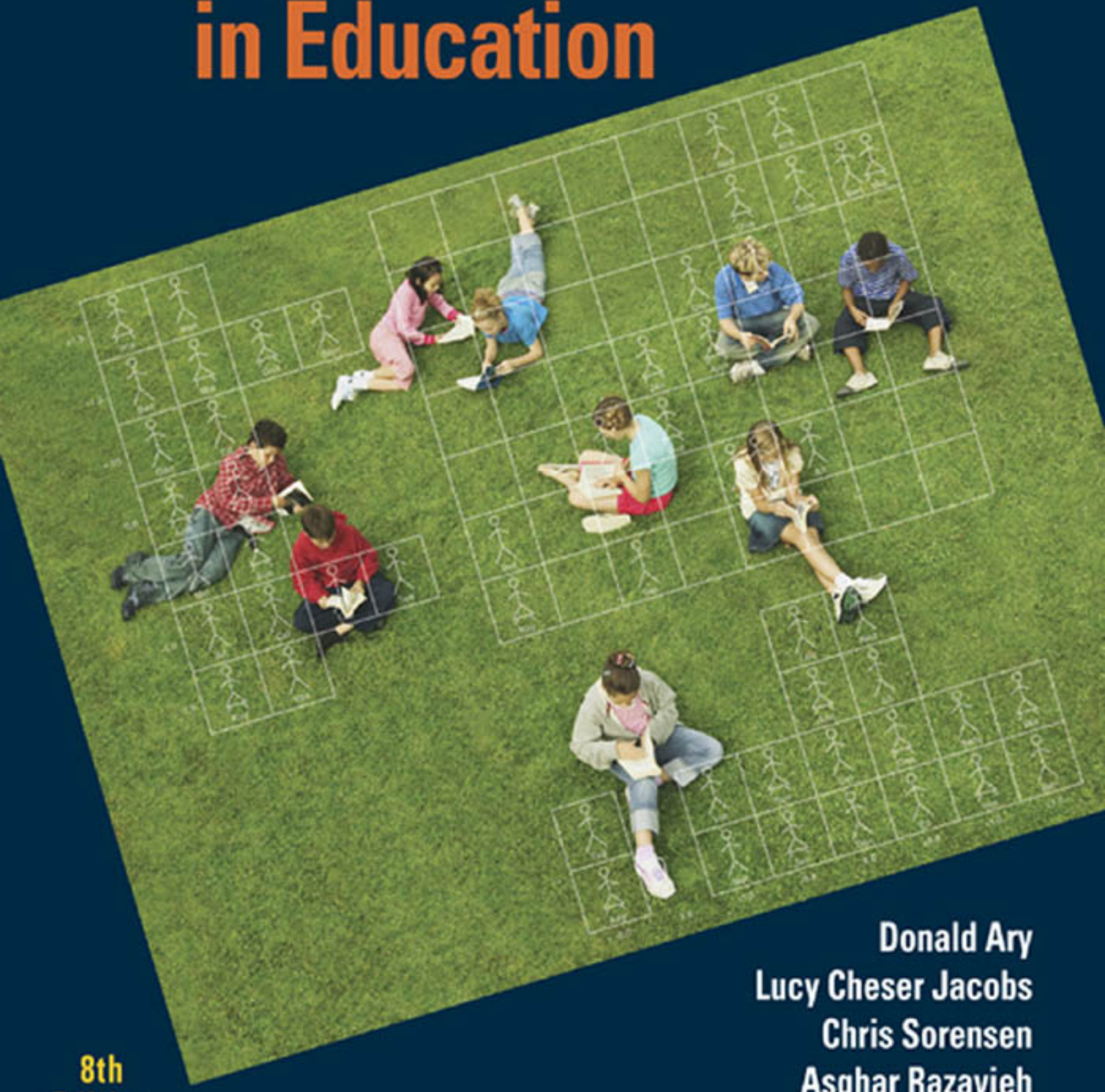
# Introduction to Research in Education



Donald Ary
Lucy Cheser Jacobs
Chris Sorensen
Asghar Razavieh

8th Edition

# Introduction to Research in Education

*This page intentionally left blank*

EIGHTH EDITION

# Introduction to Research in Education

DONALD ARY

LUCY CHESER JACOBS

CHRIS SORENSEN

WADSWORTH
CENGAGE Learning™

Australia • Brazil • Japan • Korea • Mexico
Singapore • Spain • United Kingdom • United States

WADSWORTH
CENGAGE Learning™

**Introduction to Research in Education**
**Eighth Edition**
**Donald Ary**
**Lucy Cheser Jacobs**
**Christine K. Sorensen**

Acquisitions Editor: Chris Shortt
Developmental Editor: Tangelique William
Assistant Editor: Caitlin Cox
Editorial Assistant: Linda Stewart
Associate Media Editor: Ashley Cronin
Marketing Manager: Kara Parsons
Marketing Assistant: Dimitri Hagnere
Marketing Communications Manager: Martha Pfeiffer
Content Project Manager: Samen Iqbal
Creative Director: Rob Hugel
Art Director: Maria Epes
Print Buyer: Karen Hunt
Rights Acquisitions Account Manager, Text: Tim Sisler
Rights Acquisitions Account Manager, Image: Mandy Groszko
Production Service: Aaron Downey, Matrix Productions, Inc.
Copy Editor: Dan Hays
Cover Designer: Bartay Studio
Cover Image: © Martin Barraud/Getty Images
Compositor: Integra



Library of Congress Control Number: 2008937509

Student Edition:

ISBN-13: 978-0-495-60122-7

ISBN-10: 0-495-60122-5

**Wadsworth**
10 Davis Drive
Belmont, CA 94002-3098
USA

Cengage Learning is a leading provider of customized learning solutions with office locations around the globe, including Singapore, the United Kingdom, Australia, Mexico, Brazil, and Japan. Locate your local office at **www.cengage.com/international.**

Cengage Learning products are represented in Canada by Nelson Education, Ltd.

To learn more about Wadsworth, visit **www.cengage.com/wadsworth**

Purchase any of our products at your local college store or at our preferred online store **www.ichapters.com.**

Printed in Canada
1 2 3 4 5 6 7 13 12 11 10 09

*To Sheila, Marion, and Steve*

*This page intentionally left blank*

# Contents

## *Part Three*
## Statistical Analysis

# Preface

Educational research is a vigorous, dynamic enterprise. We are amazed at not only how much more knowledge there is in the field but also how many new ways of seeking knowledge are included. At the same time, there is more technology to focus, simplify, and organize research in education.

Many changes have taken place since the first edition was published in 1972. Never once did the phrase "qualitative research" occur. The only research we included that is now classified as qualitative was historical research. The current edition includes four chapters on qualitative research, including a new one on the types of qualitative research and a chapter titled "Mixed Methods Research," which discusses combining quantitative and qualitative methods.

Not only is knowledge expanding but also tools that facilitate the pursuit of knowledge are expanding our capacity to generate new knowledge with greater precision and less effort. Among these tools are computers, the Internet, and indexing and abstracting databases. The latter necessitated a major rewrite of our chapter on searching for related literature. This endeavor, once a difficult to organize, time-consuming, imprecise task, can now be accomplished quickly, precisely, and completely through searching electronic databases.

Through all this we have endeavored to continue to present a text that is reader friendly and to make even advanced concepts understandable. We think the fact that this book lasted through seven editions so far is reasonable evidence that we have succeeded.


## ACKNOWLEDGMENTS

We are grateful to Pearson Education Ltd. on behalf of the literary executor of the late Sir Ronald A. Fisher, F.R.S., and the late Dr. Frank Yates, F.R.S., for permissions to reprint Tables III, IV, and VII from *Statistical Tables for Biological, Agricultural and Medical Research* (6th ed., 1974).

We greatly appreciate the professional and personal assistance of the staff at Cengage Learning, especially our developmental editor, Tangelique Williams. In addition, we thank Chris Shortt, Acquisitions Editor; Caitlin Cox, Assistant Editor; Diane Mars, Assistant Editor; Samen Iqbal, Production Manager; and Aaron Downey at Matrix Productions. We also appreciate the professional assistance of Jane Williams, of Guardian Academic Consulting of Sycamore, IL, and the faculty of the Reference Department of Founders' Library of Northern Illinois University. We would also like to thank Dr. Tom Christ for help with the Mixed Methods chapter.

We gratefully acknowledge the contributions of the following reviewers:

George Buck, University of Alberta
Elaine Bukowiecki, Bridgewater State College
Gail C. Delicio, Clemson University
Don Good, Milligan College
Chuck Okezie, Marygrove College
Jamis Perrett, Texas A & M University
Irvin Schonfeld, Columbia University
David Tan, University of Oklahoma

*This page intentionally left blank*

# Introduction to Research in Education

*This page intentionally left blank*

Knowledge is power.

CHAPTER 1

# The Nature of Scientific Inquiry

## INSTRUCTIONAL OBJECTIVES

After studying this chapter, the student will be able to:

1. List five major sources of knowledge and comment on the strengths and weaknesses of each source.
2. Describe the characteristics of the scientific approach.
3. State the assumptions underlying science and the attitudes expected of scientists.
4. Specify the purpose and characteristics of scientific theory in the behavioral sciences.
5. Indicate the limitations involved in the application of the scientific approach in the social sciences.
6. Define educational research and give examples.

Educators are, by necessity, decision makers. Daily they face the task of deciding how to plan learning experiences, teach and guide students, organize a school system, and a myriad other matters. Unlike unskilled workers, who are told what to do and how to do it, professionals must plan for themselves. People assume that professionals have the knowledge and skills necessary to make valid decisions about what to do and how. We generally define knowledge as justified true belief. How are educators to know what is true? How do they acquire reliable information? Although there are other sources of knowledge, such as experience, authority, and tradition, scientific knowledge about the educational process makes the most valuable contribution to decision making in education. Educators can turn to this source for reliable information and suggestions to be used in decision making. This fund of knowledge has been made available to educators by scientific inquiry into educational problems. However, education has not always been influenced by the results of such careful and systematic investigations. In fact, the development of an educational science is at a comparatively early stage

## SOURCES OF KNOWLEDGE

Before we further pursue the role of scientific inquiry in education, let us review some of the ways in which human beings throughout history have sought knowledge. The major sources of knowledge can be categorized under five headings: (1) experience, (2) authority, (3) deductive reasoning, (4) inductive reasoning, and (5) the scientific approach.

### EXPERIENCE

*Experience* is a familiar and well-used source of knowledge. After trying several routes from home to work, you learn which route takes the least time or is the most free of traffic or is the most scenic. By personal experience, you can find the answers to many of the questions you face. Much wisdom passed from generation to generation is the result of experience. If people were not able to profit from experience, progress would be severely retarded. In fact, this ability to learn from experience is a prime characteristic of intelligent behavior.

Yet for all its usefulness, experience has limitations as a source of knowledge. How you are affected by an event depends on who you are. Two people will have very different experiences in the same situation. The same forest that is a delightful sanctuary to one person may be a menacing wilderness to another. Two supervisors observing the same classroom at the same time could truthfully compile very different reports if one focused on and reported the things that went right and the other focused on and reported the things that went wrong.

Another shortcoming of experience is that you so frequently need to know things that you as an individual cannot learn by experience. A child turned loose to discover arithmetic alone might figure out how to add but would be unlikely to find an efficient way to compute square roots. A teacher could learn through experience the population of a classroom on a particular day but could not personally count the population of the United States.

### AUTHORITY

For things difficult or impossible to know by personal experience, people frequently turn to an *authority;* that is, they seek knowledge from someone who has had experience with the problem or has some other source of expertise. People accept as truth the word of recognized authorities. We go to a physician with health questions or to a stockbroker with questions about investments. To learn the size of the U.S. population, we can turn to reports by the U.S. Bureau of the Census. A student can look up the accepted pronunciation of a word in a dictionary. A superintendent can consult a lawyer about a legal problem at school. A beginning teacher asks an experienced one for suggestions and may try a certain technique for teaching reading because the teacher with experience suggests that it is effective.

Throughout history you can find examples of reliance on authority for knowledge, particularly during the Middle Ages when people preferred ancient scholars, such as Plato and Aristotle, and the early Fathers of the Church as sources of information—even over direct observation or experience. Although authority

**PICTURE THIS**


Joe Rocco

is a very useful source of knowledge, you must always ask, How does authority know? In earlier days, people assumed an authority was correct simply because of the position he or she held, such as king, chief, or high priest. Today, people are reluctant to rely on an individual as an authority merely because of position or rank. They are inclined to accept the assertions of an authority only when that authority is indeed a recognized expert in the area.

Closely related to authority are *custom* and *tradition*, on which people depend for answers to many questions related to professional as well as everyday problems. In other words, people often ask, "How has this been done in the past?" and then use the answer as a guide for action. Custom and tradition have been prominent influences in the school setting, where educators often rely on past practices as a dependable guide. However, an examination of the history of education reveals that many traditions that prevailed for years were later found to be erroneous and had to be rejected. For generations, it was considered good practice to humiliate students who made mistakes with dunce caps and the like. It is wise to appraise custom and tradition carefully before you accept them as reliable sources.

Authority is a quick and easy source of knowledge. However, as a source of knowledge, authority has shortcomings that you must consider. First, authorities can be wrong. People often claim to be experts in a field when they do not really have the knowledge to back up the claim. Second, you may find that authorities disagree among themselves on issues, indicating that their authoritative statements are often more personal opinion than fact.

## DEDUCTIVE REASONING

Ancient Greek philosophers made perhaps the first significant contribution to the development of a systematic approach for gaining knowledge. Aristotle and his followers introduced the use of **deductive reasoning,** which can be described as a thinking process in which one proceeds from general to specific knowledge through logical argument. An argument consists of a number of statements standing in relation to one another. The final statement is the conclusion, and the rest, called *premises*, offer supporting evidence. A major kind of deductive reasoning is the syllogism. A syllogism consists of a major premise and a minor premise followed by a conclusion. For example, "All men are mortal" (major premise); "The king is a man" (minor premise); "Therefore, the king is mortal" (conclusion). In deductive reasoning, if the premises are true, the conclusion is necessarily true. Deductive reasoning lets you organize premises into patterns that provide conclusive evidence for a conclusion's validity. Mystery fans will recall that Sherlock Holmes frequently would say, "I deduce …" as he combined previously unconnected facts in such a way as to imply a previously unsuspected conclusion.

Deductive reasoning can answer the question, "How likely is it that a student could pass a 20-item multiple choice test with five options per item by chance alone?" Given the premise that there is a 20 percent chance of getting a single item right and an 80 percent chance of getting it wrong and the premise that these same chances are true for every item, Figure 1.1 shows the probability of getting the following outcomes with three items.

The probability of getting three right is .008. There are three ways to get two right and one wrong, so the probability of two right is (.032)(3) = .096. The probability of getting one right and two wrong is (.128)(3) = .384. There is only one way to get three wrong; the probability of that is .512.

If we extended Figure 1.1 to determine the likelihood of getting a passing 60 percent (12 correct items in a 20-item test), we would find there is approximately one chance in 10,000 of passing. The probability of passing two 20-item tests is $(1/10{,}000)^2$ or one chance in 100 million. The notion that one has a reasonable chance of passing a test through sheer guessing is a myth.

Deductive reasoning has its limitations. To arrive at true conclusions, you must begin with true premises. The conclusion of a syllogism can never exceed the content of the premises. Because deductive conclusions are necessarily elaborations on previously existing knowledge, you cannot conduct scientific inquiry through deductive reasoning alone because it is difficult to establish the universal truth of many statements dealing with scientific phenomena. Deductive reasoning can organize what people already know and can point out new relationships as you proceed from the general to the specific, but it is not sufficient as

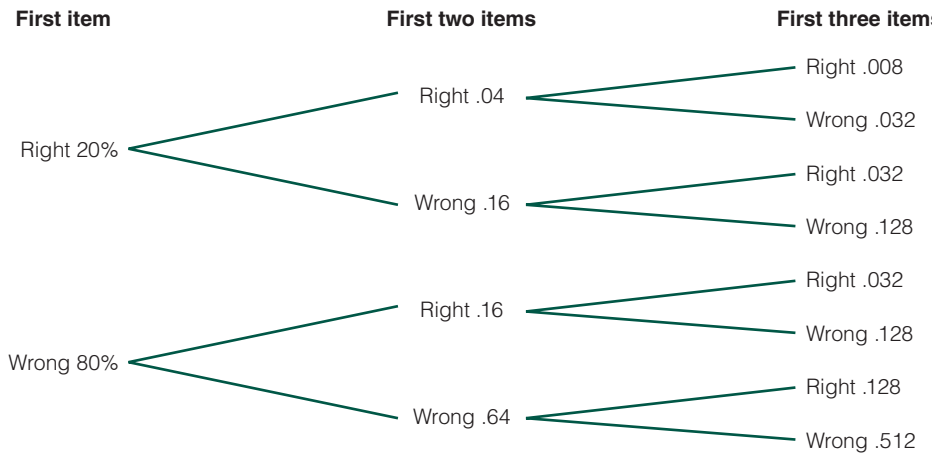



**Figure 1.1** Probabilities of Getting Various Outcomes with Three Items

a source of new knowledge. Despite its limitations, deductive reasoning is useful in research because it provides a way to link theory and observation. It lets researchers deduce from existing theory what phenomena they should observe. Deductions from theory can help build hypotheses, which are a vital part of scientific inquiry.

## INDUCTIVE REASONING

As noted previously, the conclusions of deductive reasoning are true only if the premises on which they are based are true. But how are you to know if the premises are true? In the Middle Ages, people often substituted dogma for true premises, so they reached invalid conclusions. It was Francis Bacon (1561–1626) who first called for a new approach to knowing. He held that thinkers should not enslave themselves by accepting premises handed down by authority as absolute truth. He believed that an investigator should establish general conclusions on the basis of facts gathered through direct observation. Bacon advised the seeker of truth to observe nature directly and to rid his or her mind of prejudice and preconceived ideas, which Bacon called "idols." For him, obtaining knowledge required that the thinker observe nature itself, gather particular facts, and formulate generalizations from these findings. You can see the importance of observation in the following anecdote (probably apocryphal), attributed to Bacon:

> In the year of our Lord 1432, there arose a grievous quarrel among the brethren over the number of teeth in the mouth of a horse. For 13 days the disputation raged without ceasing. All the ancient books and chronicles were fetched out, and wonderful and ponderous erudition, such as was never before heard of in this region, was made manifest. At the beginning of the 14th day, a youthful friar of goodly bearing asked his learned superiors for permission to add a word, and straightway, to the wonderment of the disputants, whose deep wisdom he sore vexed, he beseeched them to unbend in a manner coarse and unheard-of, and to look in the open mouth of a horse and find an answer to their questionings. At this, their dignity being grievously hurt, they waxed exceedingly wroth; and, joining in a mighty

> uproar, they flew upon him and smote him hip and thigh, and cast him out forthwith. For, said they, surely Satan hath tempted this bold neophyte to declare unholy and unheard-of ways of finding truth contrary to all the teachings of the fathers. After many days more of grievous strife the dove of peace sat on the assembly, and they as one man, declaring the problem to be an everlasting mystery because of a grievous dearth of historical and theological evidence thereof, so ordered the same writ down. (Mees, 1934, p. 115)

The youth in this story was calling for a new way of seeking truth: namely, seeking the facts rather than depending on authority or on sheer speculation. This became the fundamental principle of all science.

In Bacon's system, the investigator made observations on particular events in a class (or category) and then, on the basis of the observed events, made inferences about the whole class. This approach, known as **inductive reasoning**, is the reverse of the deductive method. You can see the difference between deductive and inductive reasoning in the following examples:

*Deductive:* Every mammal has lungs.
All rabbits are mammals.
Therefore, every rabbit has lungs.

*Inductive:* Every rabbit that has ever been observed has lungs.
Therefore, every rabbit has lungs.

Note that in deductive reasoning you must know the premises before you can reach a conclusion, but in inductive reasoning you reach a conclusion by observing examples and generalizing from the examples to the whole class or category. To be absolutely certain of an inductive conclusion, the investigator must observe all examples. This is known as **perfect induction** under the Baconian system; it requires that the investigator examine every example of a phenomenon. In the preceding example, to be absolutely sure that every rabbit has lungs, the investigator would have to have observations on all rabbits currently alive, as well as all past and future rabbits. Clearly, this is not feasible; you generally must rely on imperfect induction based on incomplete observation.

**Imperfect induction** is a system in which you observe a sample of a group and infer from the sample what is characteristic of the entire group. An example of a conclusion based on imperfect induction is the present thinking concerning the physical characteristics of very intelligent children. For many years, people generally believed that exceptionally bright children tended to be poor physical specimens. Even today, cartoonists usually portray the bright child as a scrawny creature with thick spectacles. Terman, a pioneer in the field of mental testing, was interested in the characteristics of exceptionally bright youngsters (Terman, 1926). In a landmark investigation, Terman intensively studied more than 1000 California children who scored higher than 140 on the Stanford–Binet intelligence test. He found the average height, weight, and general physical health of these children to be slightly above average for children of their age. From this and subsequent studies of the phenomenon, researchers have concluded that bright children, far from being scrawny, are slightly more likely to be above average in physical development than children with average IQ scores. Note that this conclusion has not been positively proved. It is simply highly probable. To

be positively sure about this conclusion, you would need physical measures for *all* children with IQ scores of 140 or higher on the Stanford–Binet. Even then, you could only be positive about the characteristics of such children today and could not be 100 percent sure that the same would be true of such children in the future. Although imperfect induction does not lead to infallible conclusions, it can provide reliable information about what is likely to be true and on which you can make reasonable decisions.

An inductive way to investigate the question, "Should you stick with your original answers on a multiple-choice test, or should you change your answers when, upon reconsideration, you think you have a better answer?" would be to go over scored exams and identify items with erasures or cross-outs. Then count the changes that go from right to wrong, wrong to right, or wrong to wrong.

Dozens of researchers have published the results of such studies, beginning with Crawford (1928). These studies have all found that more changes are from wrong to right than from right to wrong. Waddell and Blankenship (1994), through a thorough search of the literature for the years 1988–1992, found 61 studies whose results could be combined through meta-analysis (see Chapter 6). The combined results were as follows: 57 percent of changes were from wrong to right, 21 percent were from right to wrong, and 22 percent were from wrong to wrong. Therefore, the best advice is to encourage students to make changes whenever, after rethinking, they find an answer that they prefer over their original one. It is interesting to note that those studies that also asked students and professors their advice found the majority advised sticking with your original answer. The myth that you should stick with your original answer has persisted for generations, despite overwhelming evidence to the contrary.

> It's not so much what folks don't know that causes problems.
> It's what they know that ain't so.
>
> Artemus Ward

## THE SCIENTIFIC APPROACH

Exclusive use of induction often resulted in the accumulation of isolated knowledge and information that made little contribution to the advancement of knowledge. Furthermore, people found that many problems could not be solved by induction alone. In the 19th century, scholars began to integrate the most important aspects of the inductive and deductive methods into a new technique, namely the inductive–deductive method, or the **scientific approach.** This approach differs from inductive reasoning in that it uses hypotheses. A **hypothesis** is a statement describing relationships among variables that is tentatively assumed to be true. It identifies observations to be made to investigate a question.

For example, a researcher interested in increasing student on-task behavior might hypothesize that positive teacher feedback increases on-task behavior. All hypotheses indicate specific phenomena to be observed (the variables), in this case positive teacher feedback and on-task behavior.

Charles Darwin, in developing his theory of evolution, is generally recognized as the first to apply this method in the pursuit of knowledge. Darwin reported that he spent a long time making biological observations, hoping to establish some

generalizations concerning evolution. In the following passage, he describes how he arrived at a new approach:

> My first note-book (on evolution) was opened in July 1837. I worked on true Baconian principles, and without any theory collected facts on a wholesale scale, more especially with respect to domesticated productions, by printed enquiries, by conversation with skillful breeders and gardeners, and by extensive reading. When I see the list of books of all kinds which I read and abstracted, including whole series of Journals and Transactions, I am surprised at my industry. I soon perceived that selection was the keystone of man's success in making useful races of animals and plants. But how selection would be applied to organisms living in a state of nature remained for some time a mystery to me. In October 1838, that is, fifteen months after I had begun my systematic enquiry, I happened to read for amusement "Malthus on Population," and being well prepared to appreciate the struggle for existence which everywhere goes on from long-continued observation of the habits of animals and plants, it at once struck me that under these circumstances favourable variations would tend to be preserved, and unfavourable ones to be destroyed. The result of this would be the formation of new species. Here then I had at last got a theory by which to work. (Darwin, 2007, p. 68)

Darwin's procedure, involving only observation, was unproductive until reading and further thought led him to formulate a tentative hypothesis to explain the facts that he had gathered through observation. He then proceeded to test this hypothesis by making deductions from it and gathering additional data to determine whether these data would support the hypothesis. From this method of inquiry, Darwin was able to develop his theory of evolution. This use of both inductive and deductive reasoning is characteristic of modern scientific inquiry.

The scientific approach is generally described as a method of acquiring knowledge in which investigators move inductively from their observations to hypotheses and then deductively from the hypotheses to the logical implications of the hypotheses. They deduce the consequences that would follow if a hypothesized relationship were valid. If the deduced implications are compatible with the organized body of accepted knowledge, researchers then further test them by gathering empirical data. On the basis of the evidence, they accept or reject the hypotheses.

The use of hypotheses is the principal difference between the scientific approach and inductive reasoning. In inductive reasoning, you make observations first and then organize the information gained. In the scientific approach, you reason what you would find if a hypothesis were true and then you make systematic observations to confirm (or fail to confirm) the hypothesis.

## AN EXAMPLE OF THE SCIENTIFIC APPROACH

In a classic example, award-winning author Robert Pirsig provides a vivid and succinct description of the scientific approach by comparing it to the process of maintaining a motorcycle in good working order:

> Two kinds of logic are used, inductive and deductive. Inductive inferences start with observations of the machine and arrive at general conclusions. For example, if the cycle goes over a bump and the engine misfires, and then goes over another

bump and the engine misfires, and then goes over another bump and the engine misfires, and then goes over a long smooth stretch of road and there is no misfiring, and then goes over a fourth bump and the engine misfires again, one can logically conclude that the misfiring is caused by the bumps. That is induction: reasoning from particular experiences to general truths.

Deductive inferences do the reverse. They start with general knowledge and predict a specific observation. For example, if, from reading the hierarchy of facts about the machine, the mechanic knows the horn of the cycle is powered exclusively by electricity from the battery, then he can logically infer that if the battery is dead the horn will not work. That is deduction.

Solution of problems too complicated for common sense to solve is achieved by long strings of mixed inductive and deductive inferences that weave back and forth between the observed machine and the mental hierarchy of the machine found in the manuals. The correct program for this interweaving is formalized as scientific method.

Actually I've never seen a cycle-maintenance problem complex enough really to require full-scale formal scientific method. Repair problems are not that hard. When I think of formal scientific method an image sometimes comes to mind of an enormous juggernaut, a huge bulldozer—slow, tedious, lumbering, laborious, but invincible. It takes twice as long, five times as long, maybe a dozen times as long as informal mechanic's techniques, but you know in the end you're going to *get* it. There's no fault isolation problem in motorcycle maintenance that can stand up to it. When you've hit a really tough one, tried everything, racked your brain and nothing works, and you know that this time Nature has really decided to be difficult, you say, "Okay, Nature, that's the end of the nice guy," and you crank up the formal scientific method.

For this you keep a lab notebook. Everything gets written down, formally, so that you know at all times where you are, where you've been, where you're going, and where you want to get. In scientific work and electronics technology this is necessary because otherwise the problems get so complex you get lost in them and confused and forget what you know and what you don't know and have to give up. In cycle maintenance things are not that involved, but when confusion starts it's a good idea to hold it down by making everything formal and exact. Sometimes just the act of writing down the problems straightens out your head as to what they really are.

The logical statements entered into the notebook are broken down into six categories: (1) statement of the problem, (2) hypotheses as to the cause of the problem, (3) experiments designed to test each hypothesis, (4) predicted results of the experiments, (5) observed results of the experiments, and (6) conclusions from the results of the experiments. This is not different from the formal arrangement of many college and high school lab notebooks but the purpose here is no longer just busywork. The purpose now is precise guidance of thoughts that will fail if they are not accurate.

The real purpose of scientific method is to make sure Nature hasn't misled you into thinking you know something you don't actually know. There's not a mechanic or scientist or technician alive who hasn't suffered from that one so much that he's not instinctively on guard. That's the main reason why so much scientific and mechanical information sounds so dull and so cautious. If you get careless or go

romanticizing scientific information, giving it a flourish here and there, Nature will soon make a complete fool out of you. It does it often enough anyway even when you don't give it opportunities. One must be extremely careful and rigidly logical when dealing with Nature: one logical slip and an entire scientific edifice comes tumbling down. One false deduction about the machine and you can get hung up indefinitely.

In Part One of formal scientific method, which is the statement of the problem, the main skill is in stating absolutely no more than you are positive you know. It is much better to enter a statement "Solve Problem: Why doesn't cycle work?" which sounds dumb but is correct, than it is to enter a statement "Solve Problem: What is wrong with the electrical system?" when you don't absolutely know the trouble is in the electrical system. What you should state is "Solve Problem: What is wrong with cycle?" and then state as the first entry of Part Two: "Hypothesis Number One: The trouble is in the electrical system." You think of as many hypotheses as you can, then you design experiments to test them to see which are true and which are false.

This careful approach to the beginning questions keeps you from taking a major wrong turn which might cause you weeks of extra work or can even hang you up completely. Scientific questions often have a surface appearance of dumbness for this reason. They are asked in order to prevent dumb mistakes later on.

Part Three, that part of formal scientific method called experimentation, is sometimes thought of by romantics as all of science itself because that's the only part with much visual surface. They see lots of test tubes and bizarre equipment and people running around making discoveries. They do not see the experiment as part of a larger intellectual process and so they often confuse experiments with demonstrations, which look the same. A man conducting a gee-whiz science show with fifty thousand dollars' worth of Frankenstein equipment is not doing anything scientific if he knows beforehand what the results of his efforts are going to be. A motorcycle mechanic, on the other hand, who honks the horn to see if the battery works is informally conducting a true scientific experiment. He is testing a hypothesis by putting the question to Nature. The TV scientist who mutters sadly, "The experiment is a failure; we have failed to achieve what we had hoped for," is suffering mainly from a bad scriptwriter. An experiment is never a failure solely because it fails to achieve predicted results. An experiment is a failure only when it also fails adequately to test the hypothesis in question, when the data it produces don't prove anything one way or another.

Skill at this point consists of using experiments that test only the hypothesis in question, nothing less, nothing more. If the horn honks, and the mechanic concludes that the whole electrical system is working, he is in deep trouble. He has reached an illogical conclusion. The honking horn only tells him that the battery and horn are working. To design an experiment properly he has to think very rigidly in terms of what directly causes what. This you know from the hierarchy.

The horn doesn't make the cycle go. Neither does the battery, except in a very indirect way. The point at which the electrical system directly causes the engine to fire is at the spark plugs, and if you don't test here, at the output of the electrical system, you will never really know whether the failure is electrical or not.

To test properly, the mechanic removes the plug and lays it against the engine so that the base around the plug is electrically grounded, kicks the starter lever, and watches the spark-plug gap for a blue spark. If there isn't any he can conclude one of two things: (a) There is an electrical failure or (b) his experiment is sloppy. If he

is experienced he will try it a few more times, checking connections, trying every way he can think of to get that plug to fire. Then, if he can't get it to fire, he finally concludes that *a* is correct, there's an electrical failure, and the experiment is over. He has proved that his hypothesis is correct.

In the final category, conclusions, skill comes in stating no more than the experiment has proved. It hasn't proved that when he fixes the electrical system the motorcycle will start. There may be other things wrong. But he does know that the motorcycle isn't going to run until the electrical system is working and he sets up the next formal question: "Solve Problem: What is wrong with the electrical system?" He then sets up hypotheses for these and tests them. By asking the right questions and choosing the right tests and drawing the right conclusions the mechanic works his way down the echelons of the motorcycle hierarchy until he has found the exact specific cause or causes of the engine failure, and then he changes them so that they no longer cause the failure.

An untrained observer will see only physical labor and often get the idea that physical labor is mainly what the mechanic does. Actually the physical labor is the smallest and easiest part of what the mechanic does. By far the greatest part of his work is careful observation and precise thinking. That is why mechanics sometimes seem so taciturn and withdrawn when performing tests. They don't like it when you talk to them because they are concentrating on mental images, hierarchies, and not really looking at you or the physical motorcycle at all. They are using the experiment as part of a program to expand their hierarchy of knowledge of the faulty motorcycle and compare it to the correct hierarchy in their mind. They are looking at underlying form.



In Pirsig's narrative, we see five steps that are typical in scientific inquiry:

1. *Identification of the problem*. The first step is the realization that a problem exists. The problem may involve a question about something, a discrepancy in findings, or a gap in knowledge. In Pirsig's example, the fact that the motorcycle did not start constituted the problem.
2. *Statement of the problem*. The next step is the clarification of the problem. The investigator states more precisely the nature and scope of the problem that has been identified.
3. *Formulation of hypotheses*. The investigator formulates hypotheses about possible solutions of the problem. In the example, the first hypothesis was that the motorcycle did not start because of trouble in the electrical system.
4. *Prediction of consequences*. The investigator next predicts the consequences of each hypothesis; that is, what should result if the data support the hypothesis.
5. *Testing of hypotheses*. The researcher gathers objective data to evaluate the adequacy of each hypothesis formulated. If the data support the hypothesis, it is accepted as a reasonable explanation. If the data do not support the hypothesis, it is rejected.

Gribbin (1999) summed up the scientific process with the following quote from Richard Feynman, one of the great physicists of the 20th century:

> In general we look for a new law by the following process. First we guess it. Then we compute the consequences of the guess to see what would be implied if this law that we guessed is right. Then we compare the result of the computation to nature, with experiment or experience, compare it directly with observation, to see if it works. If it disagrees with experiment it is wrong. In that simple statement is the key to science. It does not make any difference how beautiful your guess is. It does not make any difference how smart you are, who made the guess, or what his name is—if it disagrees with experiment it is wrong. (p. 4)

## OTHER ASPECTS OF SCIENCE

In addition to the method scientists follow as they seek reliable knowledge, there are certain other aspects of the scientific approach, which we examine briefly. These are (1) assumptions made by scientists, (2) attitudes expected of scientists, and (3) formulation of scientific theory.

### ASSUMPTIONS MADE BY SCIENTISTS

A fundamental assumption scientists make is that the events they investigate are lawful or ordered—no event is capricious. Science is based on the belief that all natural phenomena have antecedent factors. This assumption is sometimes referred to as **universal determinism.** Primitive people proposed supernatural causes for most of the events they observed. Modern science did not develop until people began to look beyond supernatural explanations and to depend on the observation of nature itself to provide answers.

This assumption underlies any statement that declares that under specified conditions certain events will occur. For example, the chemist can state that when a mixture of potassium chlorate and manganese dioxide is heated, the process will produce oxygen. Behavioral scientists likewise assume that the behavior of organisms is lawful and predictable. Related to this first assumption is the belief that the events in nature are, at least to a degree, orderly and regular and that people can discover this order and regularity of nature through the scientific method.

A second assumption is that reliable knowledge can ultimately derive only from direct and objective observation. Reliance on empirical observation differentiates science from nonscience. The scientist does not depend on authority or tradition as sources of knowledge but insists on studying empirical evidence. In the history of science we find many examples of scientists who rejected the prevailing notions of their day and proceeded with their observations and experimentation. Galileo's early experiments with falling bodies, which may mark the beginning of modern scientific inquiry, resulted in new knowledge that contradicted notions held by the authorities of his day. A corollary of this assumption is the belief that only phenomena that are subject to observation lie within the realm of scientific investigation.

**THINK ABOUT IT 1.1**

Match the term on the left with the definition on the right.

| | |
|---|---|
| 1. Universal determinism | a. Proceeding from general to specific knowledge through logical argument |
| 2. Inductive reasoning | b. Deriving general conclusions through direct observation |
| 3. Deductive reasoning | c. A statement describing relationships among variables that is tentatively assumed to be true |
| 4. Hypothesis | d. The assumption that all natural phenomenal have antecedent factors |

**Answers**

1. d; 2. b; 3. a; 4. c

## ATTITUDES EXPECTED OF SCIENTISTS

Scientists recognize certain characteristic attitudes that they acquire as they pursue their work:

1. *Scientists are essentially doubters, who maintain a highly skeptical attitude toward the data of science*. Scientists investigate questions concerning the relationships among natural phenomena. Their findings are regarded as tentative, however, and are not accepted by themselves or other scientists unless further investigations can verify them. Verification occurs when repeated observations yield the same or similar results. Verification thus requires scientists to make their research measurements and procedures known so that others may replicate the study and verify, or fail to verify, the findings.
2. *Scientists are objective and impartial*. In conducting observations and interpreting data, scientists seek knowledge and are not trying to prove a point. They take particular care to collect data in such a way that any personal biases they may have will not influence their observations. They look for observable evidence and accept the findings even when those results are contrary to their own opinions. If the accumulated evidence upsets a favorite theory, then they either discard that theory or modify it to agree with the findings.

   In reality, scientists are human like the rest of us. Some scientists have been known to report only findings that agreed with their preconceived ideas or have even made up data to support their contentions. A notorious example occurred when Stalin ruled the Soviet Union. His secretary of agriculture, Lysenko, asserted that environment changed heredity. Those scientists who reported results supporting this contention got published, got to keep their jobs, and got promoted. Those who reported research results contrary to Lysenko's belief often lost their jobs or were sent to Siberia.

   Scientists in other countries tried to replicate these studies, but none of them got results that supported Lysenko's contention. They concluded that the phenomenon did not exist. Soon after Stalin's death, Lysenko's contentions were repudiated, and Soviet scientists admitted that they had reported what was wanted, not what they had observed.

3. *Scientists deal with facts, not values*. Scientists do not indicate any potential moral implications of their findings; they do not make decisions for other people about what is good or what is bad. Scientists provide data concerning the relationships among events, but you must go beyond scientific data if you want a decision about whether a certain consequence is desirable. Thus, although the findings of science may be of key importance in solving a problem about a value decision, the data themselves do not furnish that value judgment.
4. *Scientists are not satisfied with isolated facts but seek to integrate and systematize their findings*. They want to put the things known into an orderly system. Thus, scientists aim for theories that seek to bring together empirical findings into a meaningful pattern. However, they regard these theories as tentative or provisional, subject to revision as new evidence appears.

## FORMULATION OF SCIENTIFIC THEORY

The last aspect of the scientific approach we consider here is the construction of theory. The ultimate goal of science is theory formation. Scientists, through empirical investigation, gather many facts, but facts by themselves are of limited usefulness. As facts accumulate, scientists must integrate, organize, and classify to make the isolated findings meaningful. They must identify and explain significant relationships in the data. That is where theory comes into play. Scientists formulate theories to summarize and order the existing knowledge in a particular area. A **theory** may be defined as a set of interrelated constructs and propositions that presents an explanation of phenomena and makes predictions about relationships among variables relevant to the phenomena.

Theories knit together the results of observations, enabling scientists to make general statements about variables and the relationships among variables. Theories range from a few simple generalizations to complex formulations of laws. For example, you can observe that if you hold pressure constant, hydrogen gas expands when its temperature increases from 208°C to 408°C. You can observe that if you hold pressure constant, oxygen gas contracts when its temperature decreases from 608°C to 508°C. A familiar theory, Charles's Law, summarizes the observed effects of temperature changes on the volumes of all gases: When pressure is held constant, as the temperature of a gas increases, its volume increases; and as the temperature of a gas decreases, its volume decreases. The theory not only summarizes previous information but also predicts other phenomena by telling you what to expect of any gas under any temperature change.

### Purposes of Theories

Theories serve useful functions in the development of science. They (1) organize empirical findings and explain phenomena, (2) predict phenomena, and (3) stimulate new research. A theory organizes the findings from many separate observations and investigations into a framework that provides explanations of phenomena. We would not have progress if science were composed only of multiple separate facts. A single theory can integrate many facts by showing what variables are related and how they are related. A theory of learning, for example, might explain the relationships among the speed and efficiency of learning and

such variables as motivation, reinforcement, practice, and so on. Researchers have developed useful theories to explain motivation, intellectual and cognitive development, moral development, social development, and so on. From the explanatory framework of a theory, scientists can proceed to make predictions about what will happen in novel situations. If these predictions are supported by scientific investigation, then science proceeds finally to control. As soon as a statement (theory) was made about the relationship between the *Anopheles* mosquito and malaria in humans, scientists could (1) *explain* why malaria was endemic in some areas and not in others, (2) *predict* how changes in the environment would entail changes in the incidence of malaria, and (3) *control* malaria by changing the environment.

Researchers state and test hypotheses deduced from theories, which results in the development of new knowledge. Deductions from a theory permit predictions of phenomena, some as yet unobserved. For example, astronomers predicted the existence of the outermost planets from theory long before they were actually observed. Testing the deductions from a theory serves to confirm and elaborate the theory. If, however, research findings do not support the theory, scientists revise it and then collect more data to test the revised theory.

### Criteria for Theories

To serve its purpose in science, a theory should satisfy certain criteria. The following are some of the characteristics of a sound theory:

1. *A theory should be able to explain the observed facts relating to a particular problem*. It should be able to propose the "how" and "why" concerning the phenomena under consideration. This explanation of the events should take the simplest form possible. Scientists favor a theory that has fewer complexities and assumptions over a more complicated one. This rule is called the **principle of parsimony**.
2. *A theory should be consistent with observed facts and with the already established body of knowledge*. Scientists build on what has already been found. They look for the theory that provides the most probable or the most efficient way of accounting for the accumulated facts.
3. *A theory should provide means for its verification*. Scientists achieve this for most theories by making deductions in the form of hypotheses stating the consequences that you can expect to observe if the theory is valid. Scientists can then investigate or test these hypotheses empirically to determine whether the data support the theory. We must emphasize that it is inappropriate to speak of the "truth" or "falsity" of a theory. The acceptance or rejection of a theory depends primarily on its *utility*, or usefulness. A theory is useful or not useful depending on how efficiently it leads to predictions concerning observable consequences, which are then confirmed when the empirical data are collected. Even then, any theory is tentative and subject to revision as new evidence accumulates.

   You may recall the old theory of formal discipline, which stated that the mind is like a muscle that can be strengthened through exercise. Subjects such as logic, Latin, and Greek were once included in the curriculum because educators believed them to be best for strengthening the mind. This theory

of formal discipline prevailed until the early 20th century, when E. L. Thorndike, William James, and Charles Judd challenged and abandoned it.

4. *A theory should stimulate new discoveries and indicate further areas in need of investigation.*

The goal of theory formation has been achieved to a far greater extent in the physical sciences than in the social sciences, which is not surprising because they are older sciences. In the early days of a science, the emphasis typically is on empiricism; scientists are concerned with collecting facts in particular areas. Only with maturity does a science begin to integrate the isolated knowledge into a theoretical framework.

Although there are marked differences in the number and power of the theories that have been established in the physical and social sciences, theory has the same role to play in the progress of any science. Regardless of the subject matter, theory works in essentially the same way. It serves to summarize existing knowledge, to explain observed events and relationships, and to predict the occurrence of unobserved events and relationships. Theories represent the best efforts to understand the basic structure of the world in which we live.

## THINK ABOUT IT 1.2

Throughout history, mankind has sought to explain the source of the sun's heat. The following are among the proposed explanations:

a. The sun is a god miraculously creating heat.
b. The heat comes from combustion like a log burning in a fireplace.
c. The sun is an enormous ball of gas. The pressure created by gravity on this great mass creates great heat.
d. The sun's heat comes from atomic fusion as in the hydrogen bomb.

**Questions**

1. Which of the explanations are subject to disproof through observation?
2. Which are scientific theories?
3. Most scientific textbooks in the 19th century gave answer *c* as the best explanation of the sun's heat. Later, it was shown that if *c* was true, the sun could only produce heat for a short period of time. Should the publishers of these textbooks apologize for publishing *c* because it has now been shown to be inadequate for explaining the phenomenon?
4. Current texts present answer *d* as the best explanation of the sun's heat. Have we finally reached the correct explanation?

**Answers**

1. b, c, d
2. b, c, d
3. No. Science is dynamic, never claiming that a theory is the ultimate truth. There is no shame in embracing a theory and then discarding it when a better explanation comes along.
4. We do not know. Currently, it fits the facts. It may be the ultimate answer, but scientists remain open to the possibility that future research may produce a better explanation.

## LIMITATIONS OF THE SCIENTIFIC APPROACH IN THE SOCIAL SCIENCES

Despite their use of the scientific approach and accumulation of a large quantity of reliable knowledge, education and the other social sciences have not attained the scientific status typical of the natural sciences. The social sciences have not established generalizations equivalent to the theories of the natural sciences in scope of explanatory power or in capacity to yield precise predictions. Frequently, researchers in the social sciences disagree on what the established facts are or what explanations are satisfactory for the assumed facts. Perhaps the social sciences will never realize the objectives of science as completely as the natural sciences. Certainly, we must stress that using the scientific approach is not in itself a sufficient condition for scientific achievement. Several limitations hinder the application of the scientific approach in education and the other social sciences.

### Complexity of Subject Matter

A major obstacle is the inherent complexity of subject matter in the social sciences. Natural scientists deal with physical and biological phenomena. A limited number of variables that can be measured precisely are involved in explaining many of these phenomena, and it is possible to establish universal laws. For example, Boyle's law, summarizing the influence of pressure on gas volume, a law that deals with relatively uncomplicated variables, formulates relations among phenomena that are apparently unvarying throughout the universe.

In contrast, social scientists deal with the human subject. They are concerned with the subject's behavior and development, both as an individual and as a member of a group. They must consider many variables, acting independently and in interaction, in any attempt to understand complex human behavior. Each individual is unique in the way he or she develops, in mental ability, in social and emotional behavior, and in total personality. The behavior of humans in groups and the influence of the behavior of group members on an individual must also be dealt with by social scientists. A group of first-graders in one situation will not behave like first-graders in another situation. There are learners, teachers, and environments, each with variations that contribute to the behavioral phenomena observed in a setting. Thus, researchers must be extremely cautious about making generalizations because the data from one group or in one situation may have limited validity for other groups and other settings.

### Difficulties in Observation

Observation, the sine qua non of science, is more difficult in the social sciences than in the natural sciences. Observation in the social sciences is often less objective because it more frequently involves interpretation on the part of the observers. For example, the subject matter for investigation is often a person's responses to the behavior of others. Motives, values, and attitudes are not open to inspection. Observers must make subjective interpretations when they decide that behaviors observed indicate the presence of any particular motive, value, or attitude. The problem is that the personal values and attitudes of social scientists may influence both what they choose to observe and their assessment of the findings on which they base their conclusions. Natural scientists study phenomena that require less subjective interpretation.

### Difficulties in Replication

The chemist can objectively observe the reaction between two chemicals in a test tube. The findings can be reported and the observations can be easily replicated by others. Replication is much more difficult to achieve in the social sciences. An American educator cannot reproduce the conditions of a Russian educator's experimental teaching method with the same precision as that with which an American chemist can replicate a Russian chemist's experiment. Even within a single school building, one cannot reproduce a given situation in its entirety and with precision. Social phenomena are singular events and cannot be totally repeated for purposes of observations.

### Interaction of Observer and Subjects

An additional problem is that mere observation of social phenomena may produce changes that might not have occurred otherwise. Researchers may think that $X$ is causing $Y$, when in fact their own observation of $X$ may cause $Y$. For example, in the well-known Hawthorne experiments, changes in worker productivity stemmed not from the varying working conditions but from the mere fact that the workers knew they had been singled out for investigation. Investigators are human beings, and their presence as observers in a situation may change the behavior of their human subjects. The use of hidden video cameras and audio cassettes may help minimize this interaction in some cases, but much social science research includes the responses of human subjects to human observers.

### Difficulties in Control

The range of possibilities for controlled experiments on human subjects is much more limited than in the natural sciences. The complexities involved in research on human subjects present control problems that have no parallels in the natural sciences. In the latter, rigid control of experimental conditions is possible in the laboratory. Such control is not possible with human subjects; social scientists must deal with many variables simultaneously and must work under conditions that are much less precise. They try to identify and control as many of these variables as possible, but the task is sometimes very difficult.

### Problems of Measurement

Systematic research must provide for measurement of the variables involved. The tools for measurement in the social sciences are much less perfect and precise than the tools of the natural sciences. Social science has nothing that can compare with the precision of the ruler, the thermometer, or numerous laboratory instruments. We have already pointed out that an understanding of human behavior is complicated by the large number of determining variables acting independently and in interaction. The multivariate statistical devices available for analyzing data in the social sciences take care of relatively few of the factors that obviously are interacting. Furthermore, these devices permit you to attribute the variance only to factors operating at the time of measurement. Factors that have influenced development in the past are not measurable in the present, even though they may have significantly influenced the course of development. Because the complexity and difficulty of observation, replication, and measurement complicate social science research, researchers must exercise great caution

in generalizing from their studies. It is often necessary to conduct several studies in an area before attempting to formulate generalizations. If they consistently confirm initial findings, then researchers can be more confident in making broad generalizations.

Despite the handicaps, education and the social sciences have made great progress, and their scientific status can be expected to increase as scientific investigation and methodology become more systematic and rigorous.

## THE NATURE OF RESEARCH

Scientific research is the application of the scientific approach to studying a problem. It is a way to acquire dependable and useful information. Its purpose is to discover answers to meaningful questions by applying scientific procedures. To be classified as scientific research, an investigation must involve the approach we described in the previous section. Although it may take place in different settings and may use different methods, scientific research is universally a systematic and objective search for reliable knowledge.

### EDUCATIONAL RESEARCH

Educational research is the application of the scientific approach to the study of educational problems. Educational research is the way in which people acquire dependable and useful information about the educative process. Educators usually conduct research to find a solution to some problem or to gain insight into an issue they do not understand. The ultimate goal is to discover general principles or interpretations of behavior that people can use to explain, predict, and control events in educational situations—in other words, to formulate scientific theory.

The acceptance of the scientific approach in education and the other social sciences has lagged far behind its acceptance in the physical sciences. In 1897, J. M. Rice, a pioneer in educational research, found himself in a situation similar to that described by the quotation attributed to Bacon previously in this chapter. Rice asked the educators at the annual meeting of the National Education Association's Department of Superintendence if it would be possible to determine whether students who are given 40 minutes of spelling each day learn more than students given 10 minutes each day. Rice (1912) reported,

> To my great surprise, the question threw consternation into the camp. The first to respond was a very popular professor of psychology engaged in training teachers in the West. He said, in effect, that the question was one which could never be answered; and he gave me a rather severe drubbing for taking up the time of such an important body of educators in asking them silly questions. (pp. 17–18)

Rice did, in fact, collect empirical evidence on his question and found that the differences in achievement between those spending 10 minutes a day and those spending 40 minutes a day were negligible. He also pointed out that many words children were required to learn how to spell had little practical value. His work led other investigators, such as Edward L. Thorndike, to use documentary analysis to determine the frequency of use of words in the English language. Their work in turn led to improvements in language arts texts and curricula.

## SUMMARY

Human beings have sought to acquire knowledge through experience, authority, deductive reasoning, inductive reasoning, and the scientific approach. The scientific approach is widely regarded as the single most reliable source of new knowledge.

The scientific approach rests on two basic assumptions: (1) People can derive truth from observation, and (2) phenomena conform to lawful relationships.

Scientific inquirers seek not absolute truth but, rather, theories that explain and predict phenomena in a reliable manner. They seek theories that are parsimonious, testable, and consistent, as well as theories that are themselves stimuli for further research. The scientific approach incorporates self-correction, inasmuch as every theory is tentative and may be set aside if a new theory better fits the evidence.

Investigators have used the scientific approach to explain, predict, and control physical phenomena for centuries. As a science, educational research uses investigative methods consistent with the basic procedures and operating conceptions of scientific inquiry. The complexity of educational variables and difficulties in making reliable observations impeded scientific inquiry in education. However, since the beginning of the movement early in the 20th century, scientific inquiry in education has enjoyed increasing acceptance and increasing success in both theoretical and practical research.

## KEY CONCEPTS

deductive reasoning
hypothesis
imperfect induction
inductive reasoning
perfect induction
principle of parsimony
scientific approach
theory
universal determinism

## EXERCISES

1. Identify the source of knowledge—*deductive reasoning, inductive reasoning*, or the *scientific approach*—most prominently used in the following examples:
   - **a.** After extensive observation of reactions, Lavoisier concluded that combustion is a process in which a burning substance combines with oxygen. His work was the death blow to the old phlogiston theory of burning.
   - **b.** Dalton, after much reflection, concluded that matter must consist of small particles called *atoms*. His early assumptions became the basis for the atomic theory.
   - **c.** Later scientists took Dalton's assumptions, made deductions from them, and proceeded to gather data that confirmed these assumptions. They found support for the atomic theory.
   - **d.** Knowing that radioactive substances constantly give off particles of energy without apparently reducing their mass, Einstein developed the formula $E = mc^2$ for converting matter into energy.
   - **e.** Accepting Einstein's theory, Fermi carried on experimentation that resulted in splitting the atom.
   - **f.** After studying reinforcement theory, a teacher hypothesizes that using a tutorial computer program will lead to superior achievement in arithmetic. She devises a study in which the tutorial is used with two sixth-grade classes, whereas conventional materials are used with two other sixth-grade classes.
2. What is the role of theory in scientific inquiry?
3. What is the difference between an inductive theory and a deductive theory?
4. Give examples of the use of authority and experience as sources of knowledge.
5. Evaluate the following deductive arguments:

**a.** All graduating seniors with high GPAs study Latin. John is a senior with a high GPA. Therefore, John studies Latin.
**b.** All vertebrates have backbones. This animal has a backbone. Therefore, this animal is a vertebrate.

**6.** Evaluate the following inductive arguments:
**a.** This animal has a backbone. Animals with backbones are vertebrates. I am reasonably certain that this animal is a vertebrate.
**b.** This is a student who studies very hard. Students who make good grades tend to study hard. This student probably makes good grades.

**7.** Which characteristic attitudes expected of scientists are violated in the following statements?
**a.** This study was undertaken to prove that the use of marijuana is detrimental to academic achievement.
**b.** It proved conclusively that this is the case.
**c.** The results show that marijuana is evil.

**8.** What are the characteristics of a useful theory?

**9.** Which of the following would contribute to theory development in education?
**a.** Evidence that supports the hypothesis of a study
**b.** Evidence that refutes the hypothesis of a study
**c.** (a) only
**d.** (a) and (b)

## ANSWERS

**1. a.** Inductive reasoning
**b.** Deductive reasoning
**c.** Scientific approach
**d.** Deductive reasoning
**e.** Scientific approach
**f.** Scientific approach

**2.** Theory integrates findings, summarizes information, provides leads for new research, and enables people to explain and predict phenomena.

**3.** An inductive theory serves to explain previous observations, whereas a deductive theory is developed before extensive observations have been made.

**4.** Answers will vary.

**5. a.** The argument is flawed; the major premise is not valid.
**b.** The argument is correct.

**6. a.** The argument is correct.
**b.** The argument is flawed; cannot state that because the student studies hard, he or she makes good grades.

**7. a.** The scientist is objective and impartial.
**b.** The scientist is skeptical and regards findings as tentative.
**c.** The scientist deals with facts, not values.

**8.** A useful theory explains the phenomena in the simplest form possible, is consistent with observation and the established body of knowledge, provides means for its verification, and stimulates new investigation.

**9.** d

## REFERENCES

Crawford, C. (1928). *The technique of study*. Boston: Houghton Mifflin.

Darwin, F. (Ed.). (2007). *The life and letters of Charles Darwin* (Vol. 1). New York: Echo Library.

Gribbin, J. (2000). *Almost everyone's guide to science*: *The universe, life and everything*. New Haven, CT: Yale University Press.

Mees, C. E. K. (1934). Scientific thought and social reconstruction. *General Electric Review*, *37*, 113–119.

Pirsig, R. M. (2006). *Zen and the art of motorcycle maintenance: An inquiry into values*. New York: HarperCollins.

Rice, J. M. (1912). *Scientific management in education*. New York: Hinds, Noble & Eldredge.

Terman, L. M. (1926). The mental and physical traits of a thousand gifted children. In *Genetic studies of genius* (Vol. 1). Stanford, CA: Stanford University Press.

Waddell, D. L., & Blankenship, J. C. (1994). Answer changing: A meta-analysis of the prevalence and patterns. *Journal of Continuing Education in Nursing*, *25*, 155–158.

CHAPTER 2

# Research Approaches in Education

Real science is not about certainty but about uncertainty.

## INSTRUCTIONAL OBJECTIVES

After studying this chapter, the student will be able to:

1. Identify the major research methodologies used in educational investigations.
2. Describe the main differences between quantitative and qualitative research.
3. List the specific types of research that fall into the broad categories of quantitative and qualitative research.
4. Give an example of a research problem that might be investigated by a mixed methods approach.
5. Identify the research methodology used in given examples of both quantitative and qualitative research.
6. List the steps involved in the research process.
7. Distinguish between the characteristics of basic and applied research.
8. Explain the terms *concept, construct*, and *variable*.
9. Distinguish among types of variables: categorical versus continuous and independent versus dependent.
10. Distinguish between constitutive and operational definitions and explain the importance of the latter in research.

Educational research is typically classified into two broad categories: quantitative and qualitative research. Each approach has its own methodology and terminology. **Quantitative research** uses objective measurement to gather numeric data that are used to answer questions or test predetermined hypotheses. It generally requires a well-controlled setting. **Qualitative research,** in contrast, focuses on understanding social phenomena from the perspective of the human participants in natural settings. It does not begin with formal hypotheses, but it may result in hypotheses as the study unfolds.

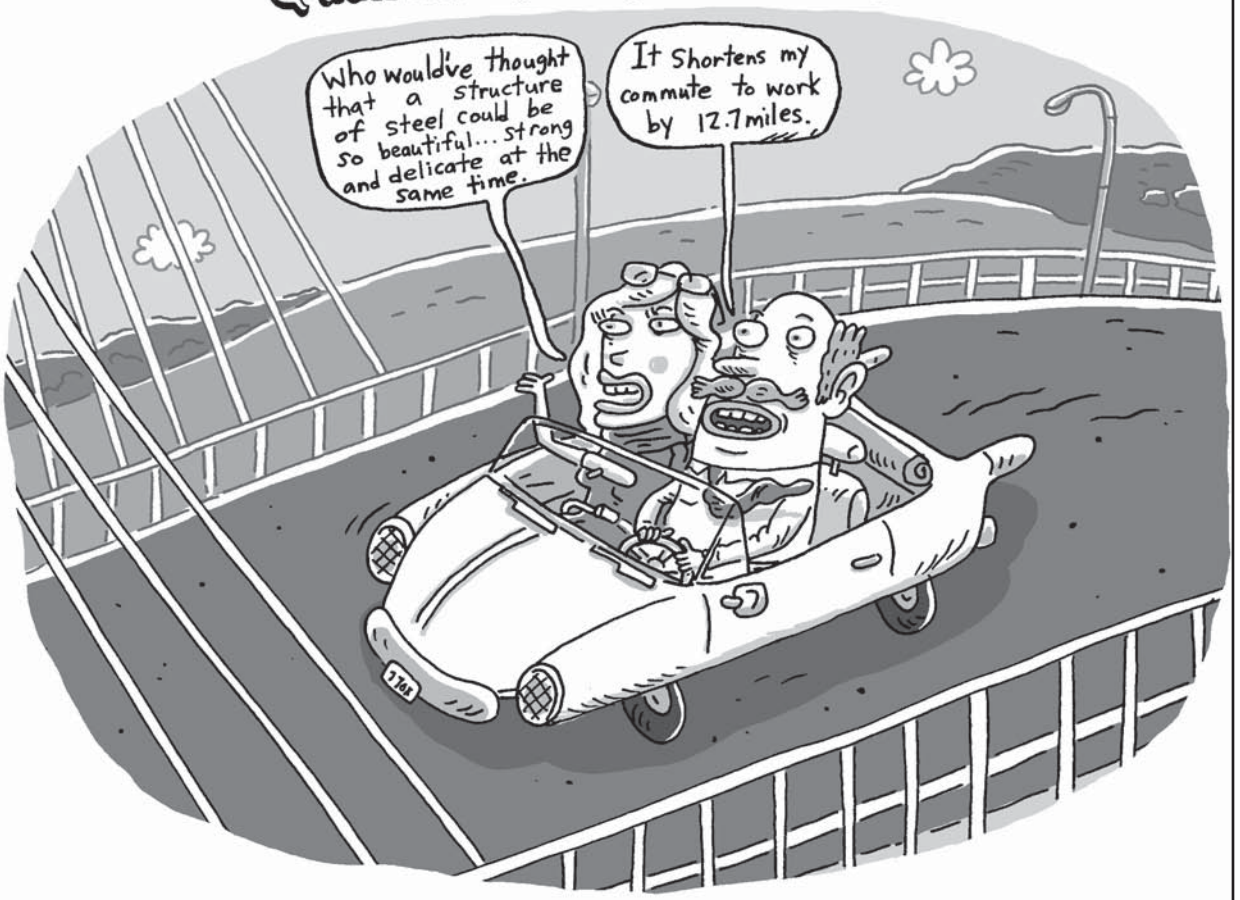
Quantitative and qualitative research stem from different philosophical assumptions that shape the ways researchers approach problems and collect and analyze data. Quantitative research originated in **positivism,** a philosophic view formulated in Europe in the 19th century. Positivists believe that general principles or laws govern the social world as they do the physical world and that through objective procedures researchers can discover these principles and apply them to understand human behavior. The positivists, such as Francis Bacon (recall the anecdote in Chapter 1), stressed observation as the primary source of dependable knowledge. Positivism is often considered the traditional scientific method, which involves hypothesis testing and objective data gathering to arrive at findings that are systematic, generalizable, and open to replication by other investigators.

Qualitative research is based on a different philosophical approach, which sees the individual and his or her world as so interconnected that essentially the one has no existence without the other. It sees social reality as unique; thus, researchers can only understand human behavior by focusing on the meanings that events have for the people involved. You must look not only at what people do but also at how they think and feel, and you must attempt to understand their reality. The intended result of a qualitative research study is a narrative report so rich and comprehensive that you can understand the social reality experienced by the participants. Furthermore, because researchers do not know in advance how naturally occurring events will unfold or what variables may be important, they do not begin a study with hypotheses.

Historically, the quantitative approach has dominated education research. In the late 20th century, however, scholars began to call for an alternative to the quantitative approach in educational research (Guba & Lincoln, 1988). They believed that using quantitative methods in highly controlled settings ignored the participants' perspectives and experiences. Qualitative research was the alternative. For a time, the relationship between quantitative and qualitative researchers was somewhat adversarial, but gradually there was a trend toward rapprochement as researchers began to see quantitative and qualitative methodology as complementary. Pring (2004) wrote, "The distinctions within the so-called paradigms are often as significant as the distinctions between them" (p. 48). A new methodology in which the same study uses both quantitative and qualitative approaches is called **mixed methods research.** The end result of mixed methods research is findings that may be more dependable and provide a more complete explanation of the research problem than either method alone could provide. All of these approaches are discussed in greater detail in later chapters.

As an example, assume you want to know the impact of a newly instituted dress code in an inner-city high school. Depending on exactly what you want to find out, you might choose either a quantitative or a qualitative approach alone or a mixed methods approach. In a quantitative study, you could measure subsequent attendance, frequency of violations of the code, and/or the number of school suspensions for failure to comply. You might administer a survey that assesses students' attitudes toward the new policy or one that asks teachers about the policy's effect on the incidence of disciplinary problems in the school. Numerical data would be gathered and analyzed. In a qualitative study, you might observe

**PICTURE THIS**




student behavior in classes and also interview groups of students to obtain in-depth information on how the new dress code has affected their motivation and learning in the classroom, their self-image, and their general attitude toward school. One might interview parents to assess their reaction to the dress code and how they think it has affected their children. The result could be a rich, verbal description of the personal and social meaning that the new policy has for students. Combining these approaches in a mixed methods study would provide the most complete information on the effect of the dress code.

Whether researchers choose quantitative, qualitative, or mixed methods depends not on their preference for one or another but on the suitability of the particular method for what they are studying and what they want to find out. One uses what will work to provide the type of data that are appropriate for answering the research question. As Johnson and Onwuegbuzie (2004) wrote, "What is most fundamental is the research question—research methods should follow

**Table 2.1** Comparison of Quantitative and Qualitative Research

| | Quantitative | Qualitative |
|---|---|---|
| **Purpose** | To study relationships, cause and effect | To examine a phenomenon as it is, in rich detail |
| **Design** | Developed prior to study | Flexible, evolves during study |
| **Approach** | Deductive; tests theory | Inductive; may generate theory |
| **Tools** | Uses preselected instruments | The researcher is primary data collection tool |
| **Sample** | Uses large samples | Uses small samples |
| **Analysis** | Statistical analysis of numeric data | Narrative description and interpretation |

research questions in a way that offers the best chance to obtain useful (and the most thorough) answers" (pp. 17–18).

A caveat is in order, however. It may be more difficult to carry out a mixed methods study because one must have knowledge and an understanding of both quantitative and qualitative methodology. A mixed methods study also typically involves more extensive data collection and analysis and thus will require more time and effort to complete. We discuss mixed methods research in Chapter 19. The reader is referred to Creswell and Plano Clark (2006) for in-depth discussion of the mixed methods approach in research. Table 2.1 summarizes briefly the major characteristics of the quantitative and qualitative research approaches.

### Think About It 2.1

A study was conducted to determine the effect on student performance of implementing a block schedule in high schools. Two high schools in a district were selected to implement block scheduling, and two other high schools continued with traditional scheduling. At the conclusion of the 2-year study, student performance on state tests (adjusted for prechange performance) was compared across the high schools. The results puzzled the researchers. One high school with block scheduling had much higher scores than the control schools, but the other high school with block scheduling had notably lower scores than the control schools. The researchers believed no valid conclusions about the benefits of block scheduling could be drawn and that further research was necessary. How would quantitative and qualitative researchers differ in their approach?

**Answer**

The quantitative researcher would suggest an experimental study to try to determine the impact of block scheduling and other demographic variables on student performance. The qualitative researcher would suggest in-depth case studies of the two block scheduling schools to determine what differences in the social context of the schools could account for the different results. Data could be collected through observation and focus groups.

We next discuss specific types of research included in the broad categories of quantitative and qualitative research.

# QUANTITATIVE RESEARCH

Quantitative research may be further classified as either experimental or nonexperimental.

## EXPERIMENTAL RESEARCH

Researchers study variables, which are characteristics that take on different values across people or things. **Experimental research** involves a study of the effect of the systematic manipulation of one variable(s) on another variable. The manipulated variable is called the **experimental treatment** or the **independent variable.** The observed and measured variable is called the **dependent variable.** For example, assume a university researcher wanted to investigate the effect of providing online feedback to students immediately following course examinations. Using two sections of economics taught by the same professor, the researcher using a random procedure would select one section to receive immediate online feedback about their performance on test questions; the other section would receive feedback during their next class session (independent variables). The researcher would compare the two sections' exam scores and their final grades in the course (dependent variables). If test scores and final grades were higher than could be accounted for by chance in the section receiving online feedback, the researcher could tentatively conclude that there is evidence the online feedback (treatment or independent variable) contributed to greater learning than the in-class feedback.

In experiments, you seek to control all other variables that might influence the dependent variable. In the foregoing example, the researcher would attempt to make sure that both groups had the same instructor; that both sections met at the same time of day but on different days; that lecture notes, readings, and exams were the same; and so forth. The researcher might also check the ability level and background of the students in the two sections to make sure one section was not superior or better prepared than the other.

To have a "true" experiment, researchers must use a random process such as a coin toss to assign available subjects to the experimental treatments. With random assignment, each subject has an equal and independent chance of being assigned to any group; thus, the assignment is independent of the researcher's personal judgment or the characteristics of the subjects themselves. Sometimes, however, researchers cannot randomly assign subjects to experimental treatments for a study. Instead, as in the preceding example, the experimenter must use already assembled groups such as classes. In this case, the research is called **quasi-experimental.** We discuss experimental research more thoroughly in Chapters 10 and 11.

## NONEXPERIMENTAL RESEARCH

In **nonexperimental quantitative research,** the researcher identifies variables and may look for relationships among them but does not manipulate the variables. Major forms of nonexperimental research are relationship studies including ex post facto and correlational research and survey research.

**Ex post facto research** is similar to an experiment, except the researcher does not manipulate the independent variable, which has already occurred in the natural course of events. The researcher simply compares groups differing on the preexisting independent variable to determine any relationship to the

dependent variable. Because there is no manipulation or control of the independent variable, one must be very careful regarding the conclusions that are drawn about any observed relationship. The variables in this type of research differ in kind rather than in amount. Later, we refer to these as **categorical variables.** For example, to answer the question, "What is the effect of part-time work on school achievement of high school students?" one would use the ex post facto method. The researcher would not manipulate the lives of high school students by having some take part-time jobs and others not; instead, the researcher would identify students who already work part-time and compare their achievement with that of students who do not work. Because researchers lack control over many factors, they must be especially careful in interpreting the results of ex post facto investigations. They do not establish a simple causal relationship among the variables of a study. We discuss ex post facto research in detail in Chapter 12.

**Correlational research** gathers data from individuals on two or more variables and then seeks to determine if the variables are related (correlated). *Correlation* means the extent to which the two variables vary directly (positive correlation) or inversely (negative correlation). The degree of relationship is expressed as a numeric index called the *coefficient of correlation.* Correlational research might ask about the relationship between the quality of writing samples produced by incoming college freshmen and their academic performance during the freshman year. Also, one might investigate the relationship between performance on a language aptitude test and success in a high school foreign language course.

Both ex post facto and correlational research investigate relationships between variables. The major distinction between the two is that in ex post facto research one categorizes the participants into at least two groups on one variable and then compares them on the other variable. In correlational research, a researcher deals with one group of individuals measured on at least two continuous variables. We discuss correlational research in Chapter 13.

### Think About It 2.2

A study investigated the difference in French grammar performance between high school freshmen taught by two different methods.

1. (a) What is the independent variable in this study? (b) What is the dependent variable? (c) What is not a variable?
2. What would the researcher have to do to make the study a true experiment?
3. How would the study be classified if intact classes were assigned to receive teaching method A or B?
4. How would the study be classified if it compared the French grammar performance of students who had already been taught using method A with the performance of those who had already been taught using method B?

**Answers:**

1. (a) teaching method, (b) French grammar performance, (c) grade level of participants
2. Randomly assign students to teaching method A or B
3. Quasi-experimental research
4. Ex post facto

**Survey research** (also called **descriptive research**) uses instruments such as questionnaires and interviews to gather information from groups of individuals. Surveys permit the researcher to summarize the characteristics of different groups or to measure their attitudes and opinions toward some issue. Researchers in education and the social sciences use surveys widely. For example, an educational

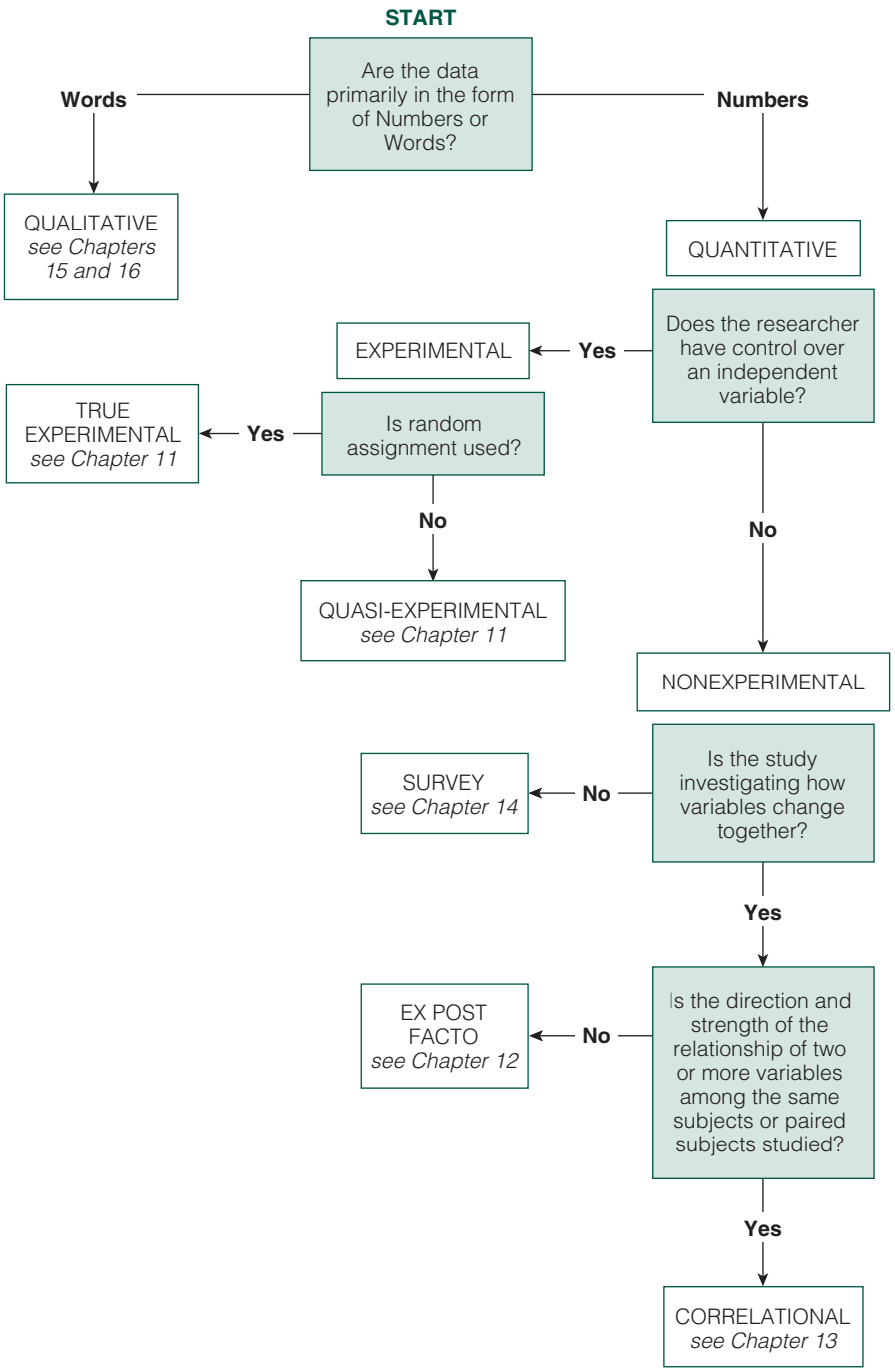



**Figure 2.1** Major Types of Quantitative Educational Research

researcher might ask a group of parents about what kind of sex education program, if any, they believe schools should provide for middle school students. A survey of teachers could reveal their perceptions of giftedness in schoolchildren. Opinion polls are surveys that researchers conduct to determine whom people are likely to vote for or what positions they take on certain issues. We discuss surveys in Chapter 14. Figure 2.1 summarizes the major types of quantitative research.

## QUALITATIVE RESEARCH

Qualitative researchers seek to understand a phenomenon by focusing on the total picture rather than breaking it down into variables. The goal is a holistic picture and depth of understanding rather than a numeric analysis of data. For example, social scientists have long observed that differences in educational background alone do not seem to account for the difficulties black students encounter in a predominantly white university. Researchers could explore the question "How do black students perceive their academic experience in a white university?" using qualitative methodology. Researchers would focus on a few black students and study them in great detail through observation and in-depth interviews. There are many different types of qualitative research; we consider briefly eight of the most widely used approaches: basic interpretative studies, case studies, document or content analysis, ethnography, grounded theory, historical studies, narrative inquiry, and phenomenological studies. We discuss qualitative research more extensively in Chapters 15 and 16.

### BASIC INTERPRETATIVE STUDIES

A **basic interpretative study** provides descriptive accounts targeted to understanding a phenomenon using data that might be collected in a variety of ways, such as interviews, observations, and document review. The purpose is to understand the world or experience of another. These are the most simple and most common qualitative studies. Describing how teachers perceive their role in selecting curriculum is an example.

### CASE STUDIES

A **case study** is a type of ethnographic research study that focuses on a single unit, such as one individual, one group, one organization, or one program. The goal is to arrive at a detailed description and understanding of the entity (the "case"). In addition, a case study can result in data from which generalizations to theory are possible. Freud, for example, used the case study extensively in building his theory of personality. Case studies use multiple methods, such as interviews, observation, and archives, to gather data. Education and psychology researchers have used the case study widely. For example, you might conduct a case study of an inner-city school in which the students have achieved at a high level on standardized tests.

### DOCUMENT OR CONTENT ANALYSIS

**Content analysis** focuses on analyzing and interpreting recorded material to learn about human behavior. The material may be public records, textbooks, letters, films, tapes, diaries, themes, reports, or other documents. Content analysis

usually begins with a question that the researcher believes can best be answered by studying documents. For example, what was the employment situation like for married women teachers in the early 20th century? Content analysis is sometimes quantitative, such as when one investigates middle school science textbooks to determine the extent of coverage given to the achievements of minority scientists.

## ETHNOGRAPHY

**Ethnography** is an in-depth study of naturally occurring behavior within a culture or social group. Social scientists sometimes call ethnography field research because it is conducted in a natural setting or "field." The researcher observes group behavior as it occurs naturally in the setting, without any simulation or imposed structure. Ethnography requires a variety of data-gathering procedures, such as prolonged observation of the setting, interviewing members of the culture, and studying documents and artifacts. Researchers interpret the data in the context of the situation in which they gathered the data.

Ethnography is rooted in anthropology. Educational researchers use ethnography, for example, to learn how the educational experience in suburban schools differs from that in inner-city schools.

## GROUNDED THEORY

**Grounded theory** research is designed to develop a theory of social phenomena based on the field data collected in a study. Experience with the data generates insights, hypotheses, and questions, which researchers pursue with further data collection.

From an inductive analysis of the data, the researcher constructs concepts. He or she then forms a theory by proposing plausible relationships among the concepts. The theory is thus said to be grounded in the data. For example, a researcher interested in mainstreaming in elementary school could observe a number of classrooms and conduct interviews with teachers and students. Analysis of the data could lead to a theory about mainstreaming in the elementary school.

## HISTORICAL RESEARCH

**Historical research** analyzes documents, and artifacts and/or uses interviews with eyewitnesses to gain insight into past events. The success of historical research depends on the accuracy and completeness of the source material. The researcher must establish the authenticity of the documents used, as well as the validity of their contents.

An educational researcher might want to investigate the trends in kindergarten education in a particular school district from its beginnings to the present. Also, one might investigate the methods used to teach reading in the past or study school practices and policies such as grade retention.

Some of the data collected in historical research might be quantitative, such as when an investigator compares a school district's reading achievement scores when one teaching method was used with the scores obtained when another method was in vogue.

**Table 2.2** Types of Qualitative Research

| Type | Major Question |
|---|---|
| Basic interpretive studies | How are events, processes, and activities perceived by the participant? |
| Case study | What are the characteristics of this individual, organization, or group? |
| Document analysis | What can be learned about this phenomenon by studying certain documents? |
| Ethnography | What are the culture and perspectives of this group of people in its natural setting? |
| Grounded theory | What theory can be derived inductively about a phenomenon from the data collected in a particular setting? |
| Historical studies | What insights or conclusions can be reached about this past event? |
| Narrative inquiry | What insights and understandings about an issue emerge from examining life stories? |
| Phenomenological study | What does this experience mean for the participants in the experience? |

### NARRATIVE INQUIRY

In **narrative inquiry,** researchers examine the stories people tell about their lives and co-construct a narrative analysis of those stories. The researcher and those telling their stories have an equal voice in determining the meanings attributed to the experiences. Narrative analysis has also been referred to using terms such as life stories. A researcher investigating teacher reflection or teacher pathways into teaching might use narrative inquiry approaches.

### PHENOMENONOLOGICAL STUDIES

**Phenomenological studies** begin with the assumption that multiple realities are rooted in subjects' perspectives. Thus, an experience has different meanings for each person. Through unstructured interviews, the investigator explores the subject's thoughts and feelings to elicit the essence of an individual's experience. A phenomenological study might be conducted to answer the questions, "What is the relationship like between a beginning teacher and his or her mentor?" and "What does the experience mean to the beginning teacher?"

Table 2.2 summarizes the major types of qualitative research.

## TYPICAL STAGES IN RESEARCH

All researchers engage in a number of activities regardless of the particular methodology chosen for the research. We focus on these activities in greater detail in later chapters; here, we briefly summarize the steps involved.

1. *Selecting a problem.* The first step is to select the problem to investigate. The problem should be consequential enough to warrant investigation. Also, the answer to the problem is not already available, but the means for finding answers are available. Quantitative researchers typically state the problem in the form of a specific question about the relationship between variables. For example, "Do children who are taught reading through the whole-language approach score higher on reading achievement than children who

are taught reading through phonics?" or "What do teachers know about attention deficit/hyperactivity disorder?" Qualitative researchers begin with a general topic of interest. The problem statement may ask the "why" or "how" of certain phenomena. For example, "How do elementary classroom teachers deal with aggressive children?" or "How do high school students become alienated from their peers?"

2. *Reviewing the literature on the problem.* Researchers should thoroughly review the relevant literature to gain more understanding and insight into the problem and to determine what research may already have been done. The beginning researcher will likely turn to the literature for help in locating and formulating a researchable problem.
3. *Designing the research.* The investigator next plans how to conduct research to answer the question. The design is the researcher's plan for the study, which includes the method to be used, what data will be gathered, where, how, and from whom. Quantitative researchers maintain that once this research plan is set forth, it must be followed. Unhypothesized observed relationships among variables may be reported and proposed as topics for future research, but they should not replace the original intent of the study. In qualitative research, the design is flexible and may change during the investigation if appropriate. The design of qualitative research is thus often described as "emergent."
4. *Collecting the data.* The next step involves executing the research plan. Quantitative researchers use a wide variety of instruments to gather data, including tests, questionnaires, ratings, and attitude scales. Qualitative researchers also have a toolbox of data-gathering techniques, including in-depth interviewing, participant observation, and document analysis.
5. *Analyzing the data.* The data collected in research must be analyzed. Quantitative data are usually in the form of numbers that researchers analyze using various statistical procedures. Even verbal data, such as compositions written by high school students, would be converted through the scoring process to a numerical form. The analysis of the numerical data in quantitative research provides evidence that supports or fails to support the hypothesis of the study. Qualitative data generally take the form of words (descriptions, observations, impressions, recordings, and the like). The researcher must organize and categorize or code the large mass of data so that they can be described and interpreted. Although the qualitative researcher does not deal with statistics, analyzing qualitative data is not easy. It is a time-consuming and painstaking process.
6. *Interpreting the findings and stating conclusions.* The researcher next tries to interpret the findings in terms of the research problem. The quantitative researcher typically makes statements about the probability that such a finding is due to chance and reaches a conclusion about the hypothesis. Qualitative researchers present their interpretations and explanations in narrative form. They do not talk about probability but try to emphasize the *trustworthiness* and *credibility* of the findings (see Chapter 16).
7. *Reporting results.* Researchers must make their procedures, findings, and conclusions available in a form intelligible to others who may be interested.

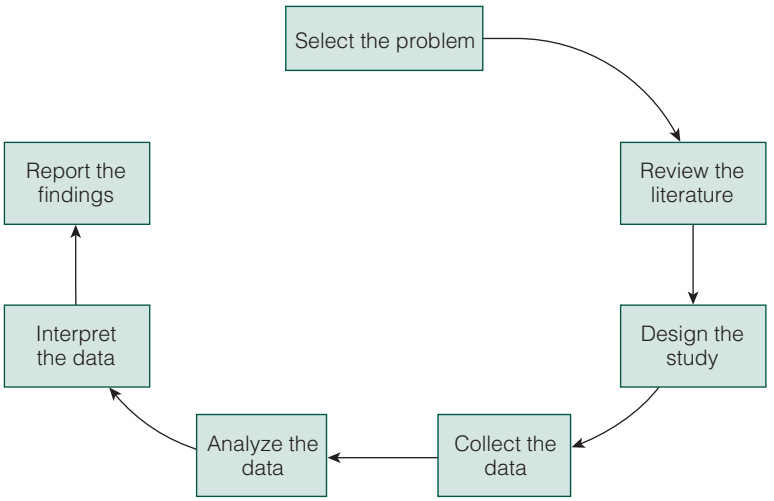



**Figure 2.2** Stages in the Research Process

This involves clearly and concisely presenting the steps in the study in sufficient detail so that another person can replicate it.

We discuss in detail each of the foregoing stages of a research study in later chapters. It is probably rare for researchers to follow precisely the sequence as we have described in the preceding discussion. These activities often overlap, and researchers may move back and forth from one stage to another. These steps are shown in Figure 2.2.

## QUESTIONS THAT EDUCATIONAL RESEARCHERS ASK

The specific question chosen for research, of course, depends on the area that interests the researchers, their background, and the particular problem they confront. However, we may classify questions in educational research as theoretical (having to do with fundamental principles) or as practical (designed to solve immediate problems of the everyday situation).

### THEORETICAL QUESTIONS

Questions of a theoretical nature are those asking "What is it?" or "How does it occur?" or "Why does it occur?" Educational researchers formulate "what" questions more specifically as "What is intelligence?" or "What is creativity?" Typical "how" questions are "How does the child learn?" or "How does personality develop?" "Why" questions might ask "Why does one forget?" or "Why are some children more achievement-oriented than other children?"

Research with a theoretical orientation may focus on either developing new theories or testing existing theories. The former involves a type of study in which researchers seek to discover generalizations about behavior, with the goal of clarifying the nature of relationships among variables. They may believe that certain variables are related and thus conduct research to describe the nature of

the relationship. From the findings, they may begin to formulate a theory about the phenomenon. Theories of learning have thus been developed because investigators have shown the relationships among certain methods, individual and environmental variables, and the efficiency of the learning process.

Probably more common in quantitative educational research are studies that aim to test already existing theories. It may be overly ambitious, especially for beginning researchers in education, to take as a goal the development of a theory. It is usually more realistic to seek to deduce hypotheses from existing theories of learning, personality, motivation, and so forth, and to test these hypotheses. If the hypotheses are logical deductions from the theory, and the empirical tests provide evidence that supports the hypotheses, then this evidence also provides support for the theory.

### PRACTICAL QUESTIONS

Many questions in educational research are direct and practical, aimed at solving specific problems that educators may encounter in everyday activities. These questions are relevant for educational research because they deal with actual problems at the level of practice and lead to an improvement in the teaching–learning process. Slavin (2004) writes that "enlightened educators look to education research for well-founded evidence to help them do a better job with the children they serve" (p. 27). Some academic researchers, however, criticize practitioner research as not being sufficiently rigorous. But Anderson (2002) also argues for a research continuum for doctoral students in education that includes practitioner research. Such practical questions are, for example, "How effective is peer tutoring in the elementary school classroom?" "How does teaching children cognitive strategies affect their reading comprehension?" "What is the relative effectiveness of the problem discussion method as compared with the lecture method in teaching high school social studies?" or "What are the most effective means of providing remediation to children who are falling behind?" The answers to such questions may be quite valuable in helping teachers make practical decisions in the classroom.

These practical questions can be investigated just as scientifically as the theoretical problems. The two types of questions differ primarily on the basis of the goals they hope to achieve rather than on the study's level of sophistication.

## BASIC AND APPLIED RESEARCH

Another system of classification concerns whether research is basic or applied. **Basic research** is research aimed at obtaining empirical data used to formulate and expand theory. Basic research is not oriented in design or purpose toward the solution of practical problems. Its essential aim is to expand the frontiers of knowledge without regard to practical application. Much early psychological investigation of reinforcement was basic research. Later, social scientists found that the reinforcement theory resulting from that research had educational applications.

**Applied research** aims to solve an immediate practical problem. It is research performed in relation to actual problems and under the conditions in which they appear in practice. Through applied research, educators can often solve

their problems at the appropriate level of complexity—that is, in the classroom teaching–learning situation. Although applied research may solve some specific question, it may not provide the general knowledge to solve other problems. For example, an elementary school teacher may study the effect of a new method of teaching fractions. She or he conducts the research to answer a practical question, not necessarily to make broad generalizations or to help develop a theory.

This classification of research is not always distinct, however, because there are varying degrees on the basic–applied continuum. Research along this basic–applied dimension is usually classified on the degree to which the findings are directly applicable to solving a practical problem. Basic research often has practical benefits in the long term. For example, advances in the practice of medicine depend on basic research in biochemistry, microbiology, and genetics. Likewise, progress in educational practice is related to progress in discovering general laws through basic psychological, educational, and sociological research.

Actually, in recent years basic and applied research has tended to merge in education and psychology. The trend has been for basic research to move more toward classroom studies because the findings of applied research may help basic researchers complete theoretical formulations. Researchers developing general theories of learning, for example, often go into classrooms because to understand how children learn investigators must consider variables such as context and social structure that are not present in the artificial environments of the laboratory. Once the theories are formulated, they can be tested through further research in the classroom.

## LANGUAGE OF RESEARCH

Any scientific discipline needs a specific language for describing and summarizing observations in that area. Scientists need terms at the empirical level to describe particular observations; they also need terms at the theoretical level for referring to hypothetical processes that may not be subject to direct observation. Scientists may use words taken from everyday language, but they often ascribe to them new and specific meanings not commonly found in ordinary usage. Or perhaps they introduce new terms that are not a part of everyday language but are created to meet special needs. One of these terms is *construct*.

### CONSTRUCTS

To summarize their observations and to provide explanations of behavior, scientists create constructs. **Constructs** are abstractions that cannot be observed directly but are useful in interpreting empirical data and in theory building. For example, people can observe that individuals differ in what they can learn and how quickly they can learn it. To account for this observation, scientists invented the construct called *intelligence*. They hypothesized that intelligence influences learning and that individuals differ in the extent to which they possess this trait. Other examples of constructs in educational research are motivation, reading readiness, anxiety, underachievement, creativity, and self-concept.

Defining constructs is a major concern for researchers. The further removed constructs are from the empirical facts or phenomena they are intended to

represent, the greater the possibility for misunderstanding and the greater the need for precise definitions. Constructs may be defined in a way that gives their general meaning, or they may be defined in terms of the operations by which they will be measured or manipulated in a particular study. The former type of definition is called a *constitutive definition;* the latter is known as an *operational definition.*

### Constitutive Definition

A **constitutive definition** is a formal definition in which a term is defined by using other terms. It is the dictionary type of definition. For example, intelligence may be defined as the ability to think abstractly or the capacity to acquire knowledge. This type of definition helps convey the general meaning of a construct, but it is not precise enough for research purposes. The researcher needs to define constructs so that readers know exactly what is meant by the term and so that other investigators can replicate the research. An operational definition serves this purpose.

### Operational Definition

An **operational definition** ascribes meaning to a construct by specifying operations that researchers must perform to measure or manipulate the construct. Operational definitions may not be as rich as constitutive definitions but are essential in research because investigators must collect data in terms of observable events. Scientists may deal on a theoretical level with such constructs as learning, motivation, anxiety, or achievement, but before studying them empirically, scientists must specify observable events to represent those constructs and the operations that will supply relevant data. Operational definitions help the researcher bridge the gap between the theoretical and the observable.

Although investigators are guided by their own experience and knowledge and the reports of other investigators, the operational definition of a concept is to some extent arbitrary. Often, investigators choose from a variety of possible operational definitions those that best represent their own approach to the problem.

Certainly an operational definition does not exhaust the full scientific meaning of any concept. It is very specific in meaning; its purpose is to delimit a term, to ensure that everyone concerned understands the particular way a term is being used. For example, a researcher might state, "For this study, intelligence is defined as the subjects' scores on the Wechsler Intelligence Scale for Children." Operational definitions are considered adequate if their procedures gather data that constitute acceptable indicators of the constructs they are intended to represent. Often, it is a matter of opinion whether they have achieved this result.

Operational definitions are essential to research because they permit investigators to measure abstract constructs and permit scientists to move from the level of constructs and theory to the level of observation, on which science is based. By using operational definitions, researchers can proceed with investigations that might not otherwise be possible. It is important to remember that although researchers report their findings in terms of abstract constructs and relate these to other research and to theory, what they have actually found is a relationship between two sets of observable and measurable data that they

selected to represent the constructs. In practice, an investigation of the relation between the construct creativity and the construct intelligence relates scores on an intelligence test to scores on a measure of creativity.

## VARIABLES

Researchers, especially quantitative researchers, find it useful to think in terms of variables. A **variable** is a construct or a characteristic that can take on different values or scores. Researchers study variables and the relationships that exist among variables. Height is one example of a variable; it can vary in an individual from one time to another, among individuals at the same time, among the averages for groups, and so on. Social class, gender, vocabulary level, intelligence, and spelling test scores are other examples of variables. In a study concerned with the relation of vocabulary level to science achievement among eighth-graders, the variables of interest are the measures of vocabulary and the measures of science achievement. There are different ways to measure science achievement. The researcher could use a standardized achievement test, a teacher-made test, grades in science class, or evaluations of completed science projects. Any of these measures could represent the variable "science achievement."

### Types of Variables

There are several ways to classify variables. Variables can be categorical, or they can be continuous. When researchers classify subjects by sorting them into mutually exclusive groups, the attribute on which they base the classification is termed a categorical variable. Home language, county of residence, father's principal occupation, and school in which enrolled are examples of categorical variables. The simplest type of categorical variable has only two mutually exclusive classes and is called a **dichotomous variable.** Male–female, citizen–alien, and pass–fail are dichotomous variables. Some categorical variables have more than two classes; examples are educational level, religious affiliation, and state of birth.

When an attribute has an infinite number of values within a range, it is a **continuous variable.** As a child grows from 40 to 41 inches, he or she passes through an infinite number of heights. Height, weight, age, and achievement test scores are examples of continuous variables.

The most important classification of variables is on the basis of their *use* within the research under consideration, when they are classified as independent variables or dependent variables. Independent variables are antecedent to dependent variables and are known or are hypothesized to influence the dependent variable, which is the outcome. In experimental studies, the treatment is the independent variable and the outcome is the dependent variable. In an experiment in which freshmen are randomly assigned to a "hands-on" unit on weather forecasting or to a textbook-centered unit and are then given a common exam at the end of the study, the method of instruction (hands-on versus textbook) antecedes the exam scores and is the independent variable in this study. The exam scores follow and are the dependent variable. The experimenter is hypothesizing that the exam scores will partially depend on how the students were taught weather forecasting. In this case, freshman status is a constant.

In nonexperimental studies, it is often more difficult to label variables as independent or dependent. We call the variable that inevitably precedes another one in time the independent variable. In a study of the relationship between teacher experience and students' achievement scores, teacher experience would be considered the independent variable. Not only does it precede student achievement but also it would be illogical to say that student achievement influenced teacher experience.

In some other cases, it is not possible to tell which variable is antecedent and which follows. In a study of the relationship between self-confidence and popularity in high school students, either variable could be considered independent or dependent. It is possible for a variable to be an independent variable in one study and a dependent variable in another. Whether a variable is independent or dependent depends on the purpose of the study. If you investigate the effect of motivation on achievement, then motivation is the independent variable.

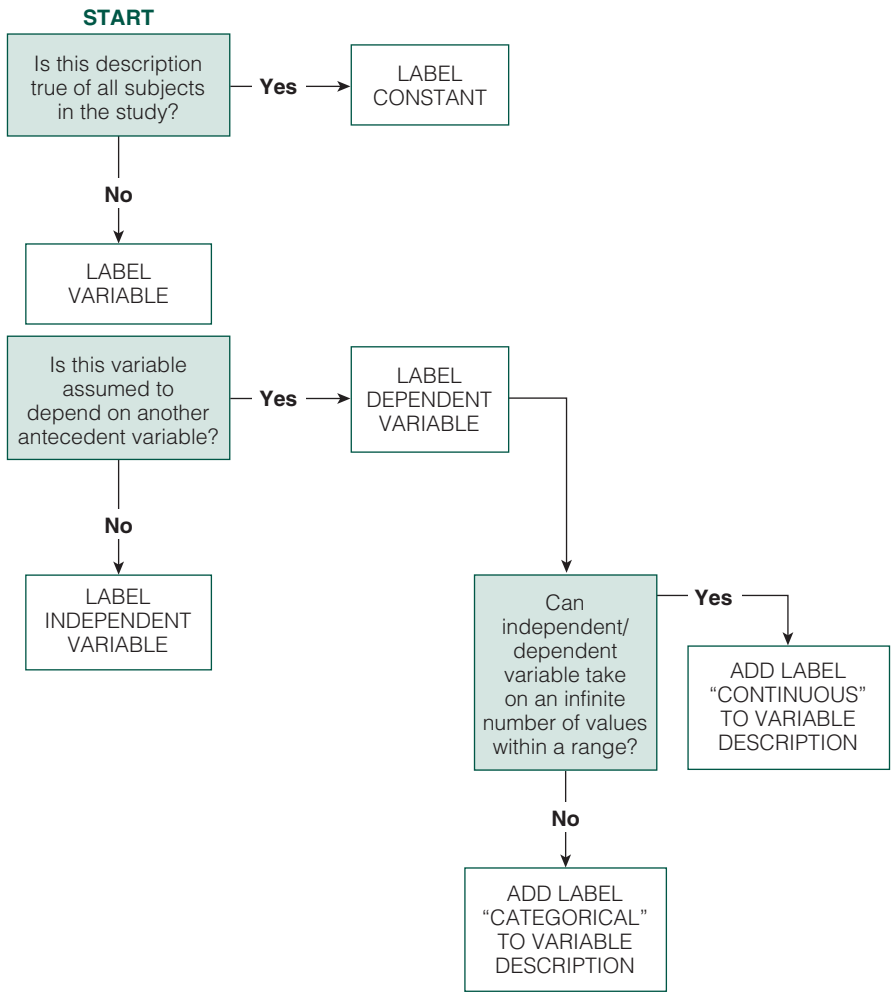



**Figure 2.3** Flow Chart for Classifying Variables and Constants

However, if you wish to determine the effect of testing procedures, classroom grouping arrangements, or grading procedures on students' motivation, then motivation becomes the dependent variable. Intelligence is generally treated as an independent variable because educators are interested in its effect on learning, the dependent variable. However, in studies investigating the effect of preschool experience on the intellectual development of children, intelligence is the dependent variable.

### CONSTANTS

The opposite of variable is **constant.** A constant is a fixed value within a study. If all subjects in a study are eighth-graders, then grade level is a constant. In a study comparing the attitudes toward school of high school girls who plan professional careers with those who do not plan professional careers, high school girls constitute a constant; whether they plan professional careers is the independent variable, and their attitudes constitute the dependent variable. Figure 2.3 illustrates a process for classifying variables and constants.

## SUMMARY

The two broad research methodologies in education are quantitative and qualitative. Quantitative research deals with questions of relationship, cause and effect, or current status that researchers can answer by gathering and statistically analyzing numeric data. It can be further classified as experimental and nonexperimental. Qualitative research focuses on understanding social phenomena and providing rich verbal descriptions of settings, situations, and participants. The qualitative approach includes a number of different methods, such as ethnography, case study, naturalistic observation, focused interviews, and historical studies. Researchers may also choose a mixed methods design, which combines quantitative and qualitative approaches in a single study. Both qualitative and quantitative methodologies are important in educational research. One chooses the method that will provide the data to answer the research question.

The typical steps in educational research are (1) selecting a problem; (2) reviewing the literature; (3) selecting a research strategy and developing instruments; (4) collecting, analyzing, and interpreting data; and (5) communicating the findings by reporting the results of the study.

Based on the objective, educational research can be classified into two major categories: basic and applied. The primary concern of basic research is to expand the frontiers of knowledge and to discover general laws. The main goal of applied research is to solve immediate practical problems.

At a theoretical level, educational scientists use terms such as *intelligence, creativity, problem-solving ability,* and *motivation,* which are abstractions from observation of certain behaviors. These are referred to as *constructs.* In quantitative research, constructs are quantified and take on different values. Thus, they are referred to as *variables.* There are two major types of variables: independent and dependent. If a variable is antecedent to another variable, it is called an *independent variable,* but if it is the consequence of another variable, it is the *dependent variable.*

In quantitative research, operational definitions are used to specify how variables will be measured in the study.

## KEY CONCEPTS

applied research
basic interpretative studies
basic research
case study
categorical variable
constant
constitutive definition
constructs
content analysis
continuous variable
correlation research
dependent variable
descriptive research
dichotomous variable
document analysis
ethnography
ex post facto research
experimental research
experimental treatment
grounded theory
historical research
independent variable
mixed methods research
narrative inquiry
operational definition
phenomenological studies
positivism
qualitative research
quantitative research
quasi-experimental research
survey research
variable

## EXERCISES

1. Based on the title of each study, classify the following research as *basic* or *applied:*
   a. The Effect of RNA (Ribonucleic Acid) Injections on the Transfer of Skills from Trained Animals to Untrained Animals
   b. Outcomes of a Remedial Arithmetic Program
   c. Conditioning as a Function of the Interval between the Conditioned and Original Stimulus
   d. Teaching Geometry to Cultivate Reflective Thinking: An Experimental Study
2. In a study designed to determine the effect of varying amounts of sleep deprivation on the learning of nonsense syllables, identify the following:
   a. What is the independent variable?
   b. What is the dependent variable?
3. Classify the following variables as *categorical* or *continuous:*
   a. Achievement
   b. Phonics method of reading versus look–say method of reading
   c. Spanish speaking, English speaking, French speaking
   d. Muscle prowess
   e. Music aptitude
4. The statement "Math achievement will be measured by the score made on the mathematics subtest of the Iowa Tests of Educational Development" is an example of which of the following?
   a. Hypothesis
   b. Theoretical principle
   c. Constitutively defined construct
   d. Operationally defined construct
5. Suggest a research study for which a mixed methods design would be the most appropriate.
6. Which research method (*experimental, ex post facto,* or *survey*) would most effectively give you answers to each of the following questions?
   a. Do children who eat breakfast get better grades in school?
   b. Does a unit on proper nutrition change children's breakfast-eating habits?
   c. How many children in school report that they do not have breakfast at home?
   d. Does the institution of a free breakfast program at school make a difference in the achievement of students?
7. Based on the titles, classify each of the following studies according to the research methodology most likely used:
   a. Gender-Based Differential Item Performance in Mathematics
   b. Improving Math Skills of High School Students
   c. College Students' Views and Ratings of an Ideal Professor
   d. Effect of Early Absence of Father on Scholastic Aptitude
   e. An Alternative High School: An In-Depth Study
   f. An Analysis of Social Studies Textbooks Used in Russian Middle Schools to Determine How America Is Portrayed
8. Give an example of how basic research in the biological sciences has improved the practice of medicine.

9. Give an example of how basic research in learning has improved the practice of teaching.
10. Give an example of applied research completed in your field of interest. List other areas in which additional research needs to be done in your field. What variables might be investigated in such studies?
11. What research methodology do you believe would be the most appropriate for investigating each of the following research questions?
    - **a.** How do parents conduct home schooling?
    - **b.** Does collaborative learning promote achievement in the college classroom?
    - **c.** What is the relationship between vocabulary and reading achievement in the primary grades?
    - **d.** Do parents support character education in the middle school curriculum?
    - **e.** What is the relationship between teachers' undergraduate background (education versus liberal arts) and certain measures of competence in the classroom?
    - **f.** How did educational reforms of the junior high school lead to development of the middle school?
    - **g.** What is the influence of family composition on children's conformity to the student role?
    - **h.** What kind of education do students receive in a church-sponsored Christian high school?
    - **i.** What is the extent of drug use among students in a suburban high school?
12. How would you operationally define science achievement?
13. The following is an abstract from the journal *Exceptional Children* (O'Connor, White, & Swanson, 2007):

    *Abstract: This research evaluated two methods to improve the reading fluency of struggling readers. Poor readers in grades 2 and 4 with (n=17) and without (n=20) learning disabilities were randomly assigned to one of two fluency practice variations or to a control group. Students in the treatments practiced reading aloud under repeated or continuous reading conditions with an adult listener in 15-min sessions, 3 days per week for 14 weeks. For students in the treatment conditions, growth curve analyses revealed significant differences in fluency and reading comprehension over students in the control. We found no significant differences between practice conditions.*

    Consider the following elements in this study: (1) poor readers, (2) students in grades 2 and 4, (3) students with and without learning disabilities, (4) reading aloud with an adult listener or not, (5) repeated or continuous reading conditions, (6) reading fluency growth, and (7) reading comprehension.
    - **a.** Which are categorical independent variables?
    - **b.** Which are continuous independent variables?
    - **c.** Which are categorical dependent variables?
    - **d.** Which are continuous dependent variables?
    - **e.** What independent variable did the researchers actively manipulate?
    - **f.** What was the population of interest?
14. Consider the following characteristics of a research study. Indicate whether each one is most likely (1) quantitative research or (2) qualitative research.
    - **a.** Researcher is objective and detached from participants
    - **b.** Develops hypotheses after data have been collected
    - **c.** Uses induction to analyze data
    - **d.** Uses large representative samples of individuals
    - **e.** Uses narrative description

## ANSWERS

1. **a.** Basic
   **b.** Applied
   **c.** Basic
   **d.** Applied
2. **a.** Amount of sleep deprivation
   **b.** Number of nonsense syllables learned
3. **a.** Continuous
   **b.** Categorical
   **c.** Categorical
   **d.** Continuous
   **e.** Continuous
4. d

**5.** Answers will vary.
**6. a.** Ex post facto
**b.** Experimental
**c.** Survey
**d.** Experimental
**7. a.** Ex post facto
**b.** Experimental
**c.** Survey
**d.** Ex post facto
**e.** Qualitative
**f.** Qualitative
**8.** Answers will vary.
**9.** Answers will vary.
**10.** Answers will vary.
**11. a.** Qualitative
**b.** Experimental
**c.** Correlational
**d.** Survey
**e.** Ex post facto
**f.** Historical
**g.** Ex post facto
**h.** Qualitative
**i.** Quantitative (survey)
**12.** Answers may vary; an example: "Science achievement is the score on the science subtest of the Iowa Test of Basic Skills."
**13. a.** 2, 3, 4, and 5
**b.** None
**c.** None
d. 6 and 7
**e.** 5
**f.** 1
**14. a.** 1
**b.** 2
**c.** 2
**d.** 1
**e.** 2

## REFERENCES


Anderson, G. L. (2002). Reflecting on research for doctoral students in education. *Educational Researcher, 31,* 22–25.

Creswell, J. W. (2003). *Research design: Qualitative, quantitative, and mixed methods research.* Thousand Oaks, CA: Sage.

Creswell, J. W., & Plano Clark, V. L. (2006). *Designing and conducting mixed methods research.* Thousand Oaks, CA: Sage.

Guba, E. G., & Lincoln, Y. S. (1988). Do inquiry paradigms imply inquiry methodologies? In D. M. Fetterman (Ed.), *Qualitative approaches to evaluation in education* (pp. 89–115). New York: Praeger.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A paradigm whose time has come. *Educational Researcher, 33*(7), 14–26.

King, L. A. (2004). Measures and meanings: The use of qualitative data in social and personality psychology. In C. Sansone, C. Morf, & A. T. Panter (Eds.), *Handbook of methods in social psychology.* Thousand Oaks, CA: Sage.

O'Connor, R., White, A., & Swanson, H. (2007). Repeated reading versus continuous reading: Influences on reading fluency and comprehension. *Exceptional Children, 74,* 31–46.

Pring, R. (2004). *Philosophy of educational research* (2nd ed.). New York: Continuum.

Slavin, R. E. (2004). Education research can and must address "What works" questions. *Educational Researcher, 33,* 27–28.

Tashhakkori, A., & Teddlie, C . (2003). *Handbook on mixed methods in the behavioral and social sciences.* Thousand Oaks, CA: Sage.

Thomas, R. M. (2003). *Blending qualitative and quantitative research methods.* Thousand Oaks, CA: Sage.

CHAPTER 3

# The Research Problem

A research problem is not a nuisance; it is a step toward new knowledge.

## INSTRUCTIONAL OBJECTIVES

After studying this chapter, the student will be able to:

1. Define a research problem.
2. Identify potential sources of problems for educational research.
3. State the criteria to use for evaluating a research problem.
4. State the characteristics of a worthwhile theory.
5. Evaluate a given problem for research using the accepted criteria.
6. Take a general problem in an area of interest in education and formulate it in a specific form ready for empirical investigation.
7. Distinguish between the types of problem statements used in quantitative research and qualitative research.
8. Define terms such as *population* and *variables* as used in a quantitative research study.
9. Identify the population and the variables in a given study.

Systematic research begins with a **research problem.** In a classic work, John Dewey (1933) spoke of the first step in the scientific method as the recognition of a felt difficulty, an obstacle, or problem that puzzles the researcher. Your first step in the research process is therefore to select a problem for investigation. Selecting and formulating a problem is one of the most important aspects of doing research in any field. Beginning researchers are often surprised to find that this initial stage can take up a large part of the total time invested in a research project. There is no way to do research until a problem is recognized, thought through, and articulated in a useful way.

A researcher must first decide on the general problem area. This step is often difficult for beginning researchers. The difficulty is not due to a shortage of problems but, rather, to the fact that beginners must select a problem very early, when their understanding of how to do research is most limited. They are uncertain about the nature of research problems and how to go about solving them. Skill in doing research is to a large extent a matter of making wise choices about what to investigate. This skill takes time and repeated effort to develop, but the willing beginner can do it.

In order to ask questions that research can answer, one should have knowledge or experience in an area. We often hear students in difficult courses say, "I don't know enough to ask questions." Similarly, unless a researcher has knowledge or experience in an area, he or she does not know what additional knowledge is needed or how to obtain it through empirical investigation.

Furthermore, the question chosen for investigation should hold deep interest or be one about which the researcher is really curious. The choice must necessarily be very personal or else the researcher may find it difficult to muster the motivation to carry the research through to its end. Find a question that intrigues you and you will enjoy the search for a solution. For example, an elementary school teacher may be interested in finding a more effective way to teach reading. A high school biology teacher may want to know if using computer simulations would improve students' problem-solving skills. An elementary school principal may want to know if a mentoring program would improve the effectiveness of beginning teachers.

After having chosen the general area of investigation, the researcher then narrows it down to a specific statement of the research question. What specifically do you want to know or what do you want to predict? Unlikely as it may seem, once the researcher has selected a problem area and clearly articulated a question or statement, he or she has accomplished one of the most difficult phases of the research process.

## SOURCES OF PROBLEMS

The first question most students ask is "How do I find a research problem?" Although there are no set rules for locating a problem, certain suggestions can help. Three important sources for research problems are experience, deductions from theory, and related literature. Noneducation sources may also be useful. These sources are appropriate in both quantitative and qualitative research.

### EXPERIENCE

Among the most fruitful sources for beginning researchers are their own experiences as educational practitioners. Teachers have intuitions or hunches about new relationships or why certain things in school happen the way they do. Teachers often question the effectiveness of certain classroom practices that have become routine but that may be based more on tradition or authority than on scientific research. They wonder if alternative procedures would be more effective. A high school teacher might have a question about strategies to improve the achievement of at-risk students, or an elementary teacher may have questions about a new method to teach reading. In this age of accountability in education, teachers want to know if programs and practices they use are the most effective. Research can provide the answers to such questions.

Most graduate students in education have been in the classroom or are currently working full- or part-time in schools. Students who have not had teaching experience can get ideas from discussions and their reading in education courses. We recommend that you make a list of ideas, noting things that you question. By studying these notes, you will soon identify a worthwhile research problem.

## THEORIES

Theories are a good source of problems for research. A *theory* may be defined as a set of interrelated statements, principles, and propositions that specify the relationships among variables. The application of the general principles embodied in a theory to specific educational problems is only hypothetical, however, until research empirically confirms them. For example, assume a researcher is interested in how adolescents form their academic self-concepts. Social comparison theory suggests that students form academic self-concepts by comparing their self-perceived academic accomplishments to some standard or frame of reference. The frame of reference for most students would be the perceived academic abilities of their classmates. One question that might arise is "Would gifted students placed in selective homogeneous classes have lower academic self-concepts over time than equally gifted students in heterogeneous or mixed-ability classes?" This question could be investigated by studying the change over time in the academic self-concept of gifted students in homogeneous classes compared with that of gifted students placed in regular, heterogeneous classes.

Another interesting theory that has implications for education is Erik Erikson's (1967) classic theory of personality development. Erikson describes psychosocial development in terms of stages throughout the life span, each of which involves a critical issue or conflict that the person must resolve. Adolescence, which is one of these stages, has as its major task the development of a positive self-concept or, to use Erikson's term, a strong sense of identity. Forming a strong personal identity is difficult because competing roles and values face the young person. Research shows that adolescents who have achieved a sense of identity are more independent, more socially competent, better able to cope with stress, and have higher self-esteem. However, if the adolescent does not resolve the identity crisis, a sense of inferiority and personal alienation may result. It is interesting that students who have committed violent acts often report feelings of alienation. Erikson's theory could become the foundation for research on school violence. A researcher interested in studying school violence might ask, "Are there school practices that may contribute to feelings of isolation in some students? "What are some positive programs that might help improve students' self-image?" "How does the school deal with reported incidences of physical bullying or cyber-bullying?" and "Would other procedures be more effective?" A qualitative researcher might conduct a case study of an adolescent who has committed an act of school violence or of one who has been a victim of bullying.

### Choosing a Theory

Not all theories are equally useful to a beginning researcher. Let us examine some of the characteristics one searches for in a good theory for a research study:

1. *An essential characteristic of a good theory is that it is testable*. The theory chosen should be one from which the researcher can make concise

predictions (hypotheses) about what will happen in new situations and can verify these predictions through empirical observation. As the hypotheses are supported in research studies, they then become part of the theory that adds to the body of knowledge. However, if the theory cannot be tested, it serves no useful purpose.

2. *A good theory is not only testable but also falsifiable*. Being falsifiable means that it is capable of being proven wrong. It is possible to gather evidence that contradicts the theory. A theory that explains why a tornado touched down in a certain area of a town by stating that the people there are being punished for their sins is not a theory that can be proven wrong. Thus, it is not a useful theory.

   Students sometimes find this concept of **falsifiability** difficult to understand. This concept derived from the philosopher Sir Karl Popper, who in *Logic of Scientific Discovery* (1965) argued that claims to knowledge "can never be proven or fully justified, they can only be refuted" (p. 40). A theory cannot ever be proved to be true because theories are generalizations that apply to all possible instances of the phenomena they are trying to explain, and it is not possible to test it against all possibilities. We say only that a theory has been supported; the more support it gets in a variety of research studies, the more confidence we have in the usefulness of the theory. However, it is possible to disprove a theory by gathering negative evidence that contradicts the theory. According to Popper, this is how most scientific progress is achieved. Neuman and Kreuger (2003) give a useful example: "If I want to test the claim that all swans are white, and I find 1000 white swans, I have not totally confirmed the causal law or pattern. All it takes is locating one black swan to refute my claim—one piece of negative evidence" (p. 40). Negative evidence indicates that the theory needs to be rejected or at least revised. To summarize, a good theory is one for which evidence can be gathered that will either support or refute the theory. Both outcomes must be possible.

3. *A good theory deals with some significant phenomenon or behavior that needs explanation, such as learning or motivation*.

4. *A good theory provides the simplest, clearest, and most plausible explanation for the phenomenon*. A good theory follows the principle of parsimony, which states that a theory should explain the largest number of facts with the smallest number of principles.

5. *A good theory has internal consistency*; its propositions do not contradict one another. For example, a "commonsense" theory of human separation may state "Absence makes the heart grow fonder" but also "Out of sight, out of mind." One could find evidence to support both of these propositions; thus, the theory would not be useful for predicting what might happen when people are separated.

In summary, think of an educational, psychological, or sociological theory that you find especially interesting. Read a summary of the theory in journals, textbooks, or primary sources, and then ask a question. A theory-based research question is beneficial because the results can be tied to a body of existing knowledge. The research can verify or fail to verify the theory, and it will most likely suggest other questions for research. You might talk to your professors to find out what they are working on or to get their suggestions.

## RELATED LITERATURE

Another valuable source of problems is the published literature in your area of interest. In published research, you will find examples of research problems and the methods used to solve them. A review of related literature may help in the following ways:

1. You may find a study that needs to be replicated. You can repeat someone else's study, not exactly, but with some variation. You might use a different age group, different setting, or a different methodology. A study might even become a cross-cultural one to determine if the conclusions from research in one culture apply in other cultures. You need not be concerned that you are merely replicating a study rather than doing some new, groundbreaking research. Replication is a worthwhile activity because it provides more evidence of the validity of the original findings. As studies are repeated at different times and in different places, with the findings supported in each study, we can have increasing confidence in the scientific validity of the findings.

   For example, researchers have conducted numerous replications of Piaget's famous studies (1999) of the development of moral judgment in children. These studies have used Piaget's basic approach but have investigated the development of moral judgment in children of different socioeconomic classes, in children of the same chronological age but differing in intelligence level, in children differing in the extent of their participation in their own age groups, in children differing in the nature of parental discipline experienced in the home, and in both boys and girls. Recently, other investigators have used techniques that differed from Piaget's in their attempts to confirm his findings and conclusions. In general, the large body of research stemming from Piaget's investigations has supported his original conclusions. Thus, a single research study, if it deals with a significant problem and if its findings are exciting, can inspire many other studies.

2. You may find a question that represents the next logical step in the research on a problem. The outcomes of one piece of research very often lead to new questions. In the concluding sections of their research reports, researchers often describe new questions that have arisen and suggest additional studies that should be done. A productive way to extend studies is to introduce new variables into a research design for further control and for determining interaction effects among variables. Many multivariate studies are extensions of earlier single variable investigations (see Chapter 11).

In conclusion, published research can be a great source of ideas for research. With some critical analysis of the research in your field and a bit of creativity, you should be able to find several potentially researchable problems. Reading research will also help you by showing how previous researchers measured variables, selected samples, analyzed data, and so on.

### Reviews of Research

Reviews of research that integrate and summarize studies on specific topics can be very useful for identifying a research problem. Those produced by the American Educational Research Association (AERA) are particularly useful.

The *Review of Educational Research* published quarterly by AERA since 1931 reviews and integrates educational literature on a different topic each volume. For example, the topic of volume 71 (2007) was "Difference, Diversity, and Distinctiveness in Education and Learning."

In 1973, AERA launched the annual *Review of Research in Education* to provide summaries of what research has been done, is being done, and needs to be done in a specific broad topic each year. Volume 82 (2008) focused on "What Counts as Knowledge in Educational Settings: Disciplinary Knowledge, Assessment, and Curriculum."

Approximately every 10 years, AERA publishes the *Handbook of Research on Teaching* (Gage, 1963; Travers, 1973; Wittrock, 1985; Richardson, 2001). These volumes list, summarize, and critically analyze research in the field of teaching. Each edition contains authoritative articles by specialists on selected topics in the field. The fourth edition is composed of 51 chapters from 81 authors, all of whom are experts in their respective fields. Comprehensive bibliographies are included, by selected topics. Among the topics in the fourth edition are policies for licensing and assessment of teachers, special education, middle school teaching, teaching as a moral activity, and the teaching of physical education.

AERA's *Encyclopedia of Educational Research* (2004), designed to present "a critical synthesis and interpretation of reported educational research," contains signed articles with bibliographies providing well-documented discussions of recent trends and developments, as well as traditional topics. This four-volume encyclopedia includes approximately 200 topics. It is a good basic source for preliminary overviews of research in various areas.

Other useful periodicals include *PsycINFO*, *Education Abstracts*, and specialized reviews such as Social Work Abstracts, Historical Abstracts, and MEDLINE (medicine). For more information on related literature, see Chapter 4.

## NONEDUCATION SOURCES

You can adapt theories or procedures you encounter in other fields to apply to education. Often, movements that originate outside a profession lead people to new paths of research. The women's movement has led researchers to study gender stereotyping in educational materials, the influence of schools on the learning of sex roles, gender differences in achievement and personality, and so forth. The civil rights movement led to many studies about the education of minority children. The AIDS (acquired immunodeficiency syndrome) epidemic has stimulated a great deal of research on the best procedures and materials to use to acquaint young people in school with the danger of the disease and how best to protect themselves from it. The inspiration for much valuable research in education has come from such noneducation sources.

## QUALITATIVE RESEARCH PROBLEMS

Just as is true for quantitative researchers, beginning qualitative researchers can look to their personal experiences and interests, to theory, to the professional literature, or to current social issues and real-world concerns to find a potential problem. You need to identify an area or a topic about which you have a real interest. For example, a beginning researcher might be interested in how learning-disabled students are integrated into regular high school classrooms.

Once researchers have selected the initial focus of inquiry, they need to identify exactly what they want to know about that topic. The focus of inquiry is thus narrowed to the aspect of the phenomenon that will be explored in the research study. The focus of inquiry mentioned previously can be stated as follows: "How do other students treat learning-disabled students?" "How do the learning-disabled respond?" Although the qualitative researcher intuitively arrives at hunches about the phenomenon, he or she does not formulate an initial hypothesis that the study tests.

Suppose one is interested in the general topic of bullying behavior in elementary schools. Elementary teachers have long observed that some students are bullies and others become their victims. In a qualitative study, a researcher might ask how and why this behavior develops and could use naturalistic observation to investigate this behavior in an elementary school. The investigator could use video cameras and remote microphones to record instances of children being exposed repeatedly to negative verbal or physical actions on the part of one or more classmates. The researcher would want to interview the bullies to find out what they are thinking and what their motives and goals are. The victims would also be interviewed to learn about their feelings. The researcher might also examine gender differences in bullying behavior and the reaction of peers to this behavior.

## EVALUATING THE PROBLEM

After you have tentatively selected a question that interests you, you need to ask if it is a question that warrants an expenditure of time and effort to investigate. The following are criteria that one can use to evaluate a research problem:

1. *The problem should have significance*—that is, it should be one whose solution will make a contribution to educational theory or practice. The problem may fill in gaps in current knowledge or help resolve some of the inconsistencies in previous research. You should be able to answer the question "So what?" with respect to your proposed study. Would the solution make any difference to educational practice? Would other educators be interested in the findings? Would the findings be useful in an educational decision-making situation?
2. *The problem should be one that will lead to new problems and so to further research*. A good study, while arriving at an answer to one question, usually generates a number of other questions that need investigation. Avoid trivial problems that have little or no relationship to theory or previous research.

   We suggest that a beginning researcher consider selecting a problem that could possibly be expanded or followed up later in a master's thesis or even a doctoral dissertation. It may be helpful if students familiarize themselves with the research efforts of their professors, who not only can suggest related problems needing investigation but also may later serve as a mentor or a doctoral committee member.
3. *The problem must be researchable*. Although this criterion would seem self-evident, in practice, many proposed problems are not researchable. A researchable problem is one that can be attacked empirically; that is, it is possible to gather data that answer the question. Many interesting questions in education cannot be answered by scientific research. Philosophic questions, for example, that ask what should be done are not researchable and should

be avoided. Questions such as "Should we offer more vocational training in the high school?" or "Should schools give more attention to character education?" cannot be answered by scientifically gathering and analyzing data.

It is possible to restate philosophic questions to make them researchable. The previous question could be restated as follows: "What is the effect of a character education program on the incidence of cheating in high school?" It would be possible to gather data on this question, which could then be used by educators to help make decisions about a character education program.

4. *The problem should be suitable for the researcher*. The problem may be excellent from the standpoint of the previous criteria but inappropriate for the individual. First, the problem should be one in which you, the researcher, have a genuine interest and about which you can be enthusiastic. It should be a problem whose solution is personally important because of what it could contribute to your own knowledge or to improving your performance as an educational practitioner. Unless the problem is meaningful and interesting, it is doubtful whether you would be willing to expend the time and energy to do a thorough job.

   In addition to interest, one must have the necessary research skills to carry the study through to completion. One may have to develop and validate instruments or do complex statistical analyses. Another consideration is whether you will have access to participants and the data necessary to answer the research question. Lastly, one should choose a problem that can be investigated in the allotted time and with the resources available. Do not select a problem that is too large or too involved, and be sure to allow adequate time for constructing instruments, administering instruments, conducting interviews or observations, analyzing data, and writing the report.

5. *The problem should be ethically appropriate*. That is, the problem should be one that you can investigate without violating ethical principles. Unlike researchers in the physical sciences, educational researchers are dealing with human subjects with feelings, sensitivities, and rights who must be treated ethically. We discuss ethics in greater detail in Chapters 15 and 20. At this point, we mention briefly three issues the researcher should consider:

   a. *Consent*. Researchers need to obtain consent from the intended subjects. Subjects should be able to choose whether they wish to participate in the study or not. Obtain consent from subjects after taking steps to ensure that they have a complete understanding of the procedures to be used, any risks involved, and any demands that will be placed on them. Obtain parental consent if minor children are to be involved in the study.

   b. *Protection from harm*. Do not plan research that may cause physical harm or psychological harm such as stress, discomfort, or embarrassment that could have lasting adverse effects. Fortunately, most educational research does not involve great risk of harm to subjects. However, the potential for harm always exists, and a researcher should be prepared if a participant requests counseling or other help after participating in the study.

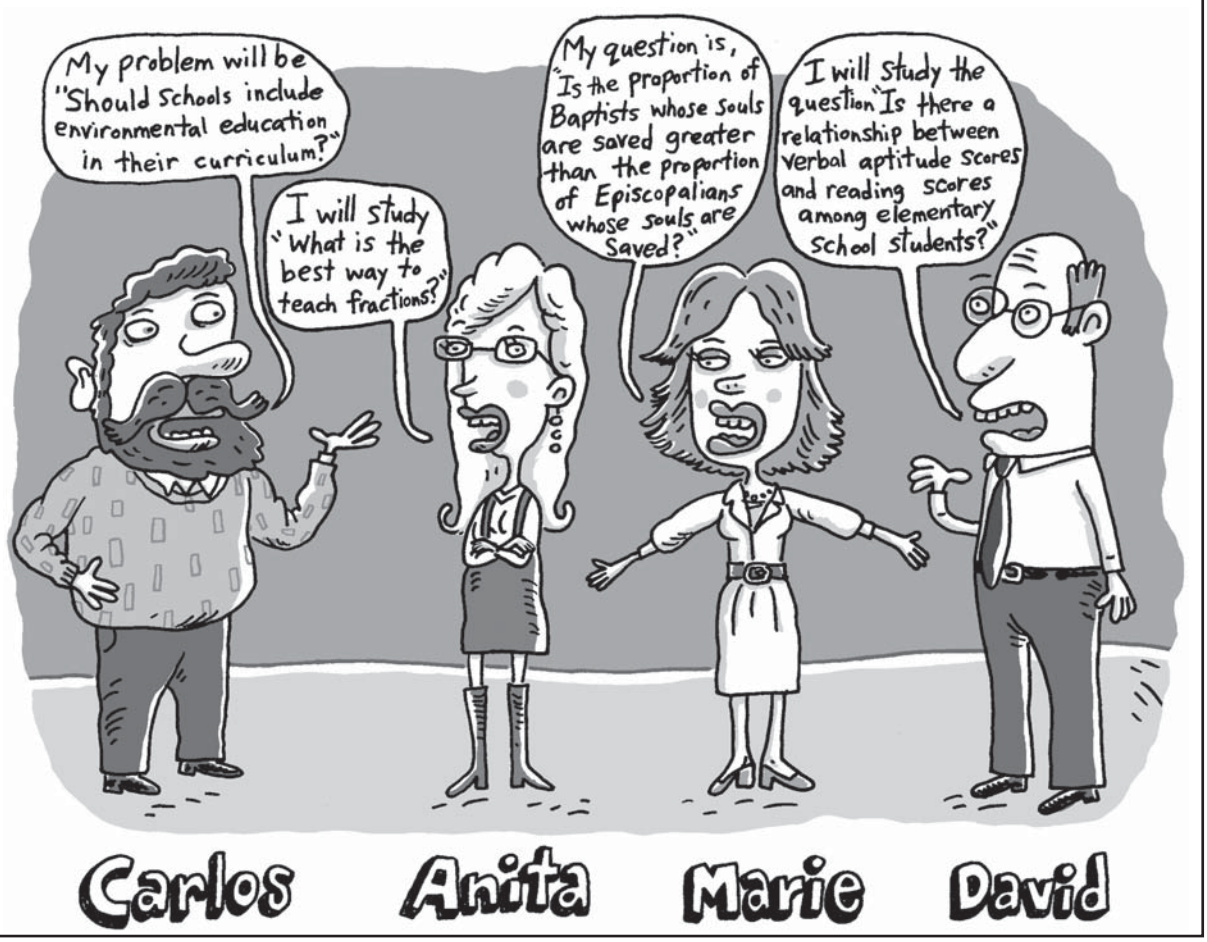
   c. *Privacy*. A researcher should invade the privacy of subjects as minimally as possible. For example, a researcher may plan to use an inventory that

asks adolescents questions about sexual experiences, religious beliefs, attitudes toward parents, or other sensitive topics. In this case, the researcher should not attach names to the inventories. Subjects have the right to expect that their anonymity will be preserved. Most educational researchers are interested in group data rather than individual responses; the scores or responses of individuals are generally pooled and reported as group averages, which tends to minimize the risk of invading privacy. Table 3.1 summarizes the criteria of a good research problem.

**Table 3.1** Characteristics of a Good Research Problem

1. The problem is significant (it will contribute to the body of knowledge in education).
2. The problem is one that will lead to further research.
3. The problem is researchable (it can be investigated through the collection of data).
4. The problem is suitable (it is interesting and suits the researcher's skills, time, and available resources).
5. The problem is ethical (it will not cause harm to subjects).

**PICTURE THIS**




Joe Rocco

**THINK ABOUT IT 3.1**

How do the questions in the cartoon rate on the criteria for evaluating research problems?

**Answers:**

1. Carlos: Research cannot answer questions of "should." The question could be rewritten as "Do students who have had a unit on environmental awareness demonstrate greater knowledge of and more positive attitudes toward environmental issues than students who have had a control unit?" Then one could randomly assign some students to have a unit on environmental awareness while others have a unit not related to the environment. At the completion of the units, one could measure the students on their knowledge of the environment, their attitudes toward environmental legislation, and environmentally appropriate behaviors such as not littering.
2. Anita: As stated, the question is not researchable. There are so many possible ways to teach fractions that one could never investigate the outcomes of them all. One could operationally define two or three methods and compare the success of students taught by the different methods, using the same test of knowledge of fractions for all groups.
3. Marie: There is no way in this world to determine whose soul has been saved. A feasible question might be "Is the proportion of Baptists who say their souls have been saved different from the proportion of Episcopalians who say their souls have been saved?"
4. David: The question is trivial because it has been investigated sufficiently in past research.

## STATING THE RESEARCH PROBLEM

After you have selected and evaluated the problem, the next task is to state the problem in a form amenable to investigation. We cannot overemphasize the importance of a clear statement of the problem. Beginning researchers often have a general idea of what they want to investigate but have trouble articulating it as a workable problem. They cannot make progress until they can state unambiguously what they are going to do. The statement of the problem varies according to the type of research. Thus, we consider quantitative and qualitative research statements separately.

### THE PROBLEM STATEMENT IN QUANTITATIVE RESEARCH

The **problem statement** in quantitative research specifies the variables and the population of interest. The problem statement can be a declarative one such as "This study investigates the effect of computer simulations on the science achievement of middle school students." The statement can ask a question about a relationship between the two (or more) variables. The previous problem might be restated as "What is the relationship between use of computer simulations and achievement in middle school science?" Some scholars prefer the question form simply because it is straightforward and psychologically seems to orient the researcher to the task at hand—namely, to find the answer to the question. But either is an acceptable way to present the research problem.

The problem can be further clarified by operationally defining the variables involved. In the previous example, you might specify what computer simulations will be used, how science achievement will be measured, and how the sample of middle school students will be selected. The problem statement then becomes "What is the effect of a computer-assisted biology course on performance on the Test of Biological Concepts of students in an eighth-grade biology class?" One can then proceed to plan an experiment that compares the scores on the Test of Biological Concepts by students having the computer instruction with those of similar students having the traditional biology curriculum.

### THE PROBLEM STATEMENT IN QUALITATIVE RESEARCH

Qualitative researchers also begin with a problem, but they state it much more broadly than in quantitative research. A qualitative problem statement or question indicates the *general* purpose of the study. Formulation of a qualitative problem begins with the identification of a general topic or an area you want to know more about. This general topic of interest is sometimes referred to by qualitative researchers as the **focus of inquiry**. This initial broad focus provides the framework but allows for changes as the study proceeds. As the researcher gathers data and discovers new meanings, the general problem narrows to more specific topics and new questions may arise. For example, Piert (2007) conducted a qualitative study to learn about a rite-of-passage program for transitioning black young people into adulthood. The program draws upon African traditional culture to impart values, improve self-concept, and develop cultural awareness. The specific problem was "What are the perceptions of former students who had experienced a rite-of-passage program while attending an African centered high school?" The findings suggest that the participants perceived the rite of passage as a community endeavor that facilitated their transition into adulthood. They viewed it as a method to inculcate social, cultural, and political values that will ensure the positive development of black young adults within the black community as well as American society.

Whereas the quantitative researcher always states the problem before collecting data, the qualitative researcher may formulate problems after beginning to collect data. In fact, the researcher often does not present the final statement of the problem—which typically specifies the setting, subjects, context, and aim of the study—until he or she has collected at least some data.

In qualitative research, the statement may be somewhat general in the beginning, but it will become more focused as the study proceeds. After exploring the sites, the people, and the situations, the researcher narrows the options and states the research problem more specifically.

## IDENTIFYING POPULATION AND VARIABLES

A good strategy for shaping a felt problem—or a vague notion of what you want to investigate—into a researchable problem is to think in terms of population and variables. For example, let us consider Ms. Burke, an elementary school principal whose question is "Does individual tutoring by upper-grade students have a positive effect on the reading achievement of younger below-average readers?"

It is usually easiest to identify the **population**—those people about whom you wish to learn something. The population here is below-average readers. Reading ability is not a variable in this question because all the children being considered have already been diagnosed as below-average readers. Having identified below-average readers as the population in the original statement, Ms. Burke should now ask herself if that is really the population she wants. She will probably decide that below-average readers is too broad a category and she should confine herself to a particular age. Thus, she selects below-average second-grade readers.

Now she is ready to identify the variables in the remainder of her original statement. "Individual tutoring" can be made into a variable by varying the type of tutoring used, varying the amount of tutoring time, or having some children receive the tutoring and others not receive tutoring. Ms. Burke decides that the last alternative concerns what she really wants to know, so she rewrites the relevant part of the question to "Does receiving a specified amount of individual tutoring versus no tutoring … ?" Thus, tutoring is the independent variable because it is antecedent to reading achievement, and the principal is predicting that the tutoring will have an effect on reading achievement, the dependent variable. Recall that the dependent variable is the outcome of interest, and the independent variable is hypothesized to influence the dependent variable. Now it becomes obvious that the word *tutoring* is too general. Unless all subjects receive the same type and amount of tutoring, the results of the study will be meaningless. Ms. Burke decides to use word flash drill as the specific type of tutoring and to specify 15 minutes per day as the amount of time.

The phrase "have a positive effect on" is quite vague until she considers it in terms of her independent variable. Does word flash drill have an effect on … what? She knows it has an effect on word flash recall, but she wants to study its effects on other aspects of reading behavior that might be observed: expressive oral reading, silent reading, positive feelings toward reading, number of books read, comprehension, and so forth. However, she is afraid that teachers might rate good word callers as comprehending more and being more positive toward reading, whereas they view the poorer word callers as more inferior on these variables than they really are. She wants a dependent variable that is independent of teacher judgment and decides to use reading scores from the California Achievement Test (CAT) as the dependent variable.

Ms. Burke's revised statement of the problem now reads "Among below-average second-grade readers, is there a difference in CAT reading scores between those who have received 15 minutes per day of individual word flash drill by upper-grade students and those who have received no word drill?" This question tells whom she is studying, what will be done differently for some students, and what she expects differential treatment to influence. Note also that the value judgment "positive effect" has been dropped from the question.

It is often useful to follow this procedure in a formal manner similar to that used for diagramming a sentence. You can begin by drawing a vertical line and writing *Population* to the left and *Variables* to the right. Then list these elements in the study below the horizontal line. For the preceding example, the diagram would be as follows:

| Population | Variables |
|---|---|
| Below-average second-grade readers | • Word flash drill for 15 minutes daily by upper-grade students versus no word flash drill (independent)<br>• Reading scores on CAT (dependent) |

Let us take another question: "What is the effect of having experienced versus not having experienced a preschool program on the reading achievement of first-graders?"

| Population | Variables |
|---|---|
| First-graders | • Having experienced versus not having experienced a preschool program (independent)<br>• Reading achievement (dependent) |

This question is complete in that it has an identified population and both independent and dependent variables. Because "preschool program" precedes "reading achievement of first-graders," the former can be identified as the independent variable and the latter as the dependent variable.

Let us look at another example: "Does high school driver education do any good?" As it stands, the question has neither a population nor variables. An investigator starting with this question might first decide to compare 18-year-old drivers who have had high school driver education with those who have not. You now have a population statement and an independent variable. Now you can turn your attention to selecting a dependent variable. What effect might having versus not having driver education have on 18-year-old drivers? Let us say you decide that "accident rate" would be a suitable dependent variable. Putting these elements into a diagram, you now have the following:

| Population | Variables |
|---|---|
| 18-year-old drivers | • Have had versus have not had high school driver education (independent)<br>• Accident rate (dependent) |

You can now state a complete question: "Do 18-year-old drivers who have had high school driver education have a lower accident rate than 18-year-old drivers who have not had high school driver education?"

The question "What is the relationship of dogmatism to political attitudes among college freshmen?" illustrates another point. Consider this diagram:

| Population | Variables |
|---|---|
| College freshmen | • Dogmatism<br>• Political attitudes |

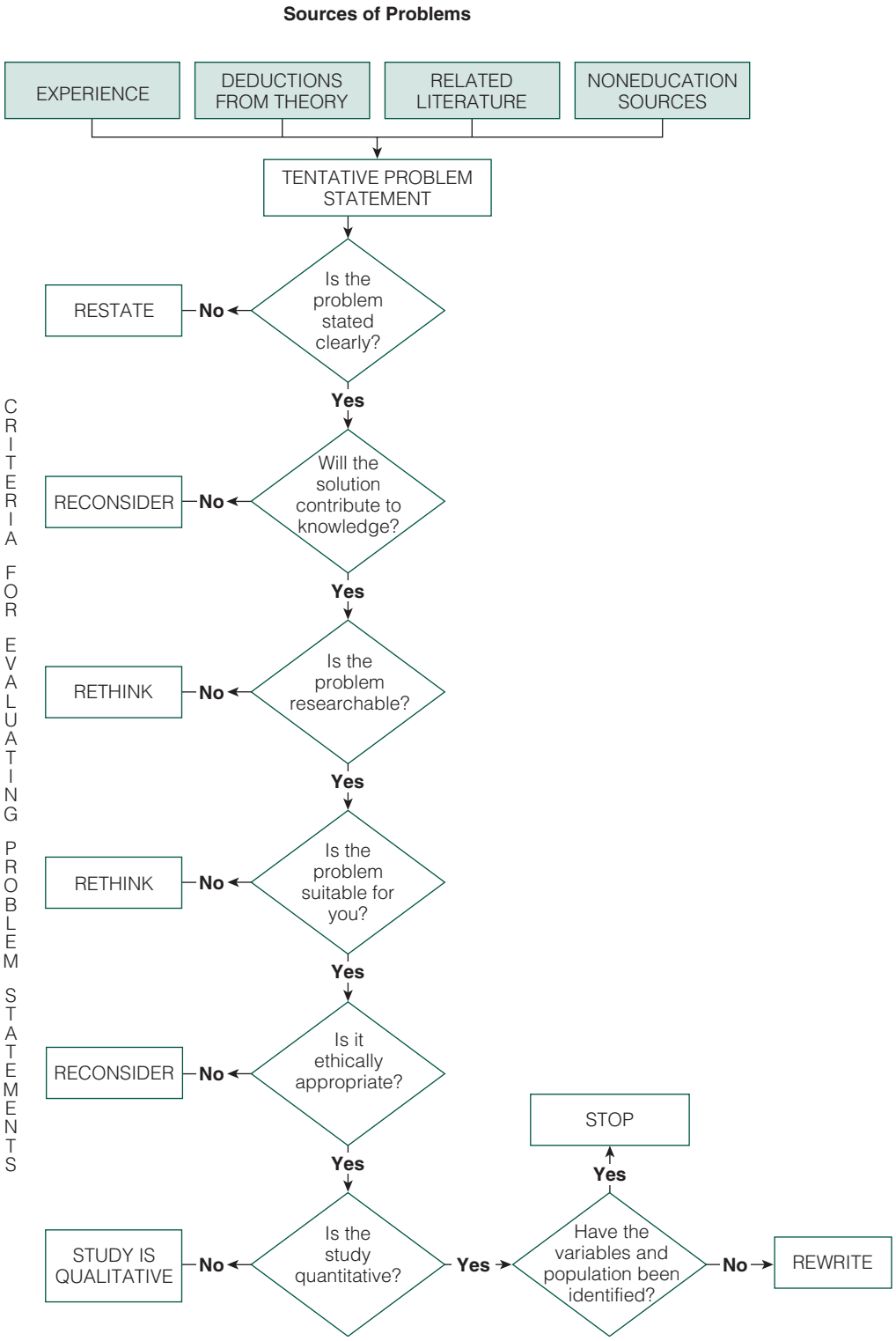



**Figure 3.1** Developing a Research Problem

This question is complete with a population and two variables. However, you cannot label the variables as independent and dependent because it cannot be determined which is antecedent to the other.

If you conduct a study to investigate status quo rather than a relationship between variables, it may be complete with only one variable. For example, you might study the opinions of college seniors concerning legalization of marijuana. In this case, the population is college seniors and the single variable is their opinions on the subject. The process of evaluating a problem for research is summarized in Figure 3.1.

### Think About It 3.2

For the revised problems of Carlos, Anita, and Marie, identify and/or add the independent variable(s), dependent variable, and population that they will need to identify in order to begin research on their questions.

**Answers**

1. Carlos

| *Population* | *Independent Variable* | *Dependent Variables* |
|---|---|---|
| High school students | Unit on environmental awareness vs. control unit | Knowledge of environment<br>Attitude toward environmental issues |

2. Anita

| *Population* | *Independent Variable* | *Dependent Variable* |
|---|---|---|
| Fourth-graders | Method of teaching fractions | Student success with fractions |

3. Marie

| *Population* | *Independent Variable* | *Dependent Variable* |
|---|---|---|
| Church members (Baptists and Episcopalians) | Religious affiliation | Whether they report their souls saved or not |

## SUMMARY

The first task facing researchers is selecting a researchable problem and stating it in a form suitable for research. To find a problem, investigators may look to their personal experiences, to theories from which questions may be deduced, to the current literature in their area of interest, or to noneducation sources. They must evaluate the significance of the proposed problem in terms of specific criteria, asking questions such as "Will the problem contribute to the present body of knowledge?" "Does it have potential for leading to further research?" "Is it testable—that is, can the variables be observed and measured?" "How appropriate is the problem with respect to my interests, experience, and knowledge in the area?" "Do I have access to the data required by the problem, and are instruments available, or could they be constructed, to measure the variables?" and "Can the data be analyzed and interpreted within

the time available?" The question should not directly involve philosophical issues, nor should it be so general that a research undertaking is impossible. A quantitative research question asks about the relationship between certain variables. The statement of the question should identify the population of interest and the variables to be investigated. A qualitative research question indicates the general purpose of the study. The criteria for evaluating qualitative problems are similar to those used for quantitative research problems.

## KEY CONCEPTS

criteria for research problems
falsifiability
focus of inquiry
population
problem statement
research problem

## EXERCISES

1. The following is an abstract taken from an article in *The Clearing House* (Springer, Pugalee, & Algozzine, 2007):

   *In U.S. schools, students must pass statewide competency tests to graduate from high school. In this article, the authors summarize the development and testing of a program implemented to improve the skills of students failing to "make the grade" on these high-stakes tests. District personnel randomly assigned 28 students who previously failed the math test to participate in an experimental (Arizona Instrument to Measure Standards [AIMS]) or to a control math class. The AIMS group used a computerized tool to generate multiple-choice problems for students to practice the content of the state's competency test. Eight AIMS students (57 percent) and two control students (14 percent) passed the retest. The outcomes offer promise for schools looking for evidence-based solutions to problems related to increasing numbers of students experiencing difficulties with high-stakes assessments.*

   **a.** What was the independent variable in this study?
   **b.** What was the dependent variable?
   **c.** What was the population?
2. Find a quantitative research report published in a journal, and answer the following questions based on your reading:
   **a.** What problem is investigated in the study?
   **b.** What are the independent and dependent variables?
   **c.** Where did you find the problem stated in the report?
   **d.** Was the problem stated with sufficient clarity so that you knew exactly what was being investigated in the study?
   **e.** Did the author suggest further research on this question?
3. Find a qualitative research report published in a journal and identify the
   **a.** Problem
   **b.** Methodology
   **c.** Findings
   **d.** Conclusions
4. Select a broad area in which you might be interested in doing research and then identify a research problem in that area. State this problem in an acceptable form for research. What was the source of this problem?
5. The following examples are inadequate statements of research problems. Restate each so that it becomes a specific question suitable for research.
   **a.** A later morning start for the high school
   **b.** Parental involvement and elementary school achievement
   **c.** Self-concept of children with learning disabilities
   **d.** Home-schooled adolescents
   **e.** Gender differences and logical thinking skills of talented preadolescents
   **f.** Teaching students with attention deficit/hyperactivity disorder (ADHD)
   **g.** Predicting achievement in a graduate education program

6. Evaluate the following research problems:
   a. Should learning-disabled students be mainstreamed in middle school English classes?
   b. Has No Child Left Behind legislation been good for U.S. education?
   c. Would it be wise to have a dress code for the city's high schools?
   d. What is the relationship between verbal aptitude scores and reading test scores?
7. State the most likely independent and dependent variables in the following studies:
   a. Peer interactions in children with autism
   b. The effect of participation in school sports on the social skills of mentally challenged adolescents
   c. The effect of an early intervention program on the academic achievement of children from low-income families
   d. The influence of an antismoking program on the attitudes of middle school students toward smoking
8. A principal wants to know if it is beneficial to keep class sizes small in kindergarten to grade 2. Write an appropriate research question designed to answer this question.
9. Classify the following studies as most likely being quantitative or qualitative:
   a. Life of an Adolescent with HIV Infection
   b. High School Principals' Perspectives on Student Expulsion
   c. The Effect of Ability Grouping on Academic Outcomes for Gifted Students
   d. Racial Stereotypes in Middle School Literature Textbooks
   e. Teaching in an Alternative High School
10. There has been an increase in the number of elementary school children diagnosed with attention deficit disorder. Based on your observation and experience, formulate a theory to explain this finding. Evaluate your theory according to the characteristics of a "good" theory. State a research problem based on your theory.

## ANSWERS

1. a. Computerized program to practice math content versus control math class
   b. Performance on a repeat of a state math competency test
   c. Students who had failed the math competency test
2. Answers will vary.
3. Answers will vary.
4. Answers will vary.
5. a. What is the effect of a later morning start time on the achievement of students at Eastern High School?
   b. What is the effect of a program to increase parental involvement on the achievement of their elementary school children?
   c. How do children with learning disabilities perceive themselves socially and academically?
   d. How do students who have been home schooled perform academically during the freshman year at Central State U.?
   e. What is the relationship between gender and logical thinking skills in talented preadolescents?
   f. What is it like to teach students with ADHD?
   g. What are the best predictors of achievement in a graduate education program?
6. a. This question involves a value judgment that is impossible to investigate empirically.
   b. This question as stated involves a value judgment and cannot be investigated empirically.
   c. Research cannot answer questions of value; it can only provide information on which decisions can be based.
   d. This question has been thoroughly investigated in previous research; it would not contribute to the body of knowledge.
7. a. *Independent*: having autism; *dependent*: peer interactions
   b. *Independent*: participation in school sports; *dependent*: social skills of mentally challenged adolescents
   c. *Independent*: experiencing an early intervention program or not; *dependent*: academic performance

**d.** *Independent*: antismoking program; *dependent*: students' attitudes toward smoking

**8.** What is the effect of class size in grades K–2 on students' academic performance?

**9. a.** Qualitative
**b.** Quantitative
**c.** Quantitative
**d.** Qualitative
**e.** Qualitative

**10.** Answers will vary.

## REFERENCES

American Educational Research Association. (2004). *Encyclopedia of educational research* (7th ed.). New York: Macmillan.

Dewey, J. (1933). *How we think*. Boston: Heath. [Reprinted 2007 by Standard Publications]

Erikson, E. H. (1967). *Identity and the life cycle: Selected papers*. Psychological Issue Monograph Series, 1(1). New York: International Universities Press.

Gage, N. L. (Ed.). (1963). *Handbook of research on teaching*. Chicago: Rand McNally.

Neuman, W. L., & Kreuger, L. (2003). *Social work research methods: Qualitative and quantitative approaches*. Boston: Allyn & Bacon.

Piaget, J. (1999). *The moral judgment of the child*. New York: Routledge. [Reprinted 2007 by Kessinger Publishing]

Piert, J. H. (2007). Transition into adulthood: The experience of a rite-of-passage program at an African centered high school. *Negro Educational Review*, *58*(3/4), 169–186.

Popper, K. (1965). *Logic of scientific discovery*. New York: Basic Books.

Richardson, V. (Ed.). (2001). *Handbook of research on teaching* (4th ed.). New York: Macmillan.

Springer, S., Pugalee, D., & Algozzine, B. (2007). Improving mathematics skills of high school students. *The Clearing House*, *81*(1), 37–44.

Travers, R. (Ed.). (1973). *Handbook of research on teaching* (2nd ed.). Chicago: Rand McNally.

Wittrock, M. C. (Ed.). (1985). *Handbook of research on teaching* (3rd ed.). New York: Macmillan.

CHAPTER 4

# Reviewing the Literature

If I have seen further it is by standing on the shoulders of giants.

Sir Isaac Newton (1642–1727)

## INSTRUCTIONAL OBJECTIVES

After studying this chapter, the student will be able to:

1. Describe the role of related literature in quantitative research.
2. Identify the main functions of a literature review in quantitative research.
3. Describe the role of related literature in qualitative and mixed methods research.
4. Use Boolean logic to prepare a search statement.
5. Apply appropriate criteria to use in judging the merit of literature you find.
6. Identify indexes and abstracting serials especially useful to educators.
7. Describe and use primary databases for efficient location of literature.
8. Describe and use aggregated databases for efficient location of literature.
9. Describe sources of statistical data.
10. Identify useful government sources.
11. Detail a systematic progression of steps in organizing the literature, explaining the purpose of each step.
12. Understand the criteria to use in judging the merit of information on the Internet.

The search for related literature plays a vital but quite different role in qualitative and quantitative research. It must be completed early in quantitative research but not in qualitative research.

## THE ROLE OF RELATED LITERATURE IN QUANTITATIVE RESEARCH

Quantitative researchers are urged not to rush headlong into conducting their study. The search for related literature should be completed before the actual conduct of the study begins in order to provide a context and background that support the conduct of the study. This literature review stage serves several important functions:

1. *Knowledge of related research enables investigators to define the frontiers of their field.* To use an analogy, an explorer might say, "We know that beyond this river there are plains for 2000 miles west, and beyond those plains a range of mountains, but we do not know what lies beyond the mountains. I propose to cross the plains, go over the mountains, and proceed from there in a westerly direction." Likewise, the researcher in a sense says, "The work of A, B, and C has discovered this much about my question; the investigations of D have added this much to our knowledge. I propose to go beyond D's work in the following manner."
2. *A thorough review of related theory and research enables researchers to place their questions in perspective.* You should determine whether your endeavors are likely to add to knowledge in a meaningful way. Knowledge in any given area consists of the accumulated outcomes of numerous studies that generations of researchers have conducted and of the theories designed to integrate this knowledge and to explain the observed phenomena. You should review the literature to find links between your study and the accumulated knowledge in your field of interest. Studies with no link to the existing knowledge seldom make significant contributions to the field. Such studies tend to produce isolated bits of information that are of limited usefulness.
3. *Reviewing related literature helps researchers to limit their research question and to clarify and define the concepts of the study.* A research question may be too broad to be carried out or too vague to be put into concrete operation; for example, "What do parenting practices have to do with mental health?" A careful review of the literature can help researchers revise their initial questions so that the final questions can be investigated. The literature review also helps in clarifying the constructs involved in the study and in translating these constructs into operational definitions. Many educational and behavioral constructs—such as stress, creativity, frustration, aggression, achievement, motivation, and adjustment—need to be clarified and operationally defined. These, as well as many other educational and behavioral constructs, do not lend themselves to research until they can be quantified. In reviewing literature, you become familiar with previous efforts to clarify these constructs and to define them operationally. Successful reviews often result in the formation of hypotheses regarding the relationships among variables in a study. The hypotheses can provide direction and focus for the study.
4. *Through studying related research, investigators learn which methodologies have proven useful and which seem less promising.* The investigator develops increasing sophistication after digging through the layers of research that the related literature represents. As you delve into your topic, you soon see that the quality of research varies greatly. Eventually, you should begin

to notice that not all studies in any one field are necessarily equal. You will soon be critiquing studies and noticing ways in which they could be improved. For example, early studies in any one particular field may seem crude and ineffective because research methodology and design are constantly being refined with each new study. Even so, many research projects fail because they use inappropriate procedures, instruments, research designs, or statistical analyses. Becoming proficient at evaluating research to determine its worth helps the investigator discover the most useful research path.

5. *A thorough search through related research avoids unintentional replication of previous studies*. Frequently, a researcher develops a worthwhile idea only to discover that a very similar study has already been made. In such a case, the researcher must decide whether to deliberately replicate the previous work or to change the proposed plans and investigate a different aspect of the problem.
6. *The study of related literature places researchers in a better position to interpret the significance of their own results*. Becoming familiar with theory in the field and with previous research prepares researchers for fitting the findings of their research into the body of knowledge in the field.

As this discussion shows, quantitative research is built on a study of earlier work in the field, which helps the researcher refine his or her problem and place it in context. For qualitative researchers, the approach is very different. They are advised not to read in their area of interest because it is important that they approach their study without any preconceived ideas that might influence their work.

## THE ROLE OF RELATED LITERATURE IN QUALITATIVE AND MIXED METHODS RESEARCH

Barney G. Glaser, a pioneer in the grounded theory school within qualitative research, wrote (1978), "In our approach we collect the data first. Then start analyzing it and generating theory. When the theory seems sufficiently grounded and developed, then we review the literature in the field and relate the theory to it through integration of ideas" (p. 31). Glaser added, "It is vital to read but in a substantive field different from the research. This maximizes the avoidance of pre-empting, preconceived concepts" (p. 31).

A grounded theory-oriented researcher may find a search for research with his descriptors in fields such as medicine or animal behavior useful. The grounded theory researcher would not seek related literature in any of the human behavioral sciences.

When the grounded theory study is complete, the researcher formulates theories to explain what has been observed. Then the researcher searches the literature to determine how his or her conclusions fit into the existing theories in the field.

Other fields of qualitative research may include a brief review of related literature at the beginning of a study to identify the theory that inspired the research or to justify the need for it. In the case of mixed methods research, the literature review may take a more dynamic and flexible form. It may be exploratory in the beginning stages of the study and explanatory at the end of the study. Or, it may take on both characteristics in iterative fashion as new research questions arise.

## EFFICIENT LOCATION OF RELATED LITERATURE

In the past, and sometimes even today, researchers had to go to their libraries and search manually for information related to their research topic. Doing a manual search requires that you first decide what key words (or **descriptors**) best fit your topic and then look at the periodicals on the library shelves to find those most likely to include your area of interest. Using your key terms, seek your topic through the periodicals' indexes to locate relevant articles. Make sure that you note at this time all relevant bibliographic details—author, title, journal name, data, volume number, and pages.

However, manual searching of this kind is time-consuming and inefficient. Currently, most universities and colleges and many public and private libraries subscribe to indexing and abstracting periodicals that are incorporated into several databases that can be searched by computer.

Computers can search for many topics simultaneously and combine them, using logical concepts known as **Boolean logic** (from the logic system developed by the 19th-century English mathematician George Boole). The essence of Boolean logic lies in the use of three terms—*AND, OR,* and *NOT*—which are used to tell the computer how the investigator wants it to deal with the various constructs relevant to the study. For example, if you are interested in the outcome of different methods of teaching spelling to early primary students, you would start by finding all appropriate descriptors for spelling instruction and for primary grades. If you connect these with AND, you will only get documents that have both descriptors. If, instead, you use OR, you will get a flood of information about spelling instruction in all grades together with a second flood about primary grades not necessarily confined to the teaching of spelling. So as not to get drowned in useless information, you can narrow your search by adding descriptors to either side of your OR equation. Thus, think of OR as an inclusive operation and AND as a selective operation. The third option, NOT, is an exclusive operation. If you wished your search to be only of U.S. publications, you would add NOT foreign countries. Thus, your search string would read Spelling Instruction AND Primary Grades NOT Foreign Countries. In practice, it is likely that you will have more descriptors, and you will have to think carefully whether they should be linked with AND or OR.

The circles in Figure 4.1 show the use of Boolean logic in its simplest form, with only two constructs. The circles on the left show the use of AND, which includes both construct A and construct B. The circles in the center show the use of OR, which includes all the documents with either construct A or construct B. The circles on the right include documents with construct A but not construct B.

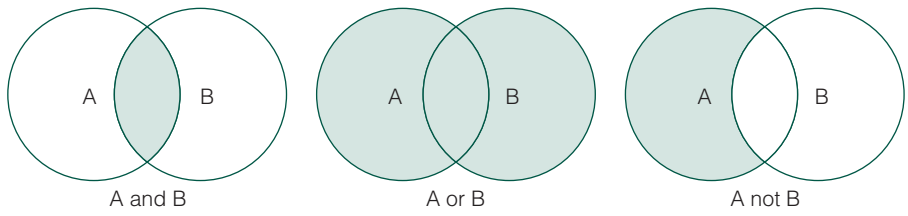



**Figure 4.1** Shaded Areas Represent What Is Retrieved by Each Logical Statement

For example, a search of the ERIC database using the Boolean descriptors (from the ERIC Thesaurus) "Changing answers" OR "Answer changing" AND "Multiple choice tests" plus a request for articles since 2004 turned up five articles dealing with the effect of changing answers in multiple choice tests, which was discussed in Chapter 1. In every study, more changes were from wrong to right, as had been the case in all the previous studies.

## INDEXING AND ABSTRACTING DATABASES

**Indexing and abstracting periodicals** are vital for locating primary sources in your field. These publications subscribe to professional journals in a given discipline. Their staff then identifies the key terms for each article, indexes them, and typically provides an abstract for each article.

Databases that combine several of these indexing and abstracting periodicals are very useful because you can ask for your key terms of interest and the database will identify the journal articles by journal, date, volume number, and pages that include your key terms.

### ERIC (Educational Resources Information Center)

There are several reasons for beginning with the **ERIC database:**

1. ERIC indexes and abstracts more education-related primary sources than any other database. It covers more than 800 journals and more than 1 million other documents.
2. It includes useful primary sources that were never published. In fact, ERIC was established in 1966 to collect, store, index, and abstract unpublished (fugitive) information. Such documents include reports submitted to the U.S. Department of Education by its contractors and grantees, reports submitted to state and local departments of education, papers presented at professional conferences, and so forth. The IDs of these documents begin with ED. Only later were professional journals added to the ERIC database. The IDs of journal articles begin with EJ. You can download the full text of ED materials. With EJ articles, only key terms (which ERIC calls descriptors) and abstracts can be downloaded.
3. It can be accessed for free from your home or office terminal at www.eric.ed.gov. The U.S. Department of Education contracts with a private contractor to maintain the ERIC system and provide its services free to the public.

The ERIC system formerly produced hard copy (print) periodicals of ED materials in *Resources in Education* and EJ documents in *Current Index to Journals in Education*. Today, it exists only in electronic form. Submissions to ERIC are now evaluated on four criteria to determine what is included and what is not:

1. Relevance of the submission to education
2. Quality of the submission (completeness, integrity, objectivity, substantive merit, and utility/importance)
3. Sponsorship by professional societies, organizations, and government agencies
4. Editorial and peer review criteria

### Using the ERIC System

As a general rule, the following steps are components of a successful ERIC search:

1. Determine the key words (descriptors) under which articles relevant to your study might be listed. These key words will typically include the population and the variables you have identified in your problem statement.
2. Since the ERIC system only recognizes descriptors from its own thesaurus, you will need to find which of your key words are used as descriptors. You may need to find synonyms for some of the key words you have listed.
3. Using ERIC descriptors and Boolean logic, perform the search and copy or save the entire reference given for any title that may be useful. This procedure simplifies the task of finding the original articles.
4. Read the abstract first. If the abstract leads you to believe the entire document would be of interest, you can download the entire document if it has an ED prefix. If it is an EJ article, first try to download the full text through a different database. If that fails, go to your library and seek out the article.
5. Search out articles in their journals at your own institution or through interlibrary loan or ask a librarian about the possibility of obtaining the articles online either through your library or through other commercial sources.

A search of the ERIC system is an important step in the quest for related literature, but the researcher cannot assume that when this step is finished, the quest is completed. Material relevant to your question may not have been included in the ERIC indexes.

Most other databases charge for their services. Fortunately, most universities, colleges, and many libraries subscribe to at least some databases and make them available to their clients and, in most cases, to anyone who wants to use the libraries' terminals on site. Check with your library to determine what databases you can access through its services. This often changes as libraries' budgets wax and wane.

## OTHER EDUCATION-FOCUSED PERIODICAL INDEXES

There are many other periodical indexes in the field of education that are useful for locating up-to-date information on research, as well as contemporary opinion. One of the standard indexes in the field is *Education Index*, which has been published regularly since 1929. This index lists articles in 478 periodicals, yearbooks, bulletins, proceedings, and monographic series. ERIC does not index 92 of these 478 periodicals. *Education Index* is the best source for locating journal articles published prior to the establishment of ERIC. A disadvantage of *Education Index* is that it does not include abstracts. H. W. Wilson, the publisher of *Education Index*, has since 1994 produced the electronic database, *Wilson Education Abstracts Full Text*. This database provides abstracts of articles in *Education Index*, as well as the full text of articles from the 150 journals considered to be the most important in the field. *H. W. Wilson Select Database*, with over 560,000 records, includes indexes and abstracts for many different professional fields, including *Social Sciences Index* and *Humanities Index*. It also includes indexes to popular

**PICTURE THIS**


Joe Rocco

magazines such as *Readers Guide to Periodical Literature*. *Education Index* is still available as a printed periodical.

*Exceptional Child Education Resources* (*ECER*) is a quarterly publication that contains abstracts of resources in special and gifted education. It uses the same thesaurus, indexing, and abstracting rules that ERIC does, and there is considerable overlap with resources indexed by ERIC and *ECER*.

The *Physical Education Index* provides a comprehensive database of all aspects of physical education.

*Child Development Abstracts and Bibliography*, published since 1927, provides an author and subject approach to the areas of infancy, clinical medicine and public health, counseling, and developmental, comparative, and experimental psychology. *Educational Administration Abstracts* provides an author and subject approach to specialized journals in the field of educational administration. *Higher Education Abstracts* is a compilation of abstracts from journals, conference proceedings, and research reports relating to college students and student services. Topics covered include counseling and housing, financial aid, and testing and measurement.

There are specialized indexes for every discipline, including business education, industrial arts, and medicine. Some are available electronically and some in print format; many are available through both. The availability of either format depends on the library and its users' needs, the quality of the electronic version of the index, cost, and other factors. By consulting the basic guides to the literature or consulting a librarian, researchers can obtain the names of the specialized indexes in other fields that they may need.

## OTHER USEFUL DATABASES

Only ERIC is more useful to educators than PsycINFO. PsycINFO is a database with nearly 2 million records covering 2000 journals, in addition to books and book chapters. The material covered in this database is related to psychology, education, business, economics, linguistics, social work, health care, and many other fields. PsycArticles is a subset of PsycINFO that offers full text for some of the journals referenced through PsycINFO. Information about these databases is available at www.psycinfo.com. The source of its information is the periodical *Psychological Abstracts*, which has been indexing and abstracting psychology and related journals since 1887. Most university libraries subscribe to the PsycINFO service.

The procedure for using PsycINFO is the same as for ERIC, but you must use PhycINFO's thesaurus.

## CITATION INDEXES

Having access to the *Social Science Citation Index* (*SSCI*) (published since 1973), the *Science Citation Index* (*SCI*) (published since 1955), and the *Arts and Humanities Citation Index* (*A&HCI*) (published since 1976) is somewhat akin to having access to a time machine, which can go forward or backward in time. If you have read a particularly useful article that was published in 1996, through subsequent indexes you can identify more recent articles that cite this article and list it in their references or you can find out which earlier writers were cited by this author. The Institute for Scientific Information (ISI) publishes *SSCI*, *SCI*, and *A&HCI* quarterly, yearly, and in 5-year compilations using the same format for each. ISI subscribes to all the important journals and somewhat important journals in each of its three general areas. For example, *SSCI* currently receives more than 2000 journals to produce its annual serial. We use *SSCI* in our description because most education literature is referenced in that index.

*SSCI* identifies which authors have been cited during the year in all areas of social science, including education, and what has been written in various areas. It also includes the necessary bibliographic information for both cited and citing authors. This information is made available by way of four indexes:

1. The *Source Index* is an alphabetical list of all authors published in the journals covered by *SSCI* during the year. Complete bibliographic information is provided for the articles published by these authors, followed by an alphabetical list of the first authors of each of the sources cited in each article. The *Source Index* is cross-referenced to secondary authors. This index identifies papers cited by the authors listed, enabling you to go backwards in time. Some authors will have several articles in the same year. If you are interested in the work of a particular author, the *Source Index* is the place to go.
2. The *Citation Index* presents an alphabetic list of the first authors of all works cited in the articles included in the *Source Index*, followed by the year, journal, volume, and first page for each of that author's cited articles. For each article, the *Citation Index* lists the names of other authors who cited that article, followed by the journal, volume, first page, and year of the article in which the citation occurred. Thus, the *Citation Index* lets you follow the work of a particular article or author forward in time. For example, in the seventh edition of this book published in 2005, we included the entire article "Toward a Prototype of Expertise in Teaching: A Descriptive Case Study" by Tracy W. Smith and David Straham (2004) in the *Journal of Teacher Education, 55*(4), 357–372. The dissertation upon which this article was based won the American Association of Colleges of Teacher Education 2001 Outstanding Dissertation Award. We anticipated that it would have considerable impact. Figure 4.2 shows the result of our search through the citation section of the *SSCI* subdivision of the Web of Science.
3. The *Permuterm Subject Index* takes every significant word and pairs it with every other significant word in each title. Each word in a title is then listed as a primary term combined with each of the other terms as co-terms. An alphabetical listing of the names of authors whose titles contain the words is provided for each paired primary term and co-term. You can then find bibliographic information for each author in the *Source Index*.
4. Bound in with the *Permuterm Subject Index*, the *Corporate Address Index* is an alphabetical listing of organizations with which authors publishing during the year are affiliated. Under each corporate entry is a list of authors with complete bibliographic information.

*SSCI* is available in both paper and online formats. Most libraries have neither, as they are fairly expensive. However, *SSCI*, *SCI*, and *AHCI* can be accessed through the electronic aggregate database Web of Science (see pages 72 and 73). Many academic libraries subscribe to Web of Science. After determining what indexes or databases you need consult your librarian on what library resources to use to access them.

### Books in Print

As you would expect, *Books in Print* indexes those books that are currently in print both by author and by title. It enables you not only to learn what books are and are not in print but also which books have new editions and which have added or dropped authors or changed publishers. Brief reviews are frequently included.

### Tests in Print and Mental Measurement Yearbook

*Tests in Print*, produced by the Buros Institute at the University of Nebraska, does the same for commercially available tests. If you find a test that interests

| Web of Science First Author | Journal Title/ Work Being Cited | Year | Vol | First page |
|---|---|---|---|---|
| SMITH TW | JOURNAL OF TEACHER EDUCATION<br>Title: Toward a prototype of expertise in teaching - A descriptive case study | 2004 | 55 | 357 |

==========

Citing Author

AU Sakamoto, A
AF Sakamoto, Atsushi

Title

TI How do in-service teachers learn from their teaching experiences?

Journal

SO JAPANESE JOURNAL OF EDUCATIONAL PSYCHOLOGY

Abstract

AB The present study reviewed research about in-service teacher learning that was published in the U.S. since 2000, as well as recent research published in Japan. In-service teacher learning was examined comprehensively from 3 perspectives: (1) learning from teaching experience, (2) school-wide contexts that support teacher learning, and (3) the longitudinal process of teacher development. Seeing teachers as "reflective practitioners" means that reflecting on teaching experiences constitutes a central point of in-service teacher learning. The present review of studies about pre-service and new teachers revealed relationships between reflection and both teaching beliefs and knowledge construction. The present review of studies on the teacher community and lesson study suggested that teachers learn from their interactions and conflicts; they also learn how collective learning occurs through lesson study as part of their school-wide professional development. The present review of research on teacher development and expertise suggested that teachers personalize their teaching, and that teachers are "adaptive experts." Further research is needed on the beliefs of in-service teachers, the development of empirical research methods for studying the development of collegiality, and teacher learning in Japan.

Citing Author

AU Oriand-Barak, L

Title

TI Lost in translation mentors learning to participate in competing discourses of practice

Journal

SO JOURNAL OF TEACHER EDUCATION

Abstract

AB Situated in the context of Israeli in-service education, this article explores the development of the author's understanding of the process of learning to mentor from the acquisition of communicative competencies (as identified in an initial study), toward a more discursive view of the process as "participation in competing discourses of practice" (as identified in subsequent studies). Recent work has revealed the intricacies and complexities entailed in translating from one language (teaching) to another (mentoring), often positioning mentors as "lost in translation." Specifically, studies shed light on issues of morality, expertise, context, and conditions for learning to mentor. To discuss the author's evolving understandings, she attends to three interrelated themes: Mentoring as connected to teaching, mentoring as distinct from teacher, and conditions for learning to mentor. Finally, the author addresses the question, So that? to consolidate her emergent understanding of the metaphor, framed as assertions for thinking about the practice of mentoring.

**Figure 4.2** Example of SSCI Citation Entry (from 2008)

you in *Tests in Print*, you are referred to the appropriate volume of *Mental Measurements Yearbook* (*MMY*) (also published by the Buros Institute) for more information on the test (test purpose, test publisher, price, test acronym, intended test population, administration times, publication date(s), test authors and in print status). Descriptions of each test are followed by critical reviews and references to studies in which the test has been used.

*Tests in Print* and *MMY* are both in hard copy and electronic form. A useful strategy is to first determine if your chosen test is in *Tests Reviews Online*, which is continuously updated. Then seek the detail in the *MMY* hard copy at your library.

If your library has neither the hard copy nor electronic *MMY*, try getting it through interlibrary loan or other sources. If you are in a hurry, *Test Reviews Online* will send you the reviews for a $15 fee. Fax services are also available.

If you cannot find the test you want in *Test Reviews Online*, the Buros Institute recommends searching the Educational Testing Service (ETS) website. ETS maintains information on more than 20,000 commercial and research instruments, although it does not provide evaluations.

## STATISTICAL SOURCES

For educational statistics, the federal government, followed by states and local governments, accounts for the greatest number of statistical documents. *Statistics Sources* (Wasserman-O'Brien, 2001), a basic **statistical sources** reference work, is an alphabetic guide to sources of information on more than 20,000 subjects, with 10,000 citations from more than 2000 sources, print and nonprint, in print and out of print, and domestic and international. It is a subject guide to data on various topics, including education, for the United States and other nations. The 2000 edition of this work also includes listings of the Federal Statistical Databases and Federal Statistical Telephone contacts. Several other commonly known reference works—such as *World Almanac and Book of Facts*, *Information Please Almanac*, and *Statistical Abstract of the United States*—contain statistics from educational fields and other subjects. In general, the data are reliable and sources for many of the statistics are also given.

Three indexes to statistics published by the Washington-based Congressional Information Service are the *American Statistics Index* (*ASI*), *Statistical Reference Index* (*SRI*), and *Index to International Statistics* (*IIS*). The *ASI* indexes and abstracts almost every statistical source issued by the federal government, and the *SRI* indexes and abstracts state documents. The latter also includes many nongovernmental statistics, ranging from those issued by private concerns and businesses to those issued by nonprofit organizations and associations. The *IIS* includes major governmental statistics from throughout the world. It is an excellent source of United Nations statistical data. These three indexes are also available in CD-ROM and online.

Since 1962, the *Digest of Education Statistics* has covered the broad field of U.S. education from kindergarten through graduate school. This publication includes a selection of data from many sources, both government and private, and draws especially on the results of surveys and activities carried out by the National Center for Education Statistics (NCES). As noted previously, the full text of the digest and other important statistical publications are searchable through the NCES website (http://nces.ed.gov).

## GOVERNMENT PUBLICATIONS

The federal government, a major source of education information, sponsors more research, conducts more surveys, and collects more statistics of all kinds than any other organization in the United States. The U.S. Department of Education disseminates a vast number of publications, including research reports, surveys, administrative actions, and program descriptions.

For locating specific U.S. government publications, the U.S. Superintendent of Documents' *Monthly Catalog of U.S. Government Publications* (or its online counterpart *GPO Monthly Catalog*, published by the U.S. Government Printing Office [GPO]) is the prime index to consult. Its main section lists documents published by each agency, and it also includes title, author, subject, and title–key word indexes. Annual compilations by title and subject were included in each December issue until 1975; since 1976, 6-month compilations have been published. There are also 5-year cumulative indexes for faster searching back through the years. The online GPO catalog has become very prevalent in academic libraries because it is easy to search. It goes back to 1976.

You can locate publications of state departments of education and other state agencies through the home pages of each state's department of education. As is true of nearly all indexes useful to scholars, most government indexes are now available on the Internet or on CD-ROM. The gateways for government information and publications on education are the U.S. Department of Education website at www.ed.gov, and the National Center for Educational Statistics website at http://nces.ed.gov. From these sites, you can find links to reports on current research, policy papers, and the searchable text of past and current popular paper resources, such as the *Digest of Education Statistics*, *Projections of Education Statistics*, and the *Condition of Education*. The purpose of the *Digest* is to provide a compilation of statistical information covering the broad field of U.S. education from kindergarten through graduate school. The 1998 *Digest* includes 428 tables from many different data sources. *Projections of Education Statistics to 2009* provides projections for key education statistics. It includes statistics on enrollment, graduates, classroom teachers, and expenditures in elementary and secondary schools, and institutions of higher education. The tables, figures, and text contain data on enrollment, teachers, graduates, and expenditures for the past 14 years and projections to the year 2009. The *Condition of Education* describes the current status and recent progress of education in the United States. The 1999 compendium features an overview essay and 60 indicators in five major substantive areas of education.

## PROQUEST DIGITAL DISSERTATIONS

This database includes abstracts of doctoral dissertations and master's theses dating back to 1861. A wealth of information from more than 1 million dissertations and theses completed at more than 1000 accredited colleges and universities worldwide can be found through this source. It can be searched by key words, titles, and authors. Its hard copy form is *Dissertation Abstracts*.

## AGGREGATE DATABASES

These are large databases that combine single-topic databases. They are convenient to use and often provide access to databases to which your library websites do not subscribe.

### Professional Development Collection

Designed for professional educators, this database provides a specialized collection of more than 550 important education journals; it also includes more than 200 educational reports. (Available at www.ebsco.com.)

### Academic Search Premier

Academic Search Premier, the world's largest academic multidisciplinary database, provides full-text articles of more than 4700 peer-reviewed scholarly publications and indexes and more than 9000 peer-reviewed scholarly publications. This database covers academic areas such as the social sciences, education, the humanities, language and linguistics, computer sciences, arts and literature, and engineering since 1975. (Available at www.epnet.com /academic/acasearchprem.asp.)

### Web of Knowledge

This database has included the Science and Social Sciences Citation Indexes since 1987. It also includes many other useful indexes and abstracting databases, including MEDLINE.

### JSTOR

This not-for-profit organization's search and store database can be accessed for a fee by way of a JSTOR username and password or through a participating institution. Since 1995, it has been building a high-quality interdisciplinary archive of scholarship in the humanities, social sciences, and sciences. More than 1000 academic journals and other materials are included in its continuously expanding collection, all of which are full-text searchable.

### Google Scholar

Available at http://scholar.google.com as of November 2004, Google Scholar is an online search engine that targets scholarly materials, such as peer-reviewed publications, book chapters, and conference proceedings, across many disciplines. With this database, you can identify relevant material throughout the world of scholarly research. Search results in Google Scholar are ordered by relevance, so when a search is conducted, those links listed first should be most relevant to your search. The relevance ranking is determined by the text in the article, the article's author, the publication, and the frequency of the citation of the article in the scholarly literature.

### WorldCat

The world's largest bibliographic database, **WorldCat**, is invaluable to researchers in nearly any field. This online database indexes more than 35 million books, journals, music scores, video recordings, films, newspapers, maps, dissertations, manuscripts, and more in 400 languages. The index includes all the holdings of most academic libraries of size, larger public libraries and specialized collections, and the national libraries of France, Great Britain, Canada, and the U.S. Library of Congress, to name a few. Although no abstracts are given, the libraries that hold each item are listed as part of the item records; this can be helpful information for either the researcher or the interlibrary loan staff to use in obtaining an item. Table 4.1 provides a summary of various databases with their respective web addresses.

**Table 4.1** Major Useful Databases for Educational Research

| Databases | Information Available At |
|---|---|
| Primary databases | |
| ERIC | www.eric.ed.gov |
| PsycINFO and PsycArticles | www.apa.org/psycinfo |
| Aggregate databases | |
| Academic SearchPremier | www.ebscohost.com/thisMarket.php?marketID=1 |
| Professional Development Collection | www.ebscohost.com/thisTopic.php?topicID=123&marketID=1 |
| Web of Knowledge | www.isiwebofknowledge.com |
| Google Scholar | http://scholar.google.com |
| WorldCat | www.worldcat.org |
| JSTOR | www.jstor.org |

## THE NECESSITY OF MASTERING ONLINE DATABASE SEARCHING

It is essential to learn how to do online searching for several reasons:

1. Many important indexing and abstracting periodicals no longer exist in hard copy. For example, ERIC no longer prints *Resources in Education* and *Current Index to Journals in Education*. The ERIC database, which includes what was once in the periodicals, now exists only in electronic form. Many libraries no longer subscribe to the hard copy versions of indexing and abstracting periodicals and subscribe only to their electronic equivalents.
2. In a manual search, you must examine numerous periodical indexes for a particular topic, follow the topic through the indexes, and find a few relevant entries that combine your selected term with another interest. Computers can search for many topics at the same time and combine them using Boolean logic. Electronic database searches yield results that are more comprehensive, more precise, more accurate, and more organized than the results of manual searches.
3. Computer searching saves time because in only seconds the computer can retrieve and print information, eliminating hours or even days for a manual search.
4. Online databases are frequently updated weekly or biweekly and thus are usually more up-to-date than printed indexes.
5. With computer searching, you can narrow your search by specifying parameters such as date and type of publication, language of publication, publication date (descending or ascending order), alphabetically by author, or by title.

It does take time and effort to master the use of **electronic databases**, but in academia today it must be done. The different resources available at various libraries create complications, as do changes in resources within an individual library. You will probably need the assistance of librarians and colleagues to learn how to search electronic databases. Sometimes, you may elect to contact one website and find out what it would cost to do the search you need.

## THE INTERNET

The vast majority of college and university students have experience using the Internet. Since the Internet has been available to consumers, the growth in Internet use, and in the variety of uses, both legitimate and suspect, has been unprecedented. As reviewed in this chapter, there are a wide range of Internet sources for scholarly literature and test sources, as well as a magnitude of other types of information. The Internet can help researchers access different databases of various types of materials and subjects throughout the world.

The strength of the Internet, however—namely its size and comprehensiveness—can work against the beginning researcher and may confound even the most experienced researcher. The old problem of not finding enough information has been replaced by the equally vexing problem of finding far too much information and needing to be able to determine the relative quality of the information. Because there is no real "publication cycle" for Internet materials, there are no editors, fact checkers, or proofreaders, and there is no accurate way for the reader to determine the authority of the author, as there is in traditional publishing. A huge amount of information exists on the Internet, some erroneous, outdated, slanted, and even harmful. The websites of hate groups, Holocaust revisionists, conspiracy theorists, and other biased authors are the more extreme examples, but one can also find such things as a free online *Webster's Dictionary*—copyright 1913!

It is often more difficult to determine the worth of a website than that of a print source because many personal sites look as professional and authoritative as a governmental or educational site. One place to start is to consider the end of the address. Sites ending in *.edu* or *.gov* are education or government sites, which tend to have more credibility than sites ending in .com, .org, or .net. Many libraries and organizations provide lists of subject-specific websites for researchers.

### EVALUATING INTERNET SOURCES

Because the Internet is a means by which a great deal of information of varying quality can be presented—the worthy and the suspect side by side—it is important that the researcher has criteria by which to judge Internet sources.

Determining the quality of a given print journal's articles has traditionally focused on the following criteria:

Reputation of the journal

Stringency of its editorial policies

Reporting of primary research, as opposed to feature articles synthesizing or summarizing bodies of research for the lay audience

Use of blind reviews

Reputation of its reviewers

Journal's affiliation with distinguished learned societies

The presence or absence of advertising

Audience for which the journal is intended, scholarly versus the lay audience

All these criteria have been used as ways to determine the relative worth of a particular journal source. These criteria help a novice researcher determine whether the article he or she is looking at is likely to be of high quality. However,

when accessing research from web versions of formerly print-based journals, journals that exist solely in electronic form, or accessing the full-text of articles from a variety of sources using a database such as ERIC as a gateway to full-text content, these criteria become less easily applied, even irrelevant. A web-based magazine for the general reader may look as professional and authoritative as a governmental or educational site, and the standards or editorial policies used in the choice of articles to include can be difficult to determine.

### Authority

Is the author of the resource identified? Is a "snail mail" address or e-mail address given so that you can contact him or her? Is the author affiliated with a university, professional organization, or community organization? Is there a link to the sponsoring group's home page, a description of the group, or contact information? For information found on the web, what is the site extension? Web addresses end in a suffix such as ".edu" or ".com." This suffix gives the researcher an idea of who is hosting the website, as follows:

| | |
|---|---|
| **.com** | A commercial site |
| **.edu** | A college or university |
| **.gov** | The U.S. government |
| **.us** | Usually a state government, community college, or school district site |
| **.org** | An organization |
| **.net** | A community network or Internet service provider |

Although you cannot be assured of the quality of information on any kind of site, educational and governmental sites usually contain authoritative information. Information provided by companies may be slanted to sell their products. In fact, it may be difficult to distinguish between an advertisement and a source of objective information.

### Accuracy

Is the resource part of an edited publication; has it been selected by a reviewer or editor? Are factual statements supported with bibliographies, endnotes, or links to the sources used? Are the sources cited, with complete bibliographic information including the date?

### Timeliness

Is it clear when the information was originally published? For web-based information, is there a date listed when the page was last updated? If there are links given to outside web pages, are the links still active or are they linking to sites that have moved or changed addresses?

## ONLINE JOURNALS

In recent years, many journals have been produced in a strictly online format and are frequently referred to as e-journals. That is, these journals are not available in print but are only available on the Internet. In education, the premier research association is the American Educational Research Association (AERA).

Within AERA, subgroups address particular interests in educational research. These are called divisions and special-interest groups (SIGs). One such SIG is the Communication of Research SIG. This SIG maintains a website listing many of the online journals in the field of education. The SIG and the listing of e-journals are available at http://aera-cr.asu.edu/index.html. Through the website, the content of nearly 200 e-journals can be accessed.

## ORGANIZING THE RELATED LITERATURE

Once you are satisfied that you have carried out a reasonably comprehensive search of the literature in the field, you can proceed to the task of organizing it. A useful approach is to arrange the studies by topic and determine how each of these topics relates to your own study.

Avoid the temptation to present the literature as a series of abstracts. Rather, use it to lay a systematic foundation for the study. Present the literature in such a way as to justify carrying out your study by showing what is known and what remains to be investigated in the topic of concern. The hypotheses provide a framework for such organization. Like an explorer proposing an expedition, the researcher maps out the known territory and points the way to the unknown territory to be explored. If your study includes several facets or investigates more than a single hypothesis, the organization process is done separately for each hypothesis.

It is almost inevitable that a number of the reports you have carefully studied and included in your notes will, on reflection, prove only peripherally related to the topic. It is neither necessary nor desirable to include in a proposal every study encountered in the search through the literature. Your readers will not be impressed by mere quantity. Relevance and organization of the material are of prime importance.

The researcher who fails to approach the task of assembling the related literature in a systematic manner from the beginning can become very disorganized. The following suggestions may be of assistance. Your university, department, or research adviser may offer help sessions or minicourses, and the librarians at most institutions can also provide helpful suggestions.

1. *Begin reading the most recent studies in the field and then work backward through earlier volumes*. An obvious advantage of this approach is that you start with studies that have already incorporated the thoughts and findings of previous research. Earlier misunderstandings have been corrected, and unprofitable approaches have been identified. Another advantage is that these studies include references to earlier works and therefore direct you to sources you might not otherwise encounter. Obviously, limits must be set to the process of gathering related research. On the one hand, laying meaningful groundwork for a study entails including all the important works in the field. On the other hand, devoting excessive time to this endeavor could result in boring the readers of your own report with superfluous detail. Make sure the related literature serves, but does not dominate, your own work.
2. *Read the abstract or summary sections of a report first to determine whether it is relevant to the question*. Doing so can save much time that might be wasted reading unhelpful articles.

3. *Before taking notes, skim the report quickly to find those sections that are related to the question—another way to save reading time.*
4. *Make notes on file cards, in a word processing program, or in some format that can be accessed easily or moved around to cluster with other notes on related research.* This begins to organize the review. With the prevalence of spreadsheet and database programs such as EndNote and ProCite on virtually every computer and stand-alone software packages such as RefWorks also available, note taking and manipulation of data is significantly easier than it was in earlier versions of this text. EndNote, ProCite, and RefWorks all offer a free download trial, which will help you get started.
5. *Write out a separate complete bibliographic reference for each work. For the sake of record keeping, include the reference in the bibliography list and with the individual note card with the notes on the source.* A bibliography typically includes author, title, publisher, year, issue and volume numbers, and/or the universal resource locator (URL) or web address, the date you accessed an electronic source, and other information depending on the type of sources. Follow the most recent edition of the chosen style manual for citing references. There are websites that provide help in using the American Psychological Association and Turabian style manuals (see Chapter 19), which may be good places to begin. Add the library call number, location of the terminal, or URL of the source to facilitate finding the work again, should it be necessary.
6. *To facilitate sorting and organizing, do not put more than one reference on each page, entry, or card.* It is not possible to arrange references alphabetically or in any other way unless they are recorded singly.
7. *Be sure to indicate which parts of the notes are direct quotations from the author and which are your own paraphrases.* Failure to make this distinction can lead to inadvertent plagiarism. It is also wise to clearly separate the author's evaluation of his or her research from your own conclusions.
8. *If you searched online databases, keep the search strategies (often called "search histories") on file.* Typically, any given database will give the option of printing out a search history, the list of searches typed in, along with the results. This information will help in the retrieval of information and reduce cost and time in case an update is needed.

## SUMMARY

A thorough knowledge of related literature is important for both quantitative and qualitative studies. Quantitative researchers need to become familiar with previous research in their field of study so that their work is based in, and expands on, research that has previously been done in that area. For qualitative researchers, the study of related literature usually follows their research, when the researcher looks at the related literature to determine how his or her research fits in with what is already known.

Sources of related literature can be found by manual searching, but this is slow and tedious. Now that electronic databases hold most of the information that researchers need, you must develop electronic searching skills, using

Boolean logic. You must also learn what resources your local library has and make use of the assistance that your librarian can give you.

There are many indexing and abstracting periodicals, covering many fields. The ERIC system and PsycINFO are probably the most useful to educators, but you need to be aware of the many other resources available to you. Citation indexes should be consulted since they enable researchers to follow topics of interest both backward and forward in time. *Books in Print* and *Tests in Print* are also important sources. There are other publications covering statistical information and government publications. Aggregate databases combine single-subject databases, allowing a single search to cover many different databases.

The Internet can be a source of information for the researcher, but it must be used with caution since there is no "quality control" comparable to the peer reviews that take place before papers are published in professional journals.

In order to be helpful to the researcher and interesting for the reader of a paper or dissertation, the related literature must be organized. Studies that are most relevant to the topic of the research reported on should be discussed first, giving as much detail as is necessary to confirm their relevance. Other studies that are peripheral to the research should be mentioned briefly, especially if they suggest additional studies that might be undertaken later.

## KEY CONCEPTS

aggregate databases
Boolean logic
citation indexes
descriptors
electronic databases
ERIC database
function of related literature
government publications
indexing and abstracting periodicals
related literature
statistical sources
WorldCat

## EXERCISES

1. Which of the following are recommended strategies for organizing a literature search?
   - **a.** Organize studies by topic.
   - **b.** Begin with early articles and work forward in time.
   - **c.** Read the abstract or summary sections of a report first.
   - **d.** Skim the report to find those sections that are related to the chosen questions.
   - **e.** Write out a complete bibliographic reference for each work.
2. According to the text, what are three important roles of incorporating related literature in a research project? In your opinion, which role seems most important to you in your current or future research?
3. Conduct a search on the same general topic in both ERIC and *Education Index*. Compare the usefulness of ERIC and *Education Index* in finding related research on your topic. Compare and contrast two of the following: the quality of abstracts, the journals covered, and the subject terms or descriptors used by the index. Which index do you predict would be the most useful in finding research on your topic?
4. Explain the organization and the purpose of the *Mental Measurements Yearbook*.
5. What purposes might searching the *Social Science Citation Index* serve? In addition to information about an article's citation record, what conclusions might the *SSCI* help you draw about an author or about a journal?
6. What conclusion could be drawn regarding a work, published in 1966, that was cited in 20 articles listed in the 2009 Source Index of *SSCI*?
7. If you were interested in the effect of teachers' stereotyping on the achievement of girls in math classes, what are some terms for which you might search? Illustrate how you might combine them into online search statements with Boolean operators AND, OR, or NOT.
8. How do indexing and abstracting periodicals, databases, and aggregate databases differ?

9. Find two web resources on an education-related topic of your choice, one from a commercial (.com) site and one from an educational (.edu) site. Evaluate them using the criteria of authority, accuracy, and timeliness. Are there notable differences between the sources? Explain.

## ANSWERS

1. a, c, d, and e
2. Knowledge of related research enables the researcher to define the frontiers of the field, place the question in perspective, and avoid unintentional replication of previous studies. Defining the frontiers of research is especially important for a relatively inexperienced researcher, but any well-reasoned choice is acceptable.
3. Both provide a means to locate relevant journal articles. *Education Index* covers journals from 1929 to the present. ERIC includes more journals. Begun in 1966, it indexes articles using the ERIC thesaurus of subject descriptors and provides annotations. *Education Index* does not index unpublished literature. There are some periodicals that only *Education Index* covers, which tend to be newsletters and so on. Because *Wilson Selected Databases*, the electronic version of the index, tends to have shorter abstracts and also uses a controlled vocabulary that is less specific than the ERIC descriptors, most students will consider ERIC more useful, unless a student is doing historical research.
4. The editions of *MMY* provide the most comprehensive listing and description of standardized tests available, given its long publication history and the length and authority of its reviews. The critical reviews of the tests assist one in selecting an appropriate test.
5. *SSCI* provides a way to see what subsequent research has followed a particular article. You can draw tentative conclusions about the influence of an author by looking at how often his or her work has been cited, and in turn, you can draw conclusions about the influential journals in a field by seeing how often those articles that are heavily cited seem to come from a cluster of important journal titles.
6. If an older article becomes heavily cited several years later, there may be several explanations: The research may have been treated by a noted scholar in a new way, a seminal piece of research may have recently been disproved, or the field of study may be experiencing a surge of interest because of a news event or finding.
7. Answers will vary, but terms mentioned would include *Females* (not *Girls*) AND *Teacher Attitudes* (not *Teacher Stereotypes*, which refers to stereotypes of teacher behavior and attributes) AND *Mathematics Instruction*, *Mathematics Education*, or *Mathematics*. *Sex Discrimination* or *Sex Stereotypes* might also be used successfully.
8. Indexing and abstracting periodicals subscribe to journals and sometimes other sources in a field. Their staff read and index them and usually produce abstracts that are published on a regular basis. This content is incorporated into an electronic database that makes searching much more precise, complete, and convenient. Aggregating databases combine single-topic individual databases to make searches even more convenient and through.
9. Answers will vary.

## REFERENCES

*Child development abstracts and bibliography*. (1927–). Lafayette, IN: Purdue University, Society for Research in Child Development.

Glaser, B. G. (1978). *Theoretical sensitivity*. Mill Valley, CA: Sociology Press.

Smith, T. W., & Straham, D. (2004). Toward a prototype of expertise in teaching: A descriptive study. *Journal of Teacher Education, 55*(4), 357–372.

Wasserman-O'Brien, S. R. (Ed.). (2001). *Statistics sources* (24th ed.). Detroit: Gale Research.

CHAPTER 5

# The Hypothesis in Quantitative Research

A hypothesis transforms a general idea into a plan for what to look for.

## INSTRUCTIONAL OBJECTIVES

After studying this chapter, the student will be able to:

1. Define *hypothesis*.
2. Describe the purposes of the hypothesis(es) in quantitative and qualitative research.
3. List the criteria of a theory useful for a research study.
4. Distinguish between an inductive and a deductive hypothesis.
5. State the criteria used to evaluate hypotheses for research.
6. Define *operational definition* and give an example.
7. Identify a testable hypothesis from given examples.
8. Define *null hypothesis* and explain its purpose in a research study.
9. Write a research hypothesis and a null hypothesis for a research study.
10. Distinguish between a directional and a nondirectional hypothesis.
11. Describe the steps in testing a hypothesis.
12. State the purpose of the research plan and list the elements to be included.
13. State the purpose of a pilot study.

After stating the research question and examining the literature, the quantitative researcher is ready to state a **hypothesis** based on the question.* This should be done before beginning the research project. Recall that the quantitative problem asks about the relationship between two (or more) variables. The hypothesis presents the researcher's expectations about the relationship between variables within the question. Hence, it is put forth as a suggested answer to the question, with the understanding that the ensuing investigation may lead to either support for the hypothesis or lack of support for it. Note that we use the word *support*, not *prove*. Research may find support for a hypothesis, but it does not prove a hypothesis.

*The role of the hypothesis in qualitative research is discussed in Chapter 15.

A researcher might ask the question, "What is the effect of preschool training on the first-grade achievement of culturally disadvantaged children?" The hypothesis would read "Culturally disadvantaged children who have had preschool training achieve at a higher level in first grade than culturally disadvantaged children who have not had preschool training." You can see that the hypothesis related the variables of preschool training and first-grade achievement. The following are additional examples of hypotheses in educational research:

1. Boys in elementary school achieve at a higher level in single-sex classes than in mixed classes.
2. Students who complete a unit on problem-solving strategies will score higher on a standardized mathematics test than those who have completed a control unit.
3. Middle school students who have previously taken music lessons will have higher math aptitude scores.
4. Middle school students who have siblings will be more popular among their peers than students who do not have siblings.
5. Students who do warm-up exercises before an examination will score higher on that examination than those who do not.
6. Elementary school children who do not get adequate sleep will perform at a lower level academically than will their peers who have adequate sleep.

Although hypotheses serve several important purposes, some research studies may proceed without them. Hypotheses are tools in the research process, not ends in themselves. Studies are often undertaken in areas in which there is little accumulated background information. A researcher may not know what outcome to predict. For example, surveys that seek to describe the characteristics of particular phenomena, or to ascertain the attitudes and opinions of groups, often proceed without hypotheses.

Two reasons for stating a hypothesis before the data-gathering phase of a quantitative study are (1) a well-grounded hypothesis indicates that the researcher has sufficient knowledge in the area to undertake the investigation, and (2) the hypothesis gives direction to the collection and interpretation of the data; it tells the researcher what procedure to follow and what type of data to gather and thus may prevent a great deal of wasted time and effort on the part of the researcher.

## PURPOSES OF THE HYPOTHESIS IN QUANTITATIVE RESEARCH

Principal purposes served by the hypothesis include the following:

1. *The hypothesis brings together information to enable the researcher to make a tentative statement about how the variables in the study may be related*. By integrating information based on experience, related research, and theory, the researcher states the hypothesis that provides the most satisfactory prediction or the best solution to a problem.

2. *Because hypotheses propose tentative explanations for phenomena, they stimulate a research endeavor that results in the accumulation of new knowledge*. Hypothesis testing research permits investigators to validate or fail to validate theory through an accumulation of data from many studies. In this way, knowledge is extended.
3. *The hypothesis provides the investigator with a relational statement that is directly testable in a research study*. That is, it is possible to collect and analyze data that will confirm or fail to confirm the hypothesis. Questions cannot be tested directly. An investigation begins with a question, but only the proposed relationship between the variables can be tested. For instance, you do not test the question, "Do teachers' written comments on students' papers result in an improvement in student performance?" Instead, you test the hypothesis that the question implies: "Teachers' written comments on students' papers result in a meaningful improvement in student performance" or, specifically, "The performance scores of students who have had written teacher comments on previous papers will exceed those of students who have not had written teacher comments on previous papers." You then proceed to gather data about the relationship between the two variables (teachers' written comments and student performance).
4. *The hypothesis provides direction to the research*. The hypothesis posits a specific relationship between variables and thus determines the nature of the data needed to test the proposition. Very simply, the hypothesis tells the researcher what to do. Facts must be selected and observations made because they have relevance to a particular question, and the hypothesis determines the relevance of these facts. The hypothesis provides a basis for selecting the sampling, measurement, and research procedures to use, as well as the appropriate statistical analysis. Furthermore, the hypothesis helps keep the study restricted in scope, preventing it from becoming too broad or unwieldy.

   For example, consider again the hypothesis concerning preschool experience of culturally disadvantaged children and their achievement in first grade. This hypothesis indicates the research method required and the sample, and it even directs the researcher to the statistical test that would be necessary for analyzing the data. It is clear from the statement of the hypothesis that the researcher will conduct an ex post facto study that compares the first-grade achievement of a sample of culturally disadvantaged children who went through a preschool program and a similar group of disadvantaged children who did not have preschool experience. Any difference in the mean achievement of the two groups could be analyzed for statistical significance by the $t$ test or analysis of variance technique. (We discuss these procedures in Chapter 7.)
5. *The hypothesis provides a framework for reporting the findings and conclusions of the study*. The researcher will find it very convenient to take each hypothesis separately and state the conclusions that are relevant to it; that is, the researcher can organize this section of the written report around the provision of answers to the original hypotheses, thereby making a more meaningful and readable presentation.

## SUGGESTIONS FOR DERIVING HYPOTHESES

As explained in Chapter 3, a study might originate in a practical problem, in some observed behavioral situation in need of explanation, in previous research, or even more profitably in some educational, psychological, or sociological theory. Thus, researchers derive hypotheses inductively from observations of behavior or deductively from theory or from the findings of previous research. Induction and deduction are complementary processes. In induction, one starts with specific observations and reaches general conclusions; in deduction, one begins with generalizations and makes specific predictions.

### DERIVING HYPOTHESES INDUCTIVELY

In the inductive procedure, the researcher formulates an **inductive hypothesis** as a generalization from apparent observed relationships; that is, the researcher observes behavior, notices trends or probable relationships, and then hypothesizes an explanation for this observed behavior. This reasoning process should be accompanied by an examination of previous research to determine what findings other investigators have reported on the question.

The inductive procedure is a particularly fruitful source of hypotheses for classroom teachers. Teachers observe learning and other student behavior every day and try to relate it to their own behavior, to the behavior of other students, to the teaching methods used, to changes in the school environment, and so on. Teachers might observe, for example, that when they present particularly challenging activities in the classroom, some students get motivated and really blossom, whereas others withdraw from the challenge. Some students learn complex concepts best from primarily verbal presentations (lectures), whereas others learn best from discussions and hands-on activities. After reflecting on such experiences, teachers may inductively formulate generalizations that seek to explain the observed relationship between their methods and materials and students' learning. These tentative explanations of why things happen as they do can become the hypotheses in empirical investigations.

Perhaps a teacher has observed that classroom tests arouse a high degree of anxiety and believes this adversely affects student performance. Furthermore, the teacher has noted that when students have an opportunity to write comments about objective questions, their test performance seems to improve. The teacher reasons that this freedom to make comments must somehow reduce anxiety and, as a result, the students score better. This observation suggests a hypothesis: Students who are encouraged to write comments about test items on their answer sheets will achieve higher test scores than students who have no opportunity to make comments.

The teacher could then design an experiment to test this hypothesis. Note that the hypothesis expresses the teacher's belief concerning the relationship between the two variables (writing or not writing comments about test items and performance on the test). Note also that the variable *anxiety* that was part of the reasoning chain leading to the hypothesis is not part of the final hypothesis. Therefore, the results of the investigation would provide information concerning only the relation between writing comments and test performance. The relationships between anxiety and comments, and anxiety and test performance, could

be subjects for subsequent hypotheses to investigate. Frequently, an original idea involves a series of relationships that you cannot directly observe. You then reformulate the question to focus on relationships that are amenable to direct observation and measurement.

The following are additional examples of hypotheses that might be arrived at inductively from a teacher's observations:

- Students' learning of computer programming in the middle grades increases their development of logical thinking skills.
- Using advance organizers increases high school students' learning from computer-assisted instruction in chemistry.
- Students trained to write summaries of a lecture will perform better on an immediate posttest on lecture comprehension than will students who simply take notes.
- Children score higher on final measures of first-grade reading achievement when they are taught in small groups rather than large groups.
- The cognitive and affective development of first-grade children is influenced by the amount of prior preschool experience.
- After-school tutoring programs increase the achievement of at-risk students.

In the inductive process, the researcher makes observations, thinks about the problem, turns to the literature for clues, makes additional observations, and then formulates a hypothesis that seeks to account for the observed behavior. The researcher (or teacher) then tests the hypothesis under controlled conditions to examine scientifically the assumption concerning the relationship between the specified variables.

## DERIVING HYPOTHESES DEDUCTIVELY

In contrast to hypotheses formulated as generalizations from observed relationships, some others are derived by deduction from **theory.** These hypotheses have the advantage of leading to a more general system of knowledge because the framework for incorporating them meaningfully into the body of knowledge already exists within the theory. A science cannot develop efficiently if each study results in an isolated bit of knowledge. It becomes cumulative by building on the existing body of facts and theories. A hypothesis derived from a theory is known as a **deductive hypothesis.**

After choosing a theory of interest, you use deductive reasoning to arrive at the logical consequences of the theory. If A is true, then we would expect B to follow. These deductions then become the hypotheses in the research study. For example, social comparison theory suggests that students form academic self-concepts by comparing their self-perceived academic accomplishments to some standard or frame of reference. The frame of reference for most students would be the perceived academic abilities of their classmates. If this is true, then one might hypothesize that gifted students would have lower academic self-concepts if they were placed in selective homogeneous groups than if they were in heterogeneous or mixed-ability groups in which they compare themselves to less able students.

One could investigate this hypothesis by examining the change over time in the academic self-concept of gifted students in homogeneous classes compared to that of matched gifted students placed in regular, heterogeneous classes. The evidence gathered will support, contradict, or possibly lead to a revision of social comparison theory.

Another useful theory from which an educational researcher might make deductions is Piaget's classic theory on the development of logical thinking in children. Piaget (1968) suggested that children pass through various stages in their mental development, including the stage of concrete operations, which begins at age 7 or 8 years and marks the transition from dependence on perception to an ability to use some logical operations. These operations are on a concrete level but do involve symbolic reasoning. Using this theory as a starting point, you might therefore hypothesize that the proportion of 9-year-old children who will be able to answer correctly the transitive inference problem, "Frank is taller than George; George is taller than Robert; who is the tallest?" will be greater than the proportion of 6-year-olds who are able to answer it correctly. Such research has implications for the importance of determining students' cognitive capabilities and structuring educational tasks that are compatible with their developmental level.

Piaget's cognitive theory also emphasizes that learning is a highly active process in which learners must construct knowledge. This tenet that knowledge must be constructed by learners rather than simply being ingested from teachers is the basis for much of the research on discovery-oriented and cooperative learning.

In a study designed to test a deduction from a theory, it is extremely important to check for any logical gaps between theory and hypothesis. The researcher must ask, "Does the hypothesis logically follow from the theory?" If the hypothesis does not really follow from the theory, then the researcher cannot reach valid conclusions about the adequacy of the theory. If the hypothesis is supported but was not rigorously deduced from the theory, the researcher cannot say that the findings furnish credibility to the theory. Table 5.1 shows propositions from some well-known theories and a hypothesis based on each theory.

## CHARACTERISTICS OF A USABLE HYPOTHESIS

After tentatively formulating the hypothesis, but before attempting any actual empirical testing, you must evaluate the hypothesis. The final worth of a hypothesis cannot be judged prior to empirical testing, but there are certain useful criteria for evaluating hypotheses.

### A HYPOTHESIS STATES THE EXPECTED RELATIONSHIP BETWEEN VARIABLES

A hypothesis should conjecture the relationship between two or more variables. For example, suppose you attempt to start your car and nothing happens. It would be unprofitable to state, "The car will not start and it has a wiring system," because no relationship between variables is specified, and so there is

**Table 5.1** Well-Known Theories and a Hypothesis Based on Each Theory

| Theory | Hypothesis |
|---|---|
| Achievement motivation (McClelland, 1953) People have a tendency to strive for success and to choose goal-oriented, success/failure activities. | There is a positive relationship between achievement motivation and success in school. |
| Attribution theory (Weiner, 1994) People attempt to maintain a positive self-image; people explain their success or failure in a way that preserves their self-image. | If students are given a task and told that they failed or succeeded (even though all actually succeed), those who are told they failed say it is due to bad luck; those who are told they are successful will attribute it to skill and intelligence. |
| Theory of multiple intelligences (Gardner, 1993) People have a number of separate intelligences that may vary in strength. | Teaching science concepts using a variety of approaches will result in greater achievement than when using only linguistic and mathematical approaches. |
| Cognitive dissonance theory (Festinger, 1957) People experience discomfort when a new behavior clashes with a long-held belief or with their self-image. To resolve the discomfort, they may change their beliefs or behavior. | Requiring middle school students who smoke to write an essay on why young people should not smoke will change their attitudes about smoking. |
| Vygotsky's theory of learning (1978) Cognitive development is strongly linked to input from other people. | Tutoring by more able peers will have a positive effect on the learning of at-risk students. |
| Maslow's human needs theory (1954). In a hierarchy of needs, people must satisfy their lower level needs (hunger or safety) before they are motivated to satisfy higher level needs (self-esteem or need to know). | Children from economically disadvantaged homes who are given breakfast at school will show higher achievement than similar students not given breakfast. |
| Behaviorism (Skinner, 1953) Behavior that is positively reinforced will increase in strength. | On-task behavior will increase when teachers positively reinforce it. |

no proposed relationship to test. A fruitful hypothesis would be "The car will not start because of a fault in the wiring system." This criterion may seem patently obvious, but consider the following statement: "If children differ from one another in self-concept, they will differ from one another in social studies achievement." The statement appears to be a hypothesis until you note that there is no statement of an expected relationship. An expected relationship could be described as "Higher self-concept is a likely antecedent to higher social studies achievement." This hypothesis would then be stated as "There will be a positive relationship between self-concept and social studies achievement." If the opposite is predicted—that is, higher self-concept leads to lower social studies achievement—then the hypothesis would be "There will be a negative relationship between self-concept and social studies achievement." Either statement would meet this first criterion.

## A HYPOTHESIS MUST BE TESTABLE

The most important characteristic of a "good" hypothesis is testability. A **testable hypothesis** is verifiable; that is, deductions, conclusions, or inferences can be drawn from the hypothesis in such a way that empirical observations either support or do not support the hypothesis. If the hypothesis is on target, then

certain predictable results should be manifest. A testable hypothesis enables the researcher to determine by observation and data collection whether consequences that are deductively implied actually occur. Otherwise, it would be impossible either to confirm or not to confirm the hypothesis. In the preceding example, the hypothesis "The car's failure to start is a punishment for my sins" is obviously untestable in this world.

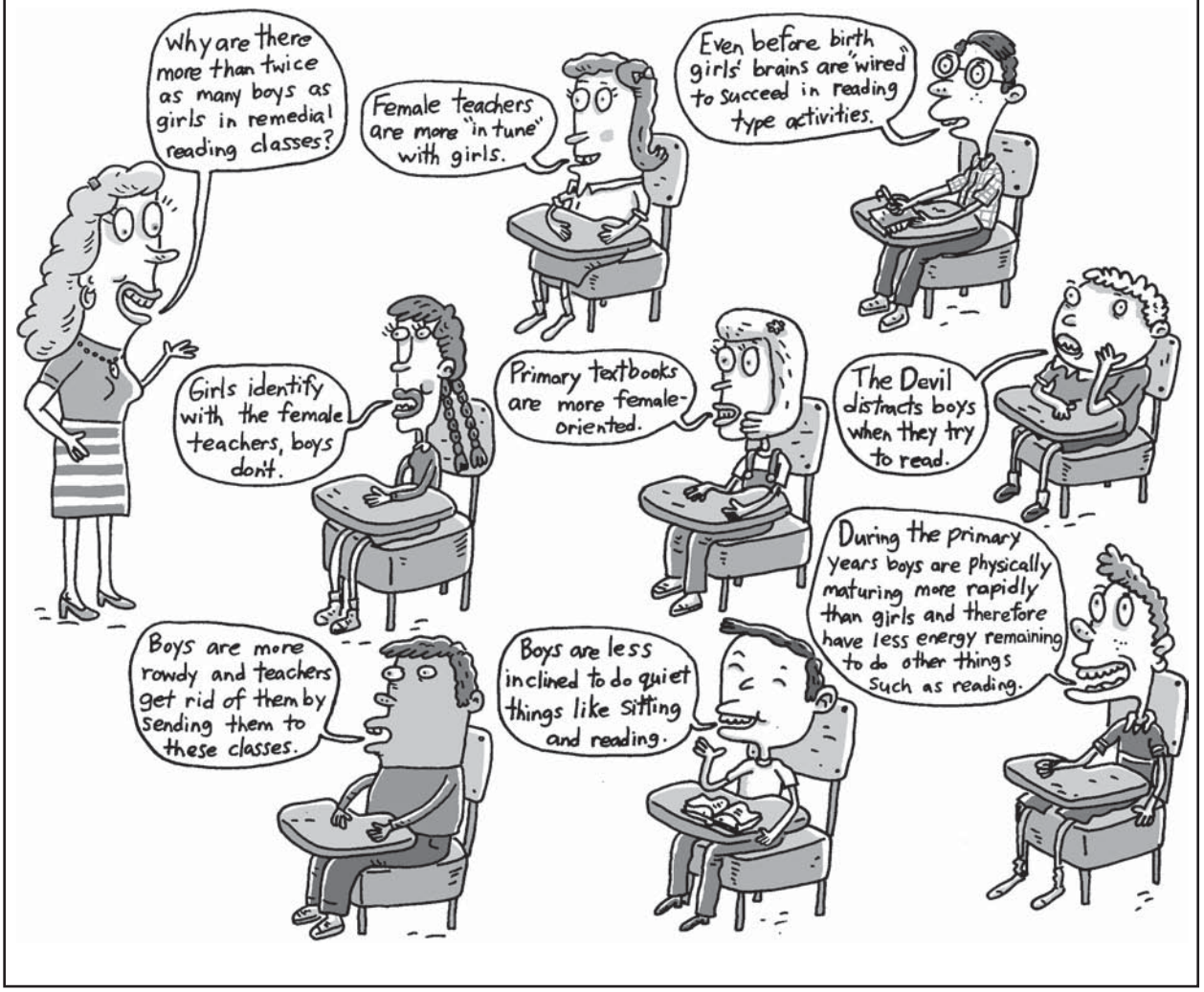
Many hypotheses—or propositions, as they may initially be stated—are essentially untestable. For instance, the hypothesis "Preschool experience promotes the all-around adjustment of the preschool child" would be difficult to test because of the difficulty of operationalizing and measuring "all-around adjustment." To be testable, a hypothesis must relate variables that can be measured. If no means are available for measuring the variables, then no one could gather the data necessary to test the validity of the hypothesis. We cannot emphasize this point too strongly. Unless you can specifically define the indicators of each variable and subsequently can measure these variables, you cannot test the hypothesis.

The indicators of the variables are referred to as **operational definitions.** Recall from Chapter 2 that variables are operationally defined by specifying the steps the investigator takes to measure the variable. Consider the hypothesis "High-stressed nursing students will perform less well on a nursing test than will low-stressed students." The operational definition of stress is as follows: One group of students is told that their performance on the nursing test will be a major determinant of whether they will remain in the nursing program (high stress), and the other group is told that they need to do as well as they can but that their scores will not be reported to the faculty or have any influence on their grades (low stress). The operational definition of test performance would be scores from a rating scale that assessed how well the students did on the various tasks making up the test. Or consider the following hypothesis: "There is a positive relationship between a child's self-esteem and his or her reading achievement in first grade." For this hypothesis to be testable, you must define the variables operationally. You might define *self-esteem* as the scores obtained on the Self-Image Profile for Children (Butler, 2001) and reading achievement as scores on the California Reading Test, or as first-grade teachers' ratings of reading achievement.

Make sure the variables can be given operational definitions. Avoid the use of constructs for which it would be difficult or impossible to find adequate measures. Constructs such as *creativity*, *authoritarianism*, and *democracy* have acquired such diverse meanings that reaching agreement on operational definitions of such concepts would be difficult, if not impossible. Remember that the variables must be defined in terms of identifiable and observable behavior.

It is important to avoid value statements in hypotheses. The statement "A counseling program in the elementary school is desirable" cannot be investigated in an empirical study because "desirable" is too vague to be measured. However, you could test the hypothesis "Elementary pupils who have had counseling will have higher scores on a measure of expressed satisfaction with school than will those who have not had counseling." You can measure verbal expressions of satisfaction, but whether they are desirable is a value judgment.

**PICTURE THIS**




Joe Rocco

**Think About It 5.1**

Which of the explanations in the cartoon are not testable hypotheses about why there are more boys than girls in remedial reading classes?

**Answer**

The one about the "wiring" in the brain and the one about the devil's activities are not testable.

## A HYPOTHESIS SHOULD BE CONSISTENT WITH THE EXISTING BODY OF KNOWLEDGE

Hypotheses should not contradict previously well-established knowledge. The hypothesis "My car will not start because the fluid in the battery has changed to gold" satisfies the first two criteria but is so contrary to what is known about the

nature of matter that you would not pursue it. The hypothesis "The car will not start because the fluid in the battery has evaporated to a low level" is consistent with previous knowledge and therefore is worth pursuing. It would probably be unprofitable to hypothesize an *absence* of relationship between the self-concept of adolescent boys and girls and their rate of physical growth because the preponderance of evidence supports the *presence* of such a relationship. Historians of science find that people such as Einstein, Newton, Darwin, and Copernicus developed truly revolutionary hypotheses that conflicted with what was accepted knowledge in their time. However, remember that the work of such pioneers was not really so much a denial of previous knowledge as a reorganization of existing knowledge into more satisfactory theory. In most cases, and especially for the beginning researcher, it is safe to suggest that the hypothesis should agree with knowledge already well established in the field. Again, this highlights the necessity for a thorough review of the literature so that hypotheses are formulated on the basis of previously reported research in the area.

## A HYPOTHESIS SHOULD BE STATED AS SIMPLY AND CONCISELY AS POSSIBLE

A hypothesis should be presented in the form of a concise declarative statement. A complete and concisely stated hypothesis makes clear what the researcher needs to do to test it. It also provides the framework for presenting the findings of the study. If a researcher is exploring more than one relationship, he or she will need to state more than one hypothesis. The general rule is to state only one relationship in any one hypothesis. For example, if you were investigating the effect of a new teaching method on student achievement and student satisfaction, you would state two hypotheses—one for effect on achievement and one for effect on satisfaction. You need not worry about the verbal redundancy inevitable in stating multiple hypotheses. Remember that the goals of testability and clarity will be served better by more specific hypotheses.

### Think About It 5.2

Which of the explanations used to explain the greater number of boys in remedial reading in the previous cartoon is not consistent with the existing body of knowledge?

**Answer**

The one that posits that in the primary grades boys mature more rapidly than girls. There is overwhelming evidence that at that stage girls mature more rapidly than boys. Boys finally catch up at approximately age 17 years.

The terms used in the hypothesis should be the simplest acceptable for conveying the intended meaning; avoid ambiguous or vague constructs. Use terms in the way that is generally accepted for referring to the phenomenon. When two hypotheses are of equal explanatory power, prefer the simpler one because it will provide the necessary explanation with fewer assumptions and variables to be defined. Remember that this principle of parsimony is important in evaluating hypotheses.

# TYPES OF HYPOTHESES

There are three categories of hypotheses: research, null, and alternate.

## THE RESEARCH HYPOTHESIS

The hypotheses we have discussed thus far are called **research hypotheses.** They are the hypotheses developed from observation, the related literature, and/or the theory described in the study. A research hypothesis states the relationship one expects to find as a result of the research. It may be a statement about the expected relationship or the expected *difference* between the variables in the study. A hypothesis about children's IQs and anxiety in the classroom could be stated "There is a positive relationship between IQ and anxiety in elementary schoolchildren" or "Children classified as having high IQs will exhibit more anxiety in the classroom than children classified as having low IQs." Research hypotheses may be stated in a **directional** or **nondirectional** form. A directional hypothesis states the direction of the predicted relationship or difference between the variables. The preceding two hypotheses about IQ and anxiety are directional. A directional hypothesis is stated when one has some basis for predicting a change in the stated direction. A nondirectional hypothesis, in contrast, states that a relationship or difference exists but without specifying the direction or nature of the expected finding—for example, "There is a relationship between IQ and anxiety in children." The literature review generally provides the basis for stating a research hypothesis as directional or nondirectional.

## THE NULL HYPOTHESIS

It is impossible to test research hypotheses directly. You must first state a **null hypothesis** (symbolized $H_0$) and assess the probability that this null hypothesis is true. The null hypothesis is a statistical hypothesis. It is called the null hypothesis because it states that there is no relationship between the variables in the population. A null hypothesis states a negation (not the reverse) of what the experimenter expects or predicts. A researcher may hope to show that after an experimental treatment, two populations will have different means, but the null hypothesis would state that after the treatment the populations' means will *not* be different.

What is the point of the null hypothesis? A null hypothesis lets researchers assess whether apparent relationships are genuine or are likely to be a function of chance alone. It states, "The results of this study could easily have happened by chance." Statistical tests are used to determine the probability that the null hypothesis is true. If the tests indicate that observed relationships had only a slight probability of occurring by chance, the null hypothesis becomes an unlikely explanation and the researcher rejects it. Researchers aim to reject the null hypothesis as they try to show there *is* a relationship between the variables of the study. Testing a null hypothesis is analogous to the prosecutor's work in a criminal trial. To establish guilt, the prosecutor (in the U.S. legal system) must provide sufficient evidence to enable a jury to reject the presumption of innocence beyond reasonable doubt. It is not possible for a prosecutor to prove guilt conclusively, nor can a researcher obtain unequivocal support for a research hypothesis. The defendant is presumed innocent until sufficient evidence indicates that he or she is not, and the null hypothesis is presumed true until sufficient evidence indicates otherwise.

For example, you might start with the expectation that children will exhibit greater mastery of mathematical concepts through individual instruction than through group instruction. In other words, you are positing a relationship between the independent variable (method of instruction) and the dependent variable (mastery of mathematical concepts). The research hypothesis is "Students taught through individual instruction will exhibit greater mastery of mathematical concepts than students taught through group instruction." The null hypothesis, the statement of no relationship between variables, will read "The mean mastery scores (population mean $\mu_i$) of all students taught by individual instruction will equal the mean mastery scores (population mean $\mu_g$) of all those taught by group instruction." $H_0$: $\mu_i = \mu_g$.*

### THE ALTERNATIVE HYPOTHESIS

Note that the hypothesis "Children taught by individual instruction will exhibit less mastery of mathematical concepts than those taught by group instruction" posits a relationship between variables and therefore is *not* a null hypothesis. It is an example of an **alternative hypothesis.**

In the example, if the sample mean of the measure of mastery of mathematical concepts is higher for the individual instruction students than for the group instruction students, and inferential statistics indicate that the null hypothesis is unlikely to be true, you reject the null hypothesis and tentatively conclude that individual instruction results in greater mastery of mathematical concepts than does group instruction. If, in contrast, the mean for the group instruction students is higher than the mean for the individual instruction students, and inferential statistics indicate that this difference is not likely to be a function of chance, then you tentatively conclude that group instruction is superior.

If inferential statistics indicate that observed differences between the means of the two instructional groups could easily be a function of chance, the null hypothesis is retained, and you decide that insufficient evidence exists for concluding there is a relationship between the dependent and independent variables. The retention of a null hypothesis is *not* positive evidence that the null hypothesis is true. It indicates that the evidence is insufficient and that the null hypothesis, the research hypothesis, and the alternative hypothesis are all possible.

## TESTING THE HYPOTHESIS

A quantitative study begins with a research hypothesis, which should be a simple, clear statement of the expected relationship between the variables. Previously, we explained that hypotheses must be testable—that is, amenable to empirical verification. When researchers speak of testing a hypothesis, however, they are referring to the null hypothesis. Only the null hypothesis can be directly tested by statistical procedures. **Hypothesis testing** involves the following steps:

1. State, in operational terms, the relationships that should be observed if the research hypothesis is true.
2. State the null hypothesis.

*The Greek letter mu, μ, is used to symbolize population mean.

3. Select a research method that will enable the hypothesized relationship to be observed if it is there.
4. Gather the empirical data and select and calculate appropriate descriptive statistics for these data (see Chapter 6).
5. Calculate inferential statistics to determine the probability that your obtained results could have occurred by chance when the null hypothesis is true (see Chapter 7).
6. If the probability of the observed findings being due to chance is very small (e.g., only 1 in 100 chances), one would have sufficient evidence to reject the null hypothesis.

Many hypotheses that are formulated are rejected after empirical testing. Their predictions are not supported by the data. Many beginning researchers believe that if the data they collect do not support their hypothesis, then their study is a failure. This is not the case. In the history of scientific research, hypotheses that failed to be supported have greatly outnumbered those that have been supported. Experienced researchers realize that unconfirmed hypotheses are an expected and useful part of the scientific experience. They can lead to reconsideration or revision of theory and the generation of new hypotheses, which often brings science closer to a correct explanation of the state of affairs. Darwin (1887/2007) wrote,

> I have steadily endeavored to keep my mind free so as to give up any hypothesis, however much beloved (and I cannot resist forming one on every subject), as soon as facts are shown to be opposed to it. Indeed, I have had no choice but to act in this manner, for with the exception of the Coral Reefs, I cannot remember a single first-formed hypothesis which had not after a time to be given up or greatly modified. (p. 293)

Although you may find support for a hypothesis, the hypothesis is not *proved* to be true. A hypothesis is never proved or disproved; it is only supported or not supported. Hypotheses are essentially probabilistic in nature; empirical evidence can lead you to conclude that the explanation is probably true or that it is reasonable to accept the hypothesis, but it never proves the hypothesis.

## CLASSROOM EXAMPLE OF TESTING A HYPOTHESIS

A teacher is interested in investigating reinforcement theory in the classroom. From her understanding of reinforcement theory, this teacher hypothesizes that teachers' positive comments on students' papers will lead to greater achievement.

Step 1. The deduced implication is stated as follows: "Teachers' positive comments on students' papers during a specified unit will result in higher scores on the end-of-unit test for those students, compared with students who received no comments." It is the relationship between the two variables—teachers' positive comments and pupil performance on the end-of-unit test—that will be investigated.

Step 2. For statistical testing, the research hypothesis must be transformed into a null hypothesis: "The population mean achievement score for students receiving positive comments (experimental group) will be the same as the population mean achievement score for students receiving no comments (control group)."

Step 3. After getting permission from parents or guardians for the children to participate, the teacher would select students to be randomly assigned to the experimental and control groups. For those students in the experimental group, she would write positive comments on their papers, whereas the students assigned to the control group would receive no comments. The comments to the experimental group should simply be words of encouragement, such as "Excellent," "Keep up the good work," or "You're doing better." These comments should have nothing to do with content or the correction of particular errors; otherwise, any improvement could be attributed to the instructional usefulness of such comments.

Step 4. After completing the specified unit, the teacher would administer a common end-of-unit test to both groups and derive average (mean) achievement scores on the test for each group.

Step 5. Inferential statistics can then be used to indicate whether any difference in mean achievement scores is real or is likely to be merely a function of chance. If the difference is not likely to be a function of chance, the researcher tentatively concludes that it results from the different treatments given to the two groups.

## THE QUANTITATIVE RESEARCH PLAN

After identifying a worthwhile problem and stating the expected outcome in the form of a research hypothesis, you are ready to develop a tentative **research plan.** The research plan at this stage is only a preliminary proposal; many changes will probably be needed before the final, formal proposal is written. Developing this tentative research plan is essential because it forces you to set down ideas in a concrete form. Many initial ideas seem promising until you must spell them out in black and white; then the difficulties or the inadequacies become obvious.

Another advantage of a written plan is that you can give it to others for their comments and criticism. In a research methods class, for example, the professor would certainly need to see what research students are planning. The director of a thesis or dissertation would want to see a written plan rather early in the process. It is much easier for another person to detect flaws and problems in a proposal that is written out than in one communicated orally. Another point to keep in mind is that the more complete and detailed the initial proposal, the more useful it will be to the researcher and the more time may be saved later.

A research plan should include the following elements: the problem, the hypothesis, the research methodology, and proposed data analysis. The following list briefly describes each component:

1. *Problem*. The plan begins with a clear statement of the research problem. A quantitative problem asks about the relationship between specified variables. Include the rationale for the study and a brief description of the background of the problem in theory and/or related research.
2. *Hypothesis*. A quantitative question is followed by a concise statement of the research hypothesis. Provide operational definitions of the variables.

3. *Methodology*. This section explains how you will conduct the study. Include the proposed research design, the population of concern, the sampling procedure, the measuring instruments, and any other information relevant to the conduct of the study.
4. *Data analysis*. Indicate how you will analyze the data to test the hypothesis and/or answer the research question. Beginning quantitative researchers may find it difficult to write this section because they are not yet familiar with statistics. You might look at the related literature to determine what type of statistical analysis other researchers used, or you might consult with your professor or an expert in statistics.

**Think About It 5.3**

State a hypothesis to test the notion that teachers assign rowdy students to remedial reading classes to get rid of them. State the null hypothesis and list the steps for testing it.

**Answer**

1. Research hypothesis: Students assessed as rowdy on a behavioral assessment scale are more often assigned to remedial reading classes than are nonrowdy students with equivalent reading skills as measured on the California Achievement Test.
2. Null hypothesis: Rowdy and nonrowdy students with the same reading skills are equally likely to be assigned to remedial reading classes.
3. Administer the Reading subtest of the California Achievement Test to all students. Match students in remedial reading classes with students with the same reading skills who are in regular classes. Use a behavioral assessment scale to identify which students are rowdy and which are not.
4. Calculate the proportion of rowdy and nonrowdy students in remedial reading classes and the proportion of rowdy and nonrowdy students in regular classes.
5. Test the null hypothesis by using a statistical test to determine if the difference in the proportions could easily be a function of chance alone.

## THE PILOT STUDY

After the tentative research plan is approved, it may be helpful to try out the proposed procedures on a few participants. This trial run, or **pilot study,** will help the researcher to decide whether the study is feasible and whether it is worthwhile to continue. At this point, one can ask a colleague to check one's procedures for any obvious flaws. The pilot study provides the opportunity to assess the appropriateness of the data-collection methods and other procedures and to make changes if necessary. It also permits a preliminary testing of the hypothesis, which may give some indication of its tenability and suggest whether further refinement is needed.

Unanticipated problems that appear can be solved at this stage, thereby saving time and effort later. A pilot study is well worth the time required and is especially recommended for the beginning researcher.

## SUMMARY

To proceed with the confirmatory phase of a quantitative research study, it is important to have one or more clearly stated hypotheses. The hypothesis is a powerful tool in scientific inquiry. It enables researchers to relate theory to observation and observation to theory. Hypotheses enable researchers, in the search for knowledge, to employ both the ideas of the inductive philosophers, with their emphasis on observation, and the logic of the deductive philosophers, with their emphasis on reason.

The hypothesis is the researcher's prediction about the outcome of the study. Hypotheses are derived inductively from observation or deductively from a known theory. Experience and knowledge in an area and familiarity with previous research are important factors in formulating a satisfactory hypothesis.

The hypothesis serves multiple functions in research. It provides direction to the researcher's efforts because it determines the research method and the type of data relevant to the solution of the problem. It also provides a framework for interpreting the results and for stating the conclusions of the study.

A good hypothesis must satisfy certain criteria. It must be testable, which means that it is possible to gather evidence that will either support or fail to support the hypothesis. It must agree with the preponderance of existing data. It must be stated as clearly and concisely as possible. Also, it must state the expected relationship between variables that can be measured.

Once formulated and evaluated in terms of these criteria, the research hypothesis is ready to be subjected to an empirical test. The researcher also states a null hypothesis—the negation of what the researcher expects—which is important in the statistical analysis of the findings. It is important to remember that a research hypothesis cannot be proved or disproved, only supported or not supported. Even if it is not supported, a hypothesis may still serve a useful purpose because it can lead the researcher to reevaluate rationale and procedures and to consider other approaches to the problem.

## KEY CONCEPTS

alternative hypothesis
deductive hypothesis
directional hypothesis
hypothesis
hypothesis testing
inductive hypothesis
nondirectional hypothesis
null hypothesis
operational definition
pilot study
purposes of hypotheses
research hypothesis
research plan
testable hypotheses
theory

## EXERCISES

1. What are the purposes of the hypothesis in research?
2. What is the difference between an inductive and a deductive hypothesis?
3. State a hypothesis based on each of the following research questions:
   - **a.** What would be the effect of using the Cuisenaire method in teaching elementary arithmetic?
   - **b.** Is there a relationship between the gender of the tutor and the gains made in reading achievement by black male elementary students?
   - **c.** Does living in interracial housing affect attitudes toward members of another race?
   - **d.** Is there any relationship between the type of reinforcement (tangible or intangible) and the amount of learning achieved by socioeconomically disadvantaged children?
   - **e.** Does preschool experience reduce the educational gap separating advantaged and disadvantaged children before they enter first grade?
   - **f.** Do teacher expectations of children's intellectual performance have any effect on the children's actual performance?
4. Rewrite the following hypothesis in null form: "Children who read below grade level

will express less satisfaction with school than those who read at or above grade level."

5. Evaluate the adequacy of each of the following hypotheses. If a hypothesis is inadequate, state the reason for the inadequacy and write an adequate hypothesis.
   - **a.** "Teachers deserve higher pay than administrators."
   - **b.** "Students who take a middle school government course will be capable of more enlightened judgments concerning local political affairs than will those who do not take the course."
   - **c.** "Computer-based drill and practice is a better way to teach slow learners multiplication combinations than is flash cards."
   - **d.** "If students differ in their socioeconomic status, they will differ in their English proficiency scores."
   - **e.** "Children who show high achievement motivation will show high anxiety as measured by the Children's Manifest Anxiety Scale."
   - **f.** "Positive verbal reinforcement of student responses by the teacher will lessen the probability of future responses."
6. Write a directional and a nondirectional hypothesis based on the research question "What is the relationship between the rate of maturation of adolescent boys and their self-concepts?"
7. Why should a hypothesis be clearly stated before a quantitative research study is initiated?
8. Label the following hypotheses as research hypotheses or null hypotheses:
   - **a.** "Students will receive lower scores on achievement tests that measure the higher levels of Bloom's taxonomy than on tests measuring lower levels of Bloom's taxonomy.
   - **b.** "There is no difference in the performance of students taught mathematics by method A and those taught mathematics by method B."
   - **c.** "The mean retention scores of children receiving experimental drug X will not differ from the scores of children who do not receive drug X."
   - **d.** "Students taught by laissez-faire teachers will show higher problem-solving skills than students taught by highly structured teachers."
9. Locate a research study stating a hypothesis and try to identify the theory from which the hypothesis originated.
10. Evaluate the following statements as possible research hypotheses:
    - **a.** "Asian high school students are better in mathematics than American high school students."
    - **b.** "Do SAT prep courses improve students' scores on the SAT?"
    - **c.** "Students who participate in the high school volunteerism program become better adult citizens than students who do not."
11. A researcher has a theory about children's ordinal position in the family and their achievement motivation. Write a research hypothesis and a hypothesis in null form.
12. Formulate a tentative research plan for your class project.
    - **a.** What is the general research problem under consideration for investigation?
    - **b.** State the preceding general research problem as a research question.
    - **c.** Explain the rationale for such a study. What are its theoretical or practical applications?
    - **d.** State the hypothesis (or hypotheses) for this study.
    - **e.** Was this hypothesis derived deductively from theory or inductively from experience and observation? Explain.
    - **f.** Identify the variables in the study and operationally define each.
    - **g.** What kind of research methodology will be required for this study?
    - **h.** What subjects (sample) will you select for the study?
    - **i.** Have you located any published research related to your problem? If so, briefly summarize the findings.
13. Which of the following evidence contributes to the development of a theory?
    - **a.** Evidence that supports a hypothesis
    - **b.** Evidence that contradicts a hypothesis
    - **c.** Both of the above
14. Select a theory that you find interesting and derive a research hypothesis from this theory. You might choose a learning theory, motivational theory, theory of cognitive dissonance, or any other educational or psychological theory.

## ANSWERS

1. The purposes of hypotheses are to provide a tentative proposition suggested as a solution to a problem or as an explanation of some phenomenon, stimulate research, provide a relational statement that is directly testable, and provide direction for research.
2. With an inductive hypothesis, the researcher makes observations of relationships and then hypothesizes an explanation for the observed behavior. With a deductive hypothesis, the researcher formulates a hypothesis based on known theory, accompanied by a rationale for the particular proposition.
3. **a.** "Elementary students taught by the Cuisenaire method will score higher on an arithmetic test than students taught by an alternative method."
   **b.** "Black male elementary students tutored by another male will achieve higher reading scores than will black male elementary students tutored by a female."
   **c.** "People living in interracial housing will express more favorable attitudes toward those of another race than will people living in segregated housing."
   **d.** "Socioeconomically disadvantaged children reinforced with tangible rewards will exhibit greater learning achievement than will children reinforced with intangible rewards."
   **e.** "Advantaged and disadvantaged children of preschool age receiving preschool training will be separated by a smaller educational gap than will advantaged and disadvantaged children of preschool age not receiving preschool training." (*Note*: The terms advantaged and disadvantaged children, preschool training, and educational gap would need to be defined.)
   **f.** "Children whose teachers have high expectations of their intellectual performance will perform at a higher level than will children whose teachers have low expectations of their intellectual performance."
4. There is no difference in the satisfaction with school expressed by children who read below grade level and children who read at or above grade level.
5. **a.** The hypothesis is inadequate because it is a value statement and cannot be investigated in a research study. A legitimate hypothesis would be "Teachers who receive higher pay than their administrators will express greater job satisfaction than will teachers who do not receive higher pay than their administrators."
   **b.** The hypothesis is inadequate because enlightened judgments is a value term. An acceptable hypothesis would be "Students who take a middle school government course will evidence more knowledge concerning local political affairs, and will more often arrive at inferences based on this knowledge, than will students who do not take a middle school government course."
   **c.** The hypothesis is inadequate because *better* is a value term and because it lacks clear and concise operational definitions. A testable hypothesis would be "Those students performing below grade level in math who practice multiplication combinations through computer drill and practice will, on average, score a higher proportion of correct answers on a criterion test than will students performing below grade level who spend the same amount of time practicing multiplication combinations with flash cards."
   **d.** The hypothesis is inadequate because there is no statement of an expected relationship between variables. An acceptable hypothesis would be "Students classified as having high socioeconomic status will have higher scores on an English proficiency test than will students classified as having low socioeconomic status."
   **e.** The hypothesis is inadequate because there are no independent or dependent variables. An acceptable hypothesis would be "Children who show high achievement motivation will have higher scores on the Children's Manifest Anxiety Scale than children with low achievement motivation."
   **f.** The hypothesis is inadequate because it is inconsistent with the existing knowledge of positive reinforcement and its effect on student responses.

6. *Directional hypothesis*: "Early maturing boys will exhibit more positive self-concepts than late-maturing boys." *Nondirectional hypothesis*: "There is a difference in the self-concepts of early and late-maturing adolescent boys."
7. The hypothesis gives direction to the collection and interpretation of data. Clearly stating the hypothesis reveals flaws that were not apparent while developing the vague idea of the study in mind.
8. **a.** Research
   **b.** Null
   **c.** Null
   **d.** Research
9. Answers will vary.
10. **a.** "Better in math" needs to be operationally defined.
    **b.** A hypothesis should not be stated in question form.
    **c.** It is not testable as stated. How would you define and measure "better adult citizens"?
11. *Research hypothesis*: "Achievement motivation and ordinal birth position in the family are positively related; or first-born children have greater achievement motivation than their siblings." *Null hypothesis*: "There is no relationship between children's birth position in the family and their achievement motivation."
12. Answers will vary.
13. c
14. Answers will vary.

## REFERENCES


Butler, R. J. (2001). *The self-image profile for children*. Oxford: Psychological Corporation.

Darwin, F. (Ed.). (2007). *The life and letters of Charles Darwin* (Vol. 1). New York: Echo Library. [Original work published 1887]

Festinger, L. (1957). *A theory of cognitive dissonance*. Evanston, IL: Row, Peterson. [Reprinted 2003 by Textbook Publishers]

Frager, R., & Fadiman, J. (1997). *Maslow's motivation and personality*. Boston: Addison-Wesley.

Gardner, H. (1993). *Multiple intelligences: The theory in practice*. New York: Basic Books.

Maslow, A. H. (1954). *Motivation and personality*. New York: Harper & Row.

McClelland, D. C. (1953). *The achievement motive*. New York: Appleton-Century-Crofts.

Piaget, J. (1968). *Six psychological studies*. New York: Vintage Books.

Skinner, B. F. (1953). *Science and human behavior*. New York: Macmillan. [Reprinted 2003 by Textbook Publishers]

Vygotsky, L. S. (1978). In M. Cole, V. John-Steiner, S. Scribner, & E. Souberman (Eds.), *Mind in society*. Cambridge, MA: Harvard University Press.

Weiner, B. (1994). Integrating social and personal theories of achievement striving. *Review of Educational Research*, *64*, 557–573.

CHAPTER 6

# Descriptive Statistics

Descriptive statistics simplify our lives by organizing and summarizing data.

## INSTRUCTIONAL OBJECTIVES

After studying this chapter, the student will be able to:

1. Describe the nature and uses of descriptive statistics.
2. Identify the characteristics, uses, and limitations of four types of measurement scales—nominal, ordinal, interval, and ratio.
3. Organize research data into frequency distributions, present them as frequency polygons and histograms, and interpret polygons and histograms occurring in the professional literature.
4. Distinguish between the measures of central tendency and the situations in which each should be used. Calculate and interpret the mean, the median, and the mode for any given data.
5. Describe appropriate applications of measures of variability and compute variance, standard deviation, and range for any given set of data.
6. Calculate and explain why *z* scores have universal meaning and how this is useful in interpreting the position of a single observation in a distribution.
7. Explain why *z* scores are often transformed into other standard scores.
8. Convert a *z* score to a stanine score and use this to give a verbal description of the score's meaning. Explain why stanine scores are easy to interpret.
9. Transform raw scores into standard scores.
10. Explain advantages and disadvantages of percentile ranks. Calculate percentile rank for a given score.
11. Identify the characteristics of the normal curve. Explain why it is useful in descriptive research.
12. Use the normal curve table to estimate the percentile rank of a given *z* score or estimate the *z* score of a given percentile rank.
13. Identify appropriate applications of Pearson *r* correlation for describing the relationship between variables. Explain why it shows both the direction and the strength of the relationship.
14. Describe the meaning of coefficient of determination and its application in interpreting the coefficient of correlation.

15 Identify the components of effect size and what factors increase and decrease effect size. Explain why number in sample does not influence effect size.

16 Explain how effect size assesses the strength of relationships between variables.

17 Calculate effect size for a difference between means. Explain why the Pearson *r* is a form of effect size.

18 Perform a meta-analysis and explain the meaning of a meta-analysis outcome.

Statistical procedures are basically methods of handling quantitative information. These procedures have two principal advantages. First, they enable researchers to organize, summarize, and describe observations. Techniques used for these purposes are called **descriptive statistics.** Second, they help determine how reliably researchers can infer that phenomena observed in a limited group—a *sample*—are likely to occur in the unobserved larger population of concern from which the sample was drawn—in other words, how accurately you can employ inductive reasoning to infer that what you observe in the part will be observed in the whole. Techniques used for such purposes are called **inferential statistics.**

Knowledge of some basic statistical procedures is essential for researchers proposing to carry out quantitative research. They need statistics to analyze and interpret their data and communicate their findings to others in education. Researchers also need an understanding of statistics in order to read and evaluate published research in their fields.

## SCALES OF MEASUREMENT

A fundamental step in the conduct of quantitative research is measurement—the process through which observations are translated into numbers. S. S. Stevens (1951) is well remembered for his definition: "In its broadest sense, measurement is the assignment of numerals to objects or events according to rules" (p. 1). Quantitative researchers first identify the variables they want to study; then they use rules to determine how to express these variables numerically. The variable *religious preference* may be measured according to the numbers indicated by students who are asked to select among (1) Catholic, (2) Jewish, (3) Protestant, (4) Muslim, or (5) other. The variable *weight* may be measured as the numbers observed when subjects step on a scale. The variable self-concept may be operationally defined as scores on the Multidimensional *Self-Concept* Scale. The nature of the measurement process that produces the numbers determines the interpretation that can be made from them and the statistical procedures that can be meaningfully used with them. The most widely quoted taxonomy of measurement procedures is Stevens' scales of measurement in which he classifies measurement as nominal, ordinal, interval, and ratio.

### NOMINAL SCALE

The most primitive scale of measurement is the **nominal scale.** Nominal measurement involves placing objects or individuals into mutually exclusive categories. Numbers are arbitrarily assigned to the categories for identification purposes

only. The numbers do not indicate any value or amount; thus, one category does not represent "more or less" of a characteristic. School District 231 is not more or less of anything than School District 103. Examples of a nominal scale are using a "0" to represent males and a "1" to represent females or the religious preference described previously.

Because the numbers in a nominal scale do not represent quantity, they cannot be arithmetically manipulated through addition, subtraction, multiplication, or division. One can only count the number of observations in each category or express the numbers in categories as a percentage of the total number of observations.

## ORDINAL SCALE

An **ordinal scale** ranks objects or individuals according to how much of an attribute they possess. Thus, the numbers in an ordinal scale indicate only the order of the categories. Neither the difference between the numbers nor their ratio has meaning. For example, in an untimed footrace, we know who came in first, second, and third, but we do not know how much faster one runner was than another. A ranking of students in a music contest is an ordinal scale. We would know who got first place, second place, and so on, but we would not know the extent of difference between them.

The essential requirement for measurement at this level is that the relationship must be such that if object X is greater than object Y and object Y is greater than object Z, then object X is greater than object Z and is written thus: If $(X > Y)$ and $(Y > Z)$, then $(X > Z)$. When appropriate, other wording may be substituted for "greater than," such as "stronger than," "precedes," and "has more of."

The lack of equal intervals in ordinal scales limits the statistical procedures available for analyzing ordinal data. We can use statistics that indicate the points below which certain percentages of the cases fall in a distribution of scores.

## INTERVAL SCALE

An **interval scale** not only places objects or events in order but also is marked in equal intervals. Equal differences between the units of measurement represent equal differences in the attribute being measured. Fahrenheit and Celsius thermometers are examples of interval scales. We can say that the difference between 60° and 70° is the same as the distance between 30° and 40°, but we cannot say that 60° is twice as warm as 30° because there is no true zero on an interval scale. Zero on an interval scale is an arbitrary point and does not indicate an absence of the variable being measured. Zero on the Celsius scale is arbitrarily set at the temperature water freezes at sea level.

Numbers on an interval scale may be manipulated by addition and subtraction, but because the zero is arbitrary, multiplication and division of the numbers are not appropriate. Thus, ratios between the numbers on an interval scale are meaningless. We may report differences between positions on an interval scale or we may add the numbers to report an average.

It is important to note that in most academic measures, the intervals are equal in terms of the measuring instrument but not necessarily in terms of the performance being measured. To illustrate, consider a spelling test with the following words: *cat, dish, ball, loquacious, schizophrenia*, and *pneumonia*. Here, the distance between one correct and three correct is the same as the distance between three correct and five correct. However, when considered in terms of spelling performance, the difference between three and five correct suggests a greater difference in ability than does the difference between one and three correct. Unless you can say that the distance between three and five on the spelling test represents the same amount of spelling performance as does the distance between one and three, then these scores indicate nothing more than the rank order of the students.

However, through careful construction it is possible to produce an instrument where the intervals observed between scores on the test give a reasonable approximation of ability intervals. The better intelligence tests are an example of this. The difference in ability between an IQ of 90 and an IQ of 95 may not be precisely the same as the difference between an IQ of 105 and an IQ of 110, but you will not be greatly misled if you assume that the two differences are approximately equal.

It has become common practice to treat many educational variables, such as classroom tests and grades (A = 4, B = 3, and so on), as if they were interval data, even when that assumption is not well justified. It would be difficult to maintain that the difference between F = 0 and D = 1 represents the same difference in academic achievement as the difference between C = 2 and B = 3, or to justify treating scores on our spelling test example as interval data. Be cautious when interpreting statistics derived from such data. The statistics imply interval-level information when the information is actually often somewhere between ordinal and interval.

However, in schools and universities, grade point average is almost always calculated as if the data are interval. Scores on teacher-made tests are usually quasi-interval, somewhere between interval and ordinal data. Although scores on teacher-made tests are probably rarely as far from interval data as our spelling test example, they almost always vary from true interval data to some extent.

## RATIO SCALE

A ratio scale, the highest level of measurement scale, has a true zero point as well as equal intervals. Ratios can be reported between any two given values on the scale. A yardstick used to measure length in units of inches or feet is a ratio scale because the origin on the scale is an absolute zero corresponding to no length at all. Thus, it is possible to state that a stick 6 feet long is twice as long as a stick 3 feet long. Other examples of ratio scales are weight, money, and distance. All types of statistical procedures are appropriate with a ratio scale. Only a few variables in education are ratio in nature. These are largely confined to motor performance and other physiological measures. A shot put score of 16 yards is twice as far as a shot put score of 8 yards, but you cannot say that a person who scores 40

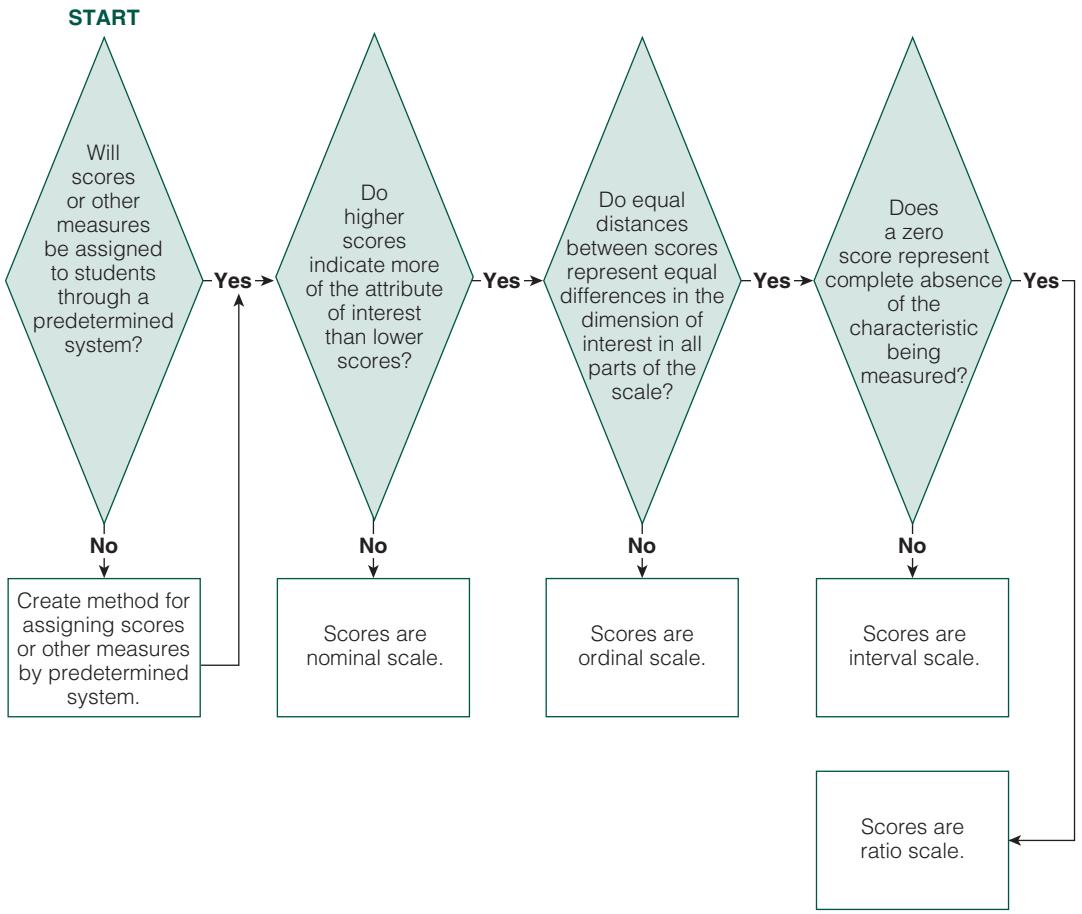



**Figure 6.1** Determining Scales of Measurement

## THINK ABOUT IT 6.1

You are buying a used car. You consider (a) its make (Ford, Toyota, etc.), (b) the miles on the odometer, (c) the year it was made, and (d) its rating in *Consumer Reports*.

1. Which of the above is nominal data?
2. Which of the above is ordinal data?
3. Which of the above is interval data?
4. Which of the above is ratio data?

**Answers**

1. a; 2. d; 3. c (the year 2008 is not twice the year 1004); 4. b

on a math test is twice as good at math as a person who scores 20 because math test scores are not ratio data. Figure 6.1 shows the decisions made to determine the scale of measurement of an observation.

# ORGANIZING RESEARCH DATA

Researchers typically collect a large amount of data. Before applying statistical procedures, the researcher must organize the data into a manageable form. The most familiar ways of organizing data are (1) arranging the measures into frequency distributions and (2) presenting them in graphic form.

## FREQUENCY DISTRIBUTIONS

A systematic arrangement of individual measures from highest to lowest is called a **frequency distribution.** The first step in preparing a frequency distribution is to list the scores in a column from highest at top to lowest at bottom. Include all possible intermediate scores even if no one scored them; otherwise, the distribution will appear more compact than it really is. Several identical scores often occur in a distribution. Instead of listing these scores separately, it saves time to add a second column in which the frequency of each measure is recorded. Table 6.1 shows the test scores of a group of 105 students in an Ed 101 lecture class. Part A of the table lists the scores in an unorganized form. Part B arranges these scores in a frequency distribution with the *f* column showing how many made each score. Now it is possible to examine the general "shape" of the distribution. With the scores so organized, you can determine their spread, whether they are distributed evenly or tend to cluster, and where clusters occur in the distribution. For example, looking over the frequency distribution of the scores presented in Part B of Table 6.1, it is easy to see that they range from 21 to 36, that 29 is the most frequent score, and that scores tend to cluster more near the top of the distribution than the bottom. None of this would be apparent had the scores not been organized. Organizing data into frequency distributions also facilitates the computation of various useful statistics.

### THINK ABOUT IT 6.2

Here are the scores that Mr. Li's 18 physics class students made on their first exam: Ali, 21; Ann, 20; Ben, 23; Cal, 20; Dan, 20; Ed, 21; Ima, 22; Jan,19; Kay, 16; Lee, 20; Mel, 18; Mia, 23; Ned, 21; Ona, 21; Sam, 22; Sue, 19; Ted,16; Van, 18. Do a frequency distribution of these scores. For the answer, see the first two columns in Table 6.2.

## GRAPHIC PRESENTATIONS

It is often helpful and convenient to present research data in graphic form. Among various types of graphs, the most widely used are the **histogram** and the **frequency polygon.** The initial steps in constructing the histogram and the frequency polygon are identical:

1. Lay out the score points on a horizontal dimension (abscissa) from the lowest value on the left to the highest on the right. Leave enough space for an additional score at both ends of the distribution.

**Table 6.1** The Test Scores of 105 Students on Ed 101 Test

**Part A. Unorganized Scores**

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 33 | 29 | 30 | 30 | 33 | 29 | 33 | 32 | 28 | 24 | 34 | 31 | 27 | 29 | 23 |
| 25 | 29 | 24 | 27 | 26 | 33 | 33 | 26 | 30 | 28 | 26 | 29 | 32 | 32 | 31 |
| 28 | 34 | 30 | 31 | 33 | 21 | 29 | 31 | 30 | 32 | 36 | 30 | 31 | 27 | 29 |
| 26 | 29 | 33 | 32 | 29 | 28 | 28 | 30 | 28 | 27 | 30 | 31 | 34 | 33 | 22 |
| 30 | 29 | 27 | 29 | 24 | 30 | 21 | 31 | 31 | 33 | 28 | 21 | 31 | 29 | 31 |
| 31 | 33 | 22 | 29 | 31 | 32 | 32 | 31 | 28 | 29 | 30 | 22 | 33 | 30 | 30 |
| 32 | 33 | 31 | 33 | 28 | 29 | 27 | 33 | 27 | 21 | 30 | 29 | 28 | 27 | 33 |

**Part B. Frequency Distribution**

| Scores ($X$) | Tallies | Frequencies ($f$) | $fX$ | $cf$ |
|---|---|---|---|---|
| 36 | / | 1 | 36 | 105 |
| 35 | | 0 | | 104 |
| 34 | /// | 3 | 102 | 104 |
| 33 | //// //// //// | 15 | 405 | 101 |
| 32 | //// /// | 8 | 256 | 80 |
| 31 | //// //// //// | 14 | 434 | 78 |
| 30 | //// //// //// | 14 | 420 | 64 |
| 29 | //// //// //// / | 16 | 464 | 50 |
| 28 | //// //// | 10 | 280 | 34 |
| 27 | //// /// | 8 | 216 | 29 |
| 26 | //// | 4 | 104 | 16 |
| 25 | / | 1 | 25 | 12 |
| 24 | /// | 3 | 72 | 11 |
| 23 | / | 1 | 23 | 8 |
| 22 | /// | 3 | 66 | 7 |
| 21 | //// | 4 | 84 | 4 |
| | | $N = 105$ | | |

We discuss the $fX$ and $cf$ columns later.

**Table 6.2** Mr. Li's Physics Class Exam Scores

| (1) $X$ | (2) $f$ | (3) $fX$ | (4) $cf$ |
|---|---|---|---|
| 23 | 2 | 46 | 18 |
| 22 | 2 | 44 | 16 |
| 21 | 4 | 84 | 14 |
| 20 | 4 | 80 | 10 |
| 19 | 2 | 38 | 6 |
| 18 | 2 | 36 | 4 |
| 17 | 0 | 0 | 2 |
| 16 | 2 | 32 | 2 |

2. Lay out the frequencies of the scores (or intervals) on the vertical dimension (ordinate).
3. Place a dot above the center of each score at the level of the frequency of that score.

From this point you can construct either a histogram or a polygon. In constructing a histogram, draw through each dot a horizontal line equal to the width representing a score, as shown in Figure 6.2. A score of 26 is thought of as ranging from 25.5 to 26.5, a score of 27 is thought of as ranging from 26.5 to 27.5, and so forth.

To construct a polygon, connect the adjacent dots, and connect the two ends of the resulting figure to the base (zero line) at the points representing 1 less than the lowest score and 1 more than the highest score, as shown in Figure 6.3. Histograms are preferred when a researcher wants to indicate the discrete nature of the data, such as when a nominal scale has been used. Polygons are preferred to indicate data of a continuous nature.

## THINK ABOUT IT 6.3

Construct a histogram and a polygon of the scores of Mr. Li's first physics exam.

**Answer**

See Figure 6.4.

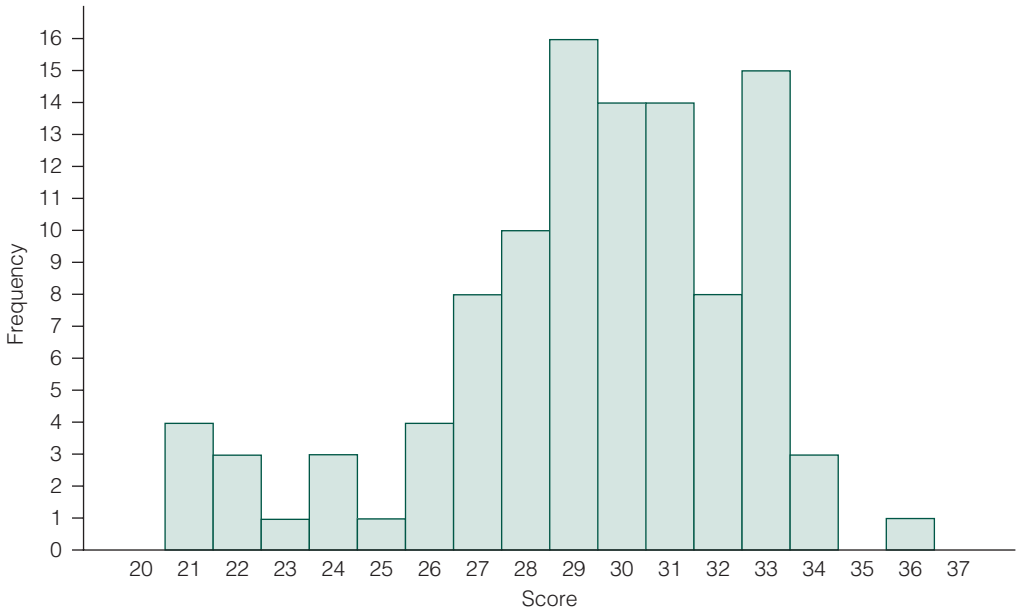



**Figure 6.2** Histogram of 105 Test Scores from Table 6.1

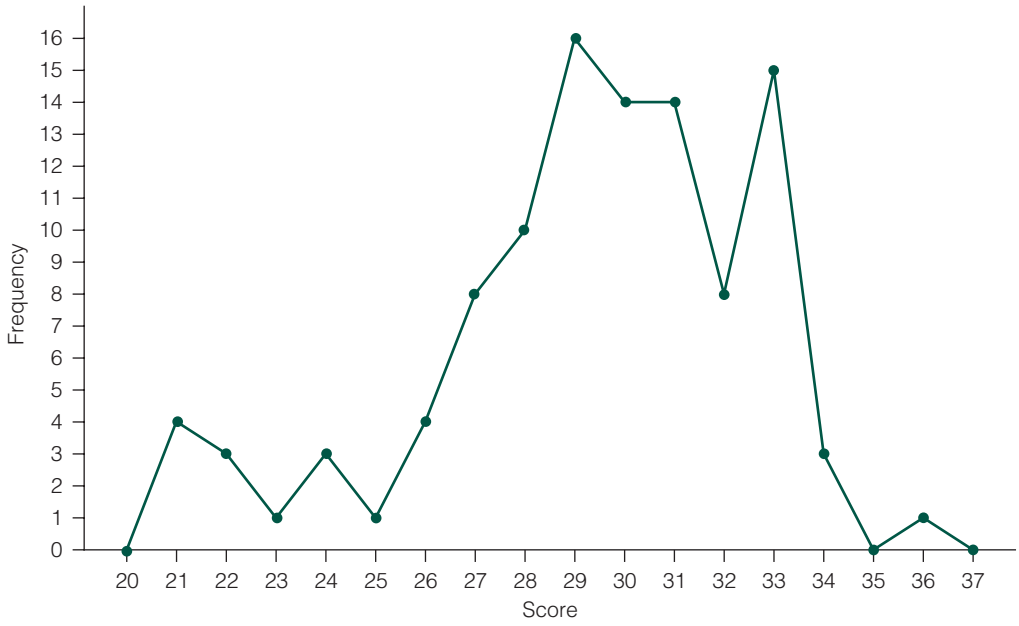



**Figure 6.3** Frequency Polygon of 105 Test Scores from Table 6.1

## MEASURES OF CENTRAL TENDENCY

A convenient way of summarizing data is to find a single index that can represent a whole set of measures. Finding a single score that can give an indication of the performance of a group of 300 individuals on an aptitude test would be useful for comparative purposes. In statistics, three indexes are available for such use. They are called **measures of central tendency,** or averages. To most laypeople, the term *average* means the sum of the scores divided by the number of scores. To a statistician, the average can be this measure, known as the mean, or one of the other two measures of central tendency, known as the *mode* and the *median*. Each of these three can serve as an index to represent a group as a whole.

### THE MEAN

The most widely used measure of central tendency is the **mean,** or arithmetic average. It is the sum of all the scores in a distribution divided by the number of cases. In terms of a formula, it is

$$\overline{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{N} \tag{6.1}$$

which is usually written as

$$\overline{X} = \frac{\Sigma X}{N} \tag{6.2}$$

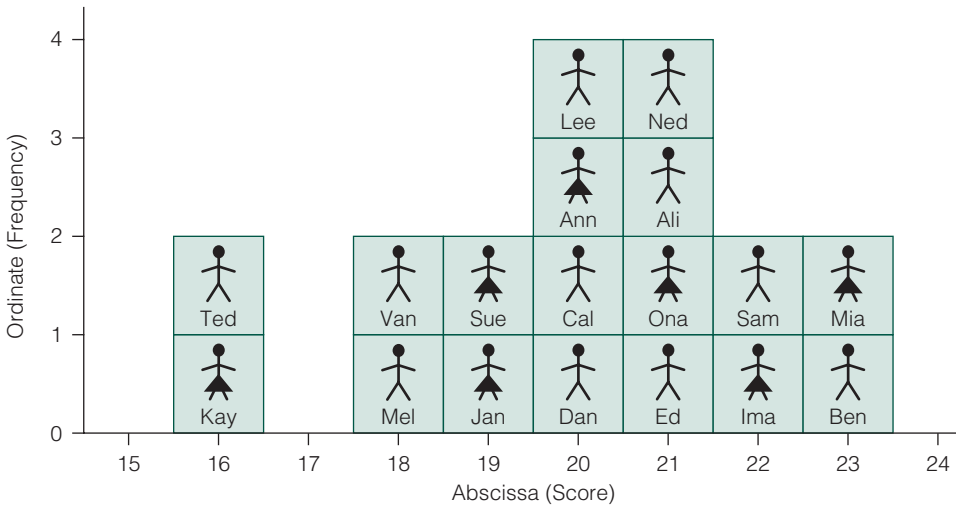
where

$\overline{X}$ = mean
$\Sigma$ = sum of
$X$ = raw score
$N$ = number of cases

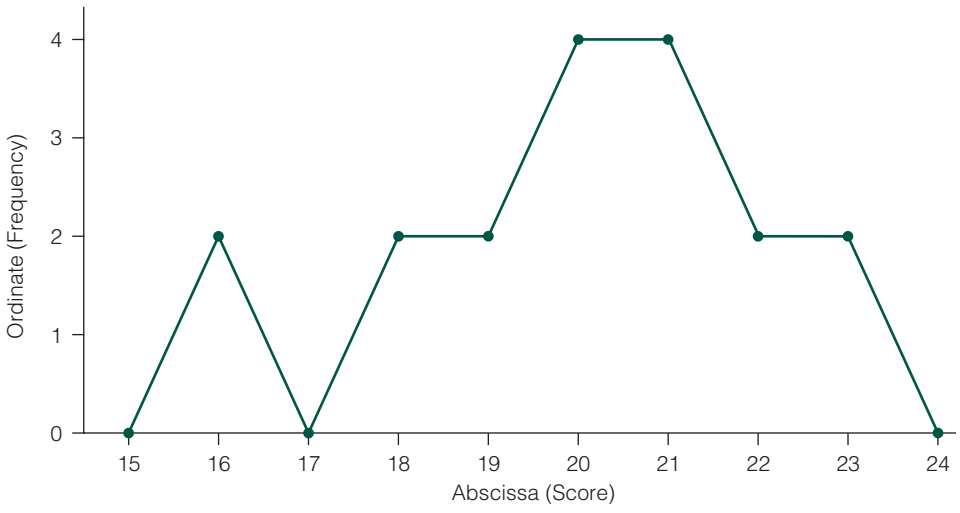
Applying Formula 6.2 to the following IQ scores, we find that the mean is 111:

IQ scores: 112 121 115 101 119 109 100

$$\overline{X} = \frac{112 + 121 + 115 + 101 + 119 + 109 + 100}{7} = \frac{777}{7} = 111$$




(A)




(B)

**Figure 6.4** (A) Histogram of Mr. Li's Physics Exam and (B) Polygon of Mr. Li's Physics Exam

Note that in this computation the scores were not arranged in any particular order. Ordering is unnecessary for calculation of the mean.

Some think of formulas as intimidating incantations. Actually, they are time savers. It is much easier to write $\overline{X} = \Sigma X/N$ than to write "add all the scores in a distribution and divide by the number of cases to calculate the mean."

Although it is not necessary to put the scores in order to calculate the mean, with larger sets of numbers it is usually convenient to start with a frequency distribution and multiply each score by its frequency. This is shown in column 3 ($fX$) in Table 6.2, Mr. Li's physics class exam scores. Adding the numbers in this column will give us the sum of the scores.

$$\Sigma X = 360$$

The mean of the physics exam scores is

$$\overline{X} = \frac{\Sigma X}{N} = \frac{360}{18} = 20$$

## THE MEDIAN

The **median** is defined as that point in a distribution of measures below which 50 percent of the cases lie (which means that the other 50 percent will lie above this point). Consider the following distribution of scores, where the median is 18:

14 15 16 17 18 19 20 21 22

In the following 10 scores we seek the point below which 5 scores fall:

14 16 16 17 18 19 20 20 21 22

The point below which 5 scores, or 50 percent of the cases, fall is halfway between 18 and 19. Thus, the median of this distribution is 18.5.

Consider the following scores:

18 20 22 25 25 30

Any point from 22.5 to 24.5 fits the definition of the median. By convention in such cases the median is defined as half way between these lowest and highest points, in this case $22.5 + 24.5/2 = 23.5$.

To find the median of Mr. Li's physics exam scores, we need to find the point below which $18/2 = 9$ scores lie. We first create a cumulative frequency column ($cf$, column 4 in Table 6.2). The cumulative frequency for each interval is the number of scores in that interval plus the total number of scores below it. Since the interval between 15.5 and 16.5 has no scores below it, its $cf$ is equal to its $f$, which is 2. Since there were no scores of 17, the $cf$ for 17 is still 2. Then adding the two scores of 18 yields a cumulative frequency of 4. Continuing up the frequency column, we get $cf$'s of 10, 14, 16, and, finally, 18, which is equal to the number of students.

The point separating the bottom nine scores from the top nine scores, the median, is somewhere in the interval 19.5 to 20.5. Most statistics texts say to partition this interval to locate the median. The $cf$ column tells us that we have six scores below 19.5. We need to add three scores to give us half the scores (9). Since there are four scores of 20, we go three-fourths of the way from 19.5 to 20.5

to report a median of 20.25. Note that many computer programs, including the Statistical Package for the Social Sciences (SPSS) and the Statistical Analysis System (SAS), simply report the midpoint of the interval—in this case 20—as the median.

Notice that the median does not take into account the size of individual scores. In order to find it, you arrange your data in rank order and find the point that divides the distribution into two equal halves. The median is an ordinal statistic because it is based on rank. You can compute a median from interval or ratio data, but in such cases the interval characteristic of the data is not being used. One circumstance in which the median may be the preferred measure of central tendency arises when there are some extreme scores in the distribution. In this case, the use of a measure of central tendency that takes into account the size of each score results in either overestimation or underestimation of the typical score. The median, because of its insensitivity to extreme scores, is the appropriate index to be applied when you want to find the typical score. For illustration, consider the following distribution:

49 50 51 53 54 55 56 60 89

The score of 54, which is the median of this distribution, is the most typical score. The mean, which takes into account the individual values of the scores 60 and 89, will certainly result in an overestimation of the typical score.

## THE MODE

The **mode** is the value in a distribution that occurs most frequently. It is the simplest to find of the three measures of central tendency because it is determined by inspection rather than by computation. Given the distribution of scores

14 16 16 17 18 19 19 19 21 22

you can readily see that the mode of this distribution is 19 because it is the most frequent score. In a histogram or polygon, the mode is the score value of the highest point (the greatest frequency), as you can see in Figures 6.2 and 6.3, where the mode is 29. Sometimes there is more than one mode in a distribution. For example, if the scores had been

14 16 16 16 18 19 19 19 21 22

you would have two modes: 16 and 19. This kind of distribution with two modes is called *bimodal*. Distributions with three or more modes are called *trimodal* or *multimodal*, respectively.

The mode is the least useful indicator of central value in a distribution for two reasons. First, it is unstable. For example, two random samples drawn from the same population may have quite different modes. Second, a distribution may have more than one mode. In published research, the mode is seldom reported as an indicator of central tendency. Its use is largely limited to inspectional purposes. A mode may be reported for any of the scales of measurement, but it is the only measure of central tendency that may legitimately be used with nominal scales.

## COMPARISON OF THE THREE INDEXES OF CENTRAL TENDENCY

Because the mean is an interval or ratio statistic, it is generally a more precise measure than the median (an ordinal statistic) or the mode (a nominal statistic). It takes into account the value of *every* score. It is also the most stable of the three measures of central tendency in that if a number of samples are randomly drawn from a parent population, the means of these samples will vary less from one another than will their medians and their modes. For these reasons, the mean is more frequently used in research than the other two measures.

The mean is the best indicator of the combined performance of an entire group. However, the median is the best indicator of *typical* performance. Consider, for example, a school board whose members have the following annual incomes: $140,000, $60,000, $50,000, $40,000, and $40,000. The mean, $66,000, is the sum of their incomes divided by the number of members, but it is higher than all but one of the board members' incomes. The median, $50,000, gives a better picture of the typical income in the group.

## SHAPES OF DISTRIBUTIONS

Frequency distributions can have a variety of shapes. A distribution is symmetrical when the two halves are mirror images of each other. In a **symmetrical distribution,** the values of the mean and the median coincide. If such a distribution has a single mode, rather than two or more modes, the three indexes of central tendency will coincide, as shown in Figure 6.5.

If a distribution is not symmetrical, it is described as **skewed,** pulled out to one end or the other by the presence of extreme scores. In skewed distributions, the values of the measures of central tendency differ. In such distributions, the value of the mean, because it is influenced by the size of extreme scores, is pulled toward the end of the distribution in which the extreme scores lie, as shown in Figures 6.6 and 6.7. The effect of extreme values is less on the median because this index is influenced not by the size of scores but by their position. Extreme values have no impact on the mode because this index has no relation with either of the ends of the distribution. Skews are labeled according to where the extreme

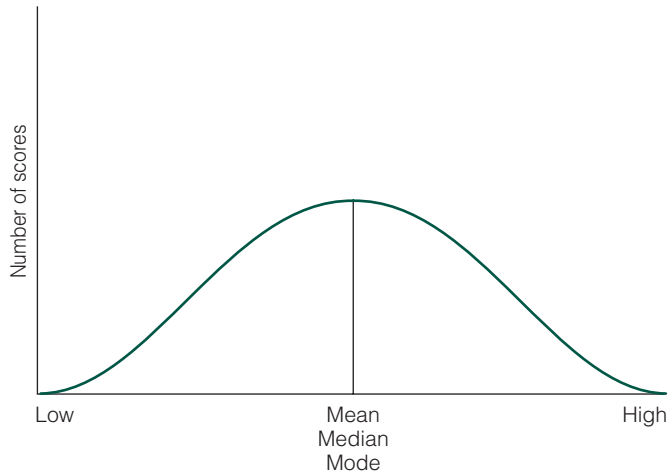



**Figure 6.5** Symmetrical Distribution

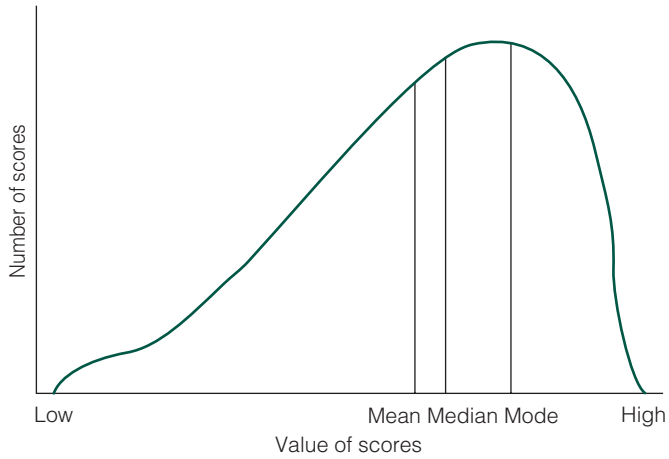



**Figure 6.6** Negatively Skewed Distribution

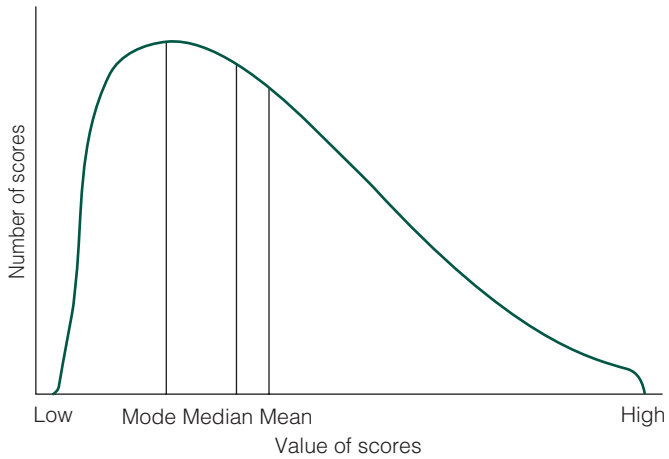



**Figure 6.7** Positively Skewed Distribution

scores lie. A way to remember this is "The tail names the beast." Figure 6.6 shows a **negatively skewed distribution,** whereas Figure 6.7 shows a **positively skewed distribution.**

## MEASURES OF VARIABILITY

Although indexes of central tendency help researchers describe data in terms of average value or typical measure, they do not give the total picture of a distribution. The mean values of two distributions may be identical, whereas the degree of dispersion, or **variability,** of their scores might be different. In one distribution, the scores might cluster around the central value; in the other, they might be scattered. For illustration, consider the following distributions of scores:

(a) 24, 24, 25, 25, 25, 26, 26 $\quad \overline{X} = 175/7 = 25$
(b) 16, 19, 22, 25, 28, 30, 35 $\quad \overline{X} = 175/7 = 25$

The value of the mean in both these distributions is 25, but the degree of scattering of the scores differs considerably. The scores in distribution (a) are obviously


much more homogeneous than those in distribution (b). There is clearly a need for indexes that can describe distributions in terms of *variation, spread, dispersion, heterogeneity*, or *scatter* of scores. Three indexes are commonly used for this purpose: range, variance, and standard deviation.

## RANGE

The simplest of all indexes of variability is the **range.** It is the difference between the upper real limit of the highest score and the lower real limit of the lowest score. In statistics, any score is thought of as representing an interval width from halfway between that score and the next lowest score (lower real limit) up to halfway between that score and the next highest score (upper real limit). For example, if several children have a recorded score of 12 pull-ups on a physical fitness test, their performances probably range from those who just barely got their chin over the bar the twelfth time and were finished (lower real limit) to those who completed 12 pull-ups, came up again, and almost got their chin over the bar, but did not quite make it for pull-up 13 (upper limit). Thus, a score of 12 is considered as representing an interval from halfway between 11 and 12 (11.5) to halfway between 12 and 13 (12.5) or an interval of 1. For example, given the following distribution of scores, you find the range by subtracting 1.5 (the lower limit of the lowest score) from 16.5 (the upper limit of the highest score), which is equal to 15. It is simpler to use Formula 6.3:

2 10 11 12 13 14 16

$$R = (X_h - X_l) + I \tag{6.3}$$

where

$R$ = range
$X_h$ = highest value in a distribution
$X_l$ = lowest value in a distribution
$I$ = interval width

Subtract the lower number from the higher and add 1 ($16 - 2 + 1 = 15$). In frequency distribution, 1 is the most common interval width.

The range is an unreliable index of variability because it is based on only two values, the highest and the lowest. It is not a stable indicator of the spread of the scores. For this reason, the use of the range is mainly limited to inspectional purposes. Some research reports refer to the range of distributions, but such references are usually used in conjunction with other measures of variability, such as variance and standard deviation.

### THINK ABOUT IT 6.4

1. a. What is the range of Mr. Li's physics exam scores?
   b. What is the range of the Ed 101 scores?

**Answers**

1. a. $23.5 - 15.5 = 8$ or, using Formula 6.3, $X_h - X_l + I = 23 - 16 + 1 = 8$.
   b. $36.5 - 20.5 = 16$ or $X_h - X_l + I = 36 - 21 + 1 = 15 + 1 = 16$. (Note that the highest occurring score was 36; the lowest occurring score was 21.)

## VARIANCE AND STANDARD DEVIATION

Variance and standard deviation are the most frequently used indexes of variability. They are both based on **deviation scores**—scores that show the difference between a raw score and the mean of the distribution. The formula for a deviation score is

$$x = X - \overline{X} \tag{6.4}$$

where

$x$ = deviation score
$X$ = raw score
$\overline{X}$ = mean

Scores below the mean will have negative deviation scores, and scores above the mean will have positive deviation scores. For example, the mean in Mr. Li's physics exam is 20; thus, Ona's deviation score is $x = 22 - 20 = 2$, whereas Ted's deviation score is $16 - 20 = -4$. By definition, the sum of the deviation scores in a distribution is always 0. Thus, to use deviation scores in calculating measures of variability, you must find a way to get around the fact that $\Sigma x = 0$. The technique used is to square each deviation score so that they all become positive numbers. If you then sum the squared deviations and divide by the number of scores, you have the mean of the squared deviations from the mean, or the **variance.** In mathematical form, variance is

$$\sigma^2 = \frac{\Sigma x^2}{N} \tag{6.5}$$

where

$\sigma^2$ = variance
$\Sigma$ = sum of
$x^2$ = deviation of each score from the mean $(X - \overline{X})$ squared, otherwise known as the deviation score squared
$N$ = number of cases in the distribution

In column 4 of Table 6.3, we see the deviation scores, differences between each score, and the mean. Column 5 shows each deviation score squared ($x^2$), and column 6 shows the frequency of each score from column 2 multiplied by $x^2$. Summing column 6 gives us the sum of the squared deviation scores $\Sigma x^2 = 72$. Dividing this by the number of scores gives us the mean of the squared deviation scores, the variance.

$$\sigma^2 = \frac{\Sigma x^2}{N} = \frac{72}{18} = 4$$

The forgoing procedure is convenient only when the mean is a whole number. This rarely occurs except in textbook examples. We have chosen to do our examples with whole number means so you can understand the concept and not get bogged down with the mathematics.

Formula 6.6 avoids the tedious task of working with squared mixed-number deviation scores such as $7.6667^2$. Using Formula 6.6 yields the desired result with much less labor. Thus, we recommend that students always use this formula for computing standard deviation if the computation must be done "by hand":

$$\sigma^2 = \frac{\Sigma X^2 - \dfrac{(\Sigma X)^2}{N}}{N} \tag{6.6}$$

where

$\sigma^2$ = variance
$\Sigma X^2$ = sum of the squares of each score (i.e., each score is first squared, and then these squares are summed)
$(\Sigma X)^2$ = sum of the scores squared (the scores are first summed, and then this total is squared)
$N$ = number of cases

**Table 6.3** Variance of Mr. Li's Physics Exam Scores

| (1) $X$ | (2) $f$ | (3) $fX$ | (4) $x$ | (5) $x^2$ | (6) $fx^2$ | (7) $X^2$ | (8) $fX^2$ |
|---|---|---|---|---|---|---|---|
| 23 | 2 | 46 | +3 | 9 | 18 | 529 | 1058 |
| 22 | 2 | 44 | +2 | 4 | 8 | 484 | 968 |
| 21 | 4 | 84 | +1 | 1 | 4 | 441 | 1764 |
| 20 | 4 | 80 | 0 | 0 | 0 | 400 | 1600 |
| 19 | 2 | 38 | −1 | 1 | 2 | 361 | 722 |
| 18 | 2 | 36 | −2 | 4 | 8 | 324 | 648 |
| 17 | 0 | 0 | | | | | |
| 16 | 2 | 32 | −4 | 16 | 32 | 256 | 512 |
| | $N = 18$ | $\Sigma X = 360$ | | | $\Sigma x^2 = 72$ | | $\Sigma X^2 = 7272$ |

Column 7 in Table 6.3 shows the square of the raw scores. Column 8 shows these raw score squares multiplied by frequency. Summing this $fX^2$ column gives us the sum of the squared raw scores:

$$\sigma^2 = \frac{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}{N} = \frac{7272 - \frac{360^2}{18}}{18} = \frac{7272 - \frac{129600}{18}}{18} = \frac{7272 - 7200}{18} = \frac{72}{18} = 4$$

Note that this result is the same as that which we obtained with Formula 6.5.

Because each of the deviation scores is squared, the variance is necessarily expressed in units that are squares of the original units of measure. For example, you might find that the variance of the heights of children in a class is 9 square inches. This would tell you that this class is more heterogeneous in height than a class with a variance of 4 square inches and more homogeneous than a class with a variance of 16 square inches.

In most cases, educators prefer an index that summarizes the data in the same unit of measurement as the original data. **Standard deviation** ($\sigma$), the positive square root of variance, provides such an index. By definition, the standard deviation is the square root of the mean of the squared deviation scores. Rewriting this definition using symbols, you obtain

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} \tag{6.7}$$

For Mr. Li's physics exam scores, the standard deviation is

$$\sqrt{\frac{72}{18}} = \sqrt{4} = 2$$

The standard deviation belongs to the same statistical family as the mean; that is, like the mean, it is an interval or ratio statistic, and its computation is based on the size of individual scores in the distribution. It is by far the most frequently used measure of variability and is used in conjunction with the mean.

Formulas 6.5, 6.6, and 6.7 are appropriate for calculating the variance and the standard deviation of a population. If scores from a finite group or sample are used to estimate the heterogeneity of a population from which that group was drawn, research has shown that these formulas more often underestimate the population variance and standard deviation than overestimate them. Mathematically, to get unbiased estimates, $N-1$ rather than $N$ is used as the denominator.

The formulas for variance and standard deviation based on sample information are

$$s^2 = \frac{\Sigma x^2}{N-1} \tag{6.8}$$

$$s = \sqrt{\frac{\Sigma x^2}{N-1}} \tag{6.9}$$

$$s = \sqrt{\frac{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}{N-1}} \tag{6.10}$$

Following the general custom of using Greek letters for population parameters and Roman letters for sample statistics, the symbols for variance and standard deviation calculated with $N-1$ are $s^2$ and $s$, respectively.

With the data in Table 6.3,

$$\frac{72}{18-1} = 4.24 \quad \text{and} \quad s = \sqrt{4.24} = 2.06$$

Formulas 6.8, 6.9, and 6.10 are often used to calculate variance and standard deviation even when there is no intention to estimate population parameters. Many computer and calculator programs calculate variance and standard deviation this way unless instructed to do otherwise.

Spread, scatter, heterogeneity, dispersion, and volatility are measured by standard deviation, in the same way that volume is measured by bushels and distance is measured by miles. A class with a standard deviation of 1.8 on reading grade level is more heterogeneous than a class with a standard deviation of 0.7. A month when the daily Dow Jones Industrial Average has a standard deviation of 40 is more volatile than a month with a standard deviation of 25. A school where the teachers' monthly salary has a standard deviation of $900 has more disparity than a school where the standard deviation is $500.

## MEASURES OF RELATIVE POSITION

Measures of relative position indicate where a score falls in relation to all other scores in the distribution. Researchers often want to assess an individual's relative position in a group or to compare the relative position of one individual on two or more measures or of two or more individuals on the same measure. The most widely used statistics for these purposes are *z* scores, stanines, other standard scores, and percentile rank.

### Z SCORE

The most widely used measure of relative position is the ***z* score,** which indicates the positive or negative difference between an individual score and the mean as measured in standard deviation units. It and other indexes derived from it are known as **standard scores.** The *z* score is defined as the distance of a score from the mean as measured by standard deviation units. The formula for finding a *z* score is

$$z = \frac{x}{\sigma} = \frac{X - \overline{X}}{\sigma} \tag{6.11}$$

where

$X$ = raw score
$\overline{X}$ = mean of the distribution
$\sigma$ = standard deviation of the distribution
$x$ = deviation score $(X - \overline{X})$

Applying this formula, a score exactly one standard deviation above the mean becomes a *z* of +1, a score exactly one standard deviation below the mean becomes a *z* of −1, and so on. A score equal to the mean will have a *z* score value of 0. For illustration, suppose a student's score on a psychology test is 72, where the mean of the distribution is 78 and the standard deviation equals 12. Suppose also that the same student has made a score of 48 on a statistics test, where the mean is 51

and the standard deviation is 6. If you substitute these figures for the appropriate symbols in Formula 6.11, you can derive a $z$ score for each test:

Psychology $$z_1 = \frac{72 - 78}{12} = -0.50$$

Statistics $$z_2 = \frac{48 - 51}{6} = -0.50$$

Both these standard scores belong to the $z$ distribution, where by definition the mean is always 0 and the standard deviation is 1, and therefore they are directly comparable. It is apparent in this example that the score of 72 on the psychology test and the score of 48 on the statistics test are equivalent—that is, both scores indicate the same relative level of performance. In other words, the standing of this student is the same in both tests when compared with the performance of the other students. It would be very difficult to make such a comparison without employing the $z$ score technique.

Let us use another example: Suppose a student who has taken the same tests has obtained a score of 81 on the psychology test and a score of 53 on the statistics test. As before, it is difficult to compare these raw scores to show on which test this student has done better. Converting the scores to $z$ scores makes the comparison easy. Using Formula 6.11, we find the values of $z_1$ and $z_2$ in this case to be as follows:

Psychology $$z_1 = \frac{81 - 78}{12} = +0.25$$

Statistics $$z_2 = \frac{53 - 51}{6} = +0.33$$

This result shows that the score of 53 on the statistics test actually indicates a slightly better relative performance than the score of 81 on the psychology test. Compared with the other students, this student has done somewhat better in statistics than in psychology.

Because the mean of the $z$ scores in any distribution is 0 and the standard deviation is 1, they have universal meaning. A $z$ score of $-0.1$ is slightly below average in a distribution of statistics test scores, a distribution of weights of people in a weight control program, a distribution of pork belly prices, or any other distribution. A $z$ score of $+2.4$ is very high, whether you are talking about achievement scores, scores on a measure of depression, corn yield per acre, or any other measure.

## OTHER STANDARD SCORES

Scores can also be transformed into other standard score scales that do not involve negative numbers or decimals. One of the most common procedures is to convert to ***T* scores** by multiplying the $z$ scores by 10 and adding 50. This results in a scale of positive whole numbers that has a mean of 50 and a standard deviation of 10. The $T$ score formula is

$$T = 10(z) + 50 = 10\left(\frac{X - \overline{X}}{\sigma}\right) + 50 \tag{6.12}$$

A score of 21 on a test for which the mean of the scores is 27 and the standard deviation is 6 would have a $z$ score of −1.00 or a $T$ score of 40:

$$T = 10\left(\frac{21 - 27}{6}\right) + 50 = 40 \tag{6.13}$$

The transformation of $z$ scores into $T$ scores not only enables one to work with whole numbers but also avoids the adverse psychological implications of describing subjects' performances with negative numbers. In the preceding example, it would be easier to report that the student's score is 40 where the mean score is 50 than to report a score of −1.00 with an average of zero.

In addition to $T$, there are other transformed standard score distributions. To transform a distribution of scores to a new standardized distribution, multiply the $z$ score by the desired standard deviation and add the desired mean. The general formula is as follows:

$$A = \sigma_A(z) + \mu_A \tag{6.14}$$

where

$A$ = standard score on the new scale
$\mu_A$ = mean for the new standard scale
$\sigma_A$ = standard deviation for the new standard scale

For example, College Entrance Examination Board (CEEB) scores have a mean of 500 and a standard deviation of 100 for its transformed distribution. If you were 1.5 standard deviations above the mean ($z = 1.5$) on the verbal section of the Scholastic Assessment Test (SAT), your score would be reported as 650, which is 500 + (100)(1.5). If your quantitative score were 500, you would have scored exactly at the mean.

The Wechsler Adult Intelligence Test scores are standard scores with a mean of 100 and a standard deviation of 15. A raw score on the mean is reported as 100. A raw score one standard deviation below the mean is reported as 85. A raw score two standard deviations above the mean is reported as 130.

Transforming a set of scores to standard scores does not alter the shape of the original distribution. If a distribution of scores is skewed, the derived standard scores also produce a skewed distribution. Only if the original distribution is normal do the standard scores produce a **normal distribution.**

## STANINE SCORES

During World War II, the U.S. Army Air Corps developed a standard system of nine scores called **stanine scores** to help its personnel interpret $z$ scores. Stanines avoid negative numbers and decimals. A stanine score of 5 represents $z$ scores that are average or slightly above or slightly below average—that is, equivalent to $z$ scores between −0.25 and +0.25. From there, stanine scores go up to 9 and down to 1 in increments of 0.5, as shown in Table 6.4. Stanines are standardized with the mean of 5 and a standard deviation of 2. The formula for stanines is $2z + 5$. You convert a $z$ score to a stanine by multiplying by 2 and adding 5. Stanines are always rounded to the nearest whole number. Whenever this formula yields a result greater than 9, the value 9 is assigned. Whenever the result is less than 1, the value 1 is assigned. Because all $z$ scores above 1.75 are assigned a stanine score of 9 and all $z$ scores below −1.75 are assigned a score of 1, stanine scores are not useful for comparing

**Table 6.4** Conversion of *z* Scores to Stanines

| *z* Score | Stanine | Interpretation | Percent in Stanine |
|---|---|---|---|
| Above +1.75 | 9 | Among the very highest scores | 4 |
| +1.25 to +1.75 | 8 | Quite well above average | 7 |
| +0.75 to +1.25 | 7 | Quite noticeably above average | 12 |
| +0.25 to +0.75 | 6 | Above average | 17 |
| −0.25 to +0.25 | 5 | Near dead average | 20 |
| −0.75 to −0.25 | 4 | Below average | 17 |
| −1.25 to −0.75 | 3 | Quite noticeably below average | 12 |
| −1.75 to −1.25 | 2 | Quite well below average | 7 |
| Below −1.75 | 1 | Among the lowest scores | 4 |

extreme scores. Stanines are easy to comprehend. Like all transformations of the *z* score, they have universal meaning. A stanine score of 4 always means below average but not too far below average. Stanines are often used in school systems for reporting students' standardized test scores.

### THINK ABOUT IT 6.5

Recall the scores in Mr. Li's physics class in Think About It 6.2.

1. What is the *z* score of a raw score of 21 on Mr. Li's exam?
2. What is the *z* score of a raw score of 18 on Mr. Li's exam?
3. What is the *z* score of a raw score of 20 on Mr. Li's exam?
4. What is the stanine for each of these scores?

**Answers**

1. +0.5; 2. −1.0; 3. 0

4. Stanines: 1. 6; 2. 3; 3. 5

## PERCENTILE RANK

A measure of relative position that most people find easy to understand and interpret is the percentile rank. A **percentile rank** (PR) indicates the percentage of scores in a distribution that fall below a given score point. It is easy to picture a score with a PR of 32 as having 32 percent of the scores in its distribution below it and a score with a PR of 89 as having 89 percent of the scores below it.

The following is a simple way to calculate PR:

1. Arrange the scores in a frequency distribution.
2. Determine the cumulative frequency of scores below the interval containing the score of interest and add one half of the frequency of scores within the interval containing the score of interest.
3. Divide this sum by the total number of scores and multiply by 100.

Consider Mr. Li's physics exam scores (see Table 6.2). To calculate the PR of 21, we start with the cumulative frequency below a score of 21 ($cf = 10$) and add one half of the number scoring 21 ($f_w = 2$). We then divide this number by the total number of students who took the test ($N = 18$). Finally, we multiply the result by 100 and round to the nearest whole number:

$$\text{PR} = \frac{cf_b + \frac{f_w}{2}}{N}(100) \tag{6.15}$$

which is rounded to 67.

A score of 18 is assigned the following PR:

$$\text{PR} = \frac{cf_b + \frac{f_w}{2}}{N}(100) = \frac{2 + \frac{2}{2}}{18}(100) = \frac{3}{18}(100) = 16.67$$

which is rounded to 17.

The major advantages of percentile ranks are as follows:

1. They have universal meaning. A score with a percentile rank of 89 is high in any distribution. A score with a percentile rank of 32 is somewhat low in any distribution.
2. The familiar concept of 0 to 100 percent applies to the interpretation of percentile rank. Schools often report percentile ranks to parents.

The major disadvantages of percentile rank are as follows:

1. As with other ordinal statistics, percentile ranks cannot be added, subtracted, multiplied, or divided.
2. As with all ordinal indexes, equal differences between percentile ranks do not represent equal differences between the scores in the original distribution. If there are many scores near a particular score, a small change in score will produce a major change in percentile rank. If there are few scores near a particular score, a considerable change in raw score will be necessary to produce a change in position relative to other scores and thus a change in percentile rank. For example, a professor has recorded the weights of the students in his Physical Education 202 class and used them to illustrate the computation of percentile rank. The result is the polygon shown in Figure 6.8.

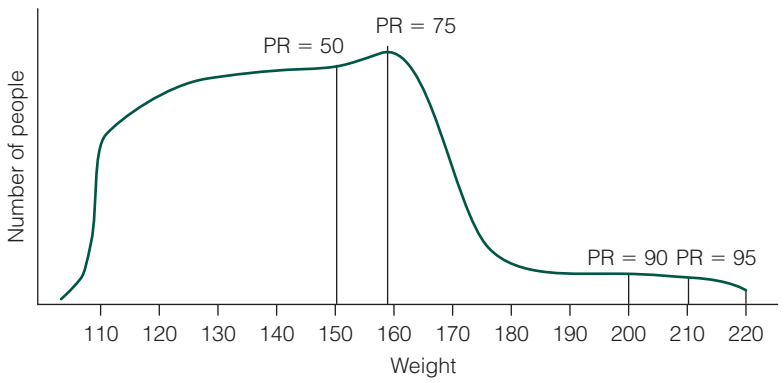



**Figure 6.8** Weights of Students in Physical Education 202 Class

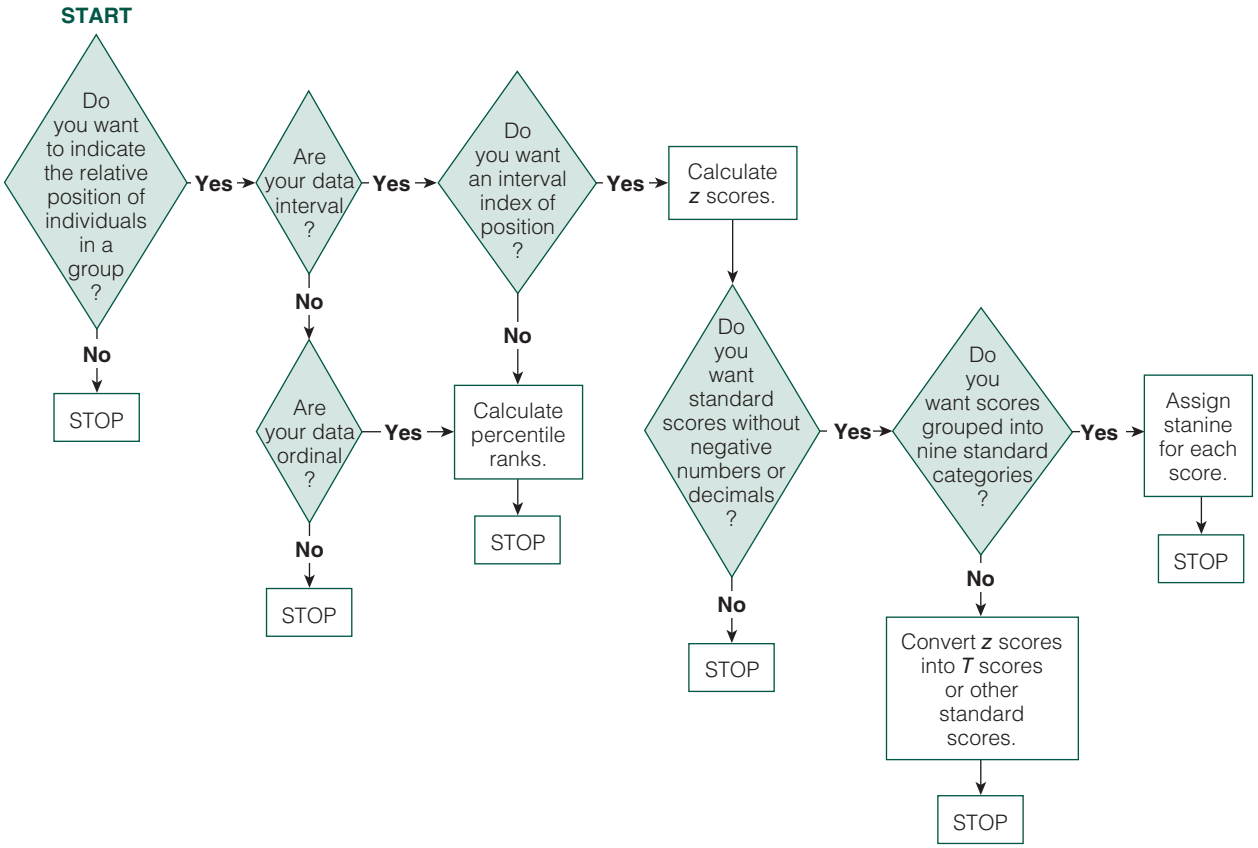



**Figure 6.9** Measures of Relative Position

A 160-pound student and a 210-pound student both resolve to lose weight and actually lose 10 pounds each. The 10-pound loss moves the 160-pound student from a percentile rank of 75 to a percentile rank of 50. The same weight loss only changes the heavier student's percentile rank from 95 to 90. Often, many cases concentrate near the middle of a distribution and then taper off with few cases occurring at the extreme ends. In such distributions, minor differences in raw scores will appear as major differences in percentile ranks among those scores that are near the center of the distribution, where a large number of the scores typically are located. At the extreme ends of the distribution, where there are few scores, major differences in raw score will have only minor effects on percentile rank. We discuss the phenomenon more closely when we consider the normal curve.

Figure 6.9 shows the process of deciding which index to choose for indicating relative position.

## THE NORMAL CURVE

Recall the example of deductive reasoning in Chapter 1, where we concluded that if the probability of having a son is 50 percent and the probability of having a daughter is 50 percent, the probability of two daughters is 25 percent, the probability of two sons is 25 percent, and the probability of one of each is

50 percent. Abraham DeMoivre (1667–1754) pondered the probabilities of various outcomes when the percent of likelihood in each trial is 50/50, as in heads and tails in honest coin flips. He came up with a formula to predict the probabilities of various number of heads (or tails) when a coin is flipped an infinite number of times. The most frequent score is half infinity, the next most frequent scores are half infinity plus one and half infinity minus one, and so forth. When a polygon of the expected proportions of various *z* scores is made, the outcome is a bell-shaped curve called the **normal curve.** This model proved very useful for gamblers interested in predicting the probability of various gaming outcomes.

Soon after the normal curve was developed, it was noticed that many naturally occurring distributions, such as height, weight, and IQ, formed polygons resembling the normal curve. If you measure U.S. boys on their tenth birthday, you will find many whose height is near the mean and slightly fewer boys who are a bit above or a bit below the mean. The further you get from the mean, the fewer boys you will find at each height. As in the normal curve probability model, *z* scores near 0 will be expected to occur more frequently than other *z* score values, and the farther from 0 a *z* score is, the less frequently it will be expected to occur.

Inasmuch as so many naturally occurring distributions resemble the normal curve, this theoretical model has proved very useful in research and other endeavors. Whenever actual data are known or believed to resemble the normal curve in distribution, you can deduce many useful estimates from the theoretical properties of the normal curve.

Note in Figure 6.10 that the normal curve is a symmetrical distribution with the same number of cases at specified *z* score distances below the mean as above the mean. Its mean is the point below which exactly 50 percent of the cases fall and above which the other 50 percent of the cases are located. Since the curve is symmetrical, the mean, the median, and the mode are identical. In a normal distribution, most of the cases concentrate near the mean. The frequency of cases decreases as you proceed away from the mean in either direction. Approximately 34 percent of the cases in a normal distribution fall between the mean and one standard deviation above the mean, and approximately 34 percent are between the mean and one standard deviation below the mean. Between one and two standard deviations from the mean on either side of the distribution are approximately 14 percent of the cases. Only approximately 2 percent of the cases fall between two and three standard deviations from the mean, and only approximately one-tenth of 1 percent of the cases fall above or below three standard deviations from the mean.

These characteristics can be seen in Figure 6.10. You can see visually that approximately one-sixth of the curve falls to the left of one standard deviation below the mean. The first line under the curve shows standard deviations from $-4$ to $+4$. These are equivalent to *z* scores from $-4.00$ to $+4.00$. The cumulative percentage line tells you that 15.9 percent of scores fall below $-1$ and 97.7 percent falls below $+2$, and so on. The line following cumulative percentage shows these cumulative percentage scores rounded to the nearest whole percentage. Multiplying each of these numbers by 100 gives you percentile rank.

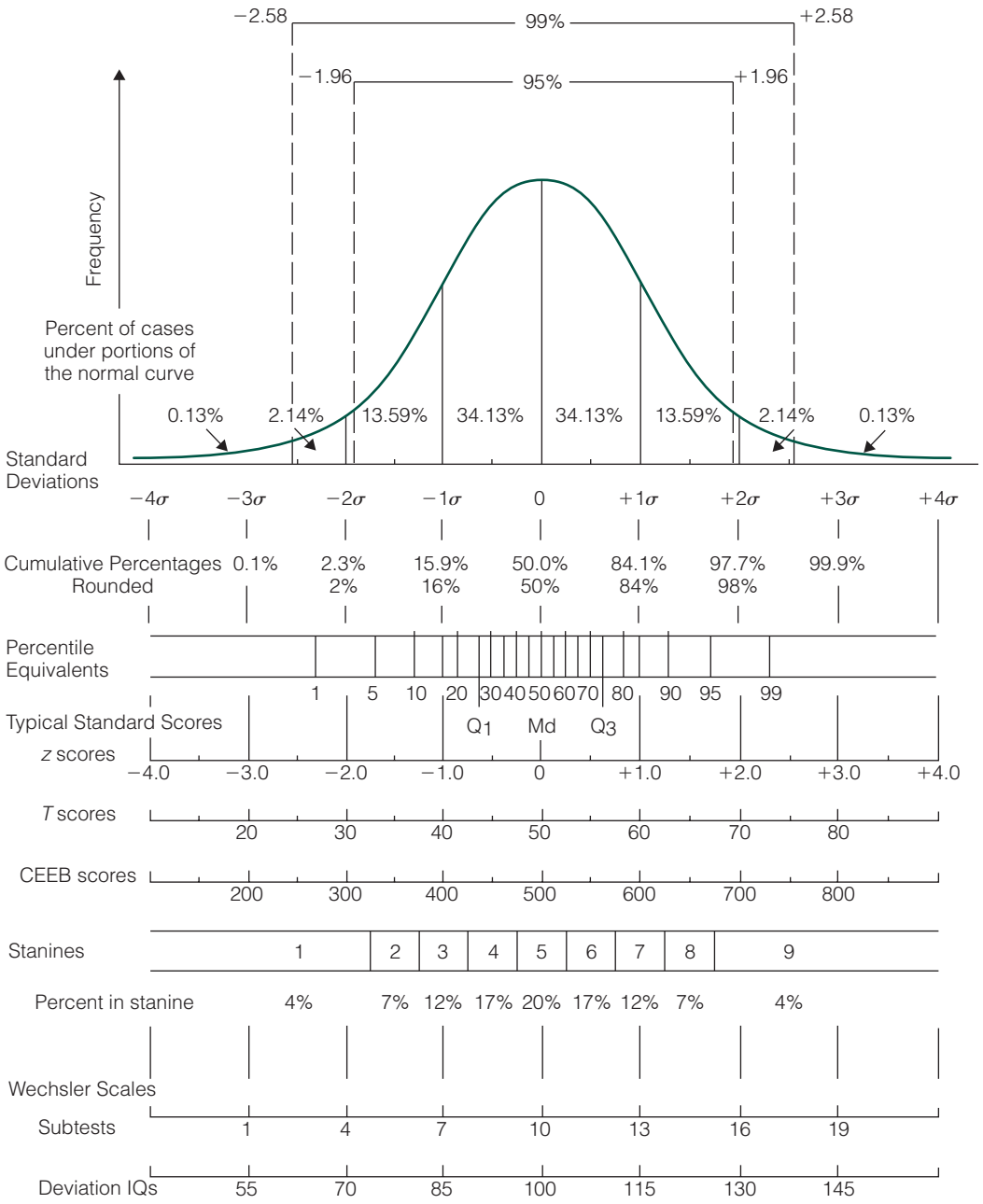



**Figure 6.10** Normal Curve and Equivalents

*Source: Test Service Bulletin No. 48,* Jan. 1955. Courtesy of The Psychological Corporation.

Percentile equivalents are shown on the next line, which also shows the first quartile ($Q_1$ sets off the lowest 25% of scores), the median (Md), and the third quartile ($Q_3$ sets off the lower 75% or upper 25% of scores). Note how slowly the percentile equivalents change below $Q_1$ and above $Q_3$ and how rapidly they change between these two points. The next line after percentile equivalents

shows $z$ scores, which are identical to the scores on the standard deviation line. Following the $z$ score line are various standard scores transformed from $z$ scores, including $T$ scores, CEEB scores, stanines, percent in stanine, Wechsler subtest scores, and Weschsler deviation IQs. Note that 95 percent of the normal curve falls between plus and minus $z = 1.96$ and 99 percent falls between plus and minus $z = 2.58$. These boundaries become important when we discuss the use of the normal curve in inferential statistics. To determine the exact percentage of the cases below and above each $z$ score in the normal distribution, consult Table A.1 in the Appendix, which gives the areas of the normal curve. Column 1 of Table A.1 contains different $z$ values. Column 2 gives the area under the curve between the mean and each $z$ value. Column 3 shows the remaining area from each $z$ score to the end of the curve. Therefore, the areas in column 2 and column 3 add up to .5000. Take as an example the $z$ value of +0.70. The area between this $z$ value and the mean can be found in column 2; it is .2580. This figure indicates that 26 percent of the cases fall between this $z$ value and the mean of the distribution. Because the mean of the normal distribution coincides with the median, 50 percent of the cases lie below the mean. Add 0.50 to the .2580, and the result tells you that you can expect 75.8 percent of the cases to fall below the $z$ value of +0.70. Column 3 indicates that the other 24.2 percent of the cases fall above the $z$ value of +0.70.

This procedure is reversed when the $z$ value is negative. Suppose you want to find the percentage of cases below the $z$ value of −0.70. The area between the mean and a $z$ score of −0.70 is .2580 or, in terms of percentage, 25.8 percent of the cases. Subtracting 25.8 from 50, you obtain 24.2. This result would indicate that only 24.2 percent of the scores lie below a $z$ value of −0.70 in a normal distribution. This value can also be found in column 3 of the table, which gives a value of .2420 for a $z$ score of 0.70. The percentage of scores above −0.70 is 100 − 24.2, or 75.8 percent. Since the normal curve is absolutely symmetrical, we do not need separate tables for positive and negative $z$ scores. You just have to remember the sign of the $z$ score with which you are working.

Among other applications, the normal curve can be used to help people who are unfamiliar with standard scores to interpret them. For example, how high is a score of 650 on the SAT? The SAT has a mean of 500 and a standard deviation of 100, so the $z$ score here is 1.5. Consulting Table A.1, column 2, you find .4332 of the normal curve falls between the mean and $z = 1.5$. Adding the 50 percent below the mean, you can say that an SAT score of 600 exceeds the scores of 93 percent of SAT scores.

Because it is known that the population distribution of SAT scores closely resembles the normal curve, PR approximations based on the normal curve will be quite near to the actual PRs. With other scores, as actual distributions become increasingly less like the normal curve, the PR approximations become increasingly less on target. Where the shape of a distribution is not known, it is usually reasonable to assume a distribution similar to the normal curve and to use the normal curve table to find reasonable approximations of the PRs of various $z$ scores. The more the actual shape differs from the normal, the less useful the approximations become.

The most common use of the normal curve in descriptive statistics is going from a given $z$ score to a percentile rank as described in the previous paragraph,

but we can also use it to go in the opposite direction, from a given percentile rank to its $z$ score equivalent.

## CORRELATION

After completing the second unit in physics class, Mr. Li gave a second exam. Table 6.5 lists his students in column 1. Their $z$ scores on test 1 are shown in column 2, and their $z$ scores on test 2 are shown in column 3. Recall that $z$ scores are a way to indicate the relative positions of scores. They have universal meaning and can be used with any interval or ratio data.

Looking at Table 6.5, you can see that there is a tendency for those who had positive $z$ scores on test 1 to have positive $z$ scores on test 2 and for those with negative $z$ scores on test 1 to have negative $z$ scores on test 2. Four students have identical $z$ scores on both tests. The others have $z$ scores with the same sign but different values, except Ali, who had a positive $z$ score on test 1 and a negative $z$ score on test 2.

Figure 6.11 shows a histogram of the first test scores on the abscissa ($x$) and a histogram of the second test scores turned sideways on the ordinate ($y$). In the upper right part of Figure 6.11 you find each student's position on both the first and the second test. This gives a picture of how the students tend to have similar $z$ scores on the two tests, but there is some shifting of their relative positions. There is a strong but not perfect positive relationship between the relative positions of each student's scores on the two tests.

**Table 6.5** Mr. Li's First and Second Test $z$ Scores

| | Test 1 $z$ Scores | Test 2 $z$ Scores | $z$ Score Products |
|---|---|---|---|
| Student | $z_x$ | $z_y$ | $(z_x z_y)$ |
| Ali | +0.5 | −1 | −0.5 |
| Ann | 0 | 0 | 0 |
| Ben | +1.5 | +1 | 1.5 |
| Cal | 0 | −1 | 0 |
| Dan | 0 | +0.5 | 0 |
| Ed | +0.5 | +0.5 | 0.25 |
| Ima | +1.0 | +1.5 | 1.5 |
| Jan | −0.5 | 0 | 0 |
| Kay | −2.0 | −1.5 | 3.0 |
| Lee | 0 | −1 | 0 |
| Mel | −1 | −0.5 | 0.5 |
| Mia | +1.5 | +1.0 | 1.5 |
| Ned | +0.5 | +1.5 | 0.75 |
| Ona | +0.5 | +1 | 0.5 |
| Sam | +1.0 | +1 | 1.0 |
| Sue | −0.5 | −0.5 | 0.25 |
| Ted | −2.0 | −1.0 | 2 |
| Van | −1.0 | −1.5 | 1.5 |
| | | | $\Sigma = 13.75$ |

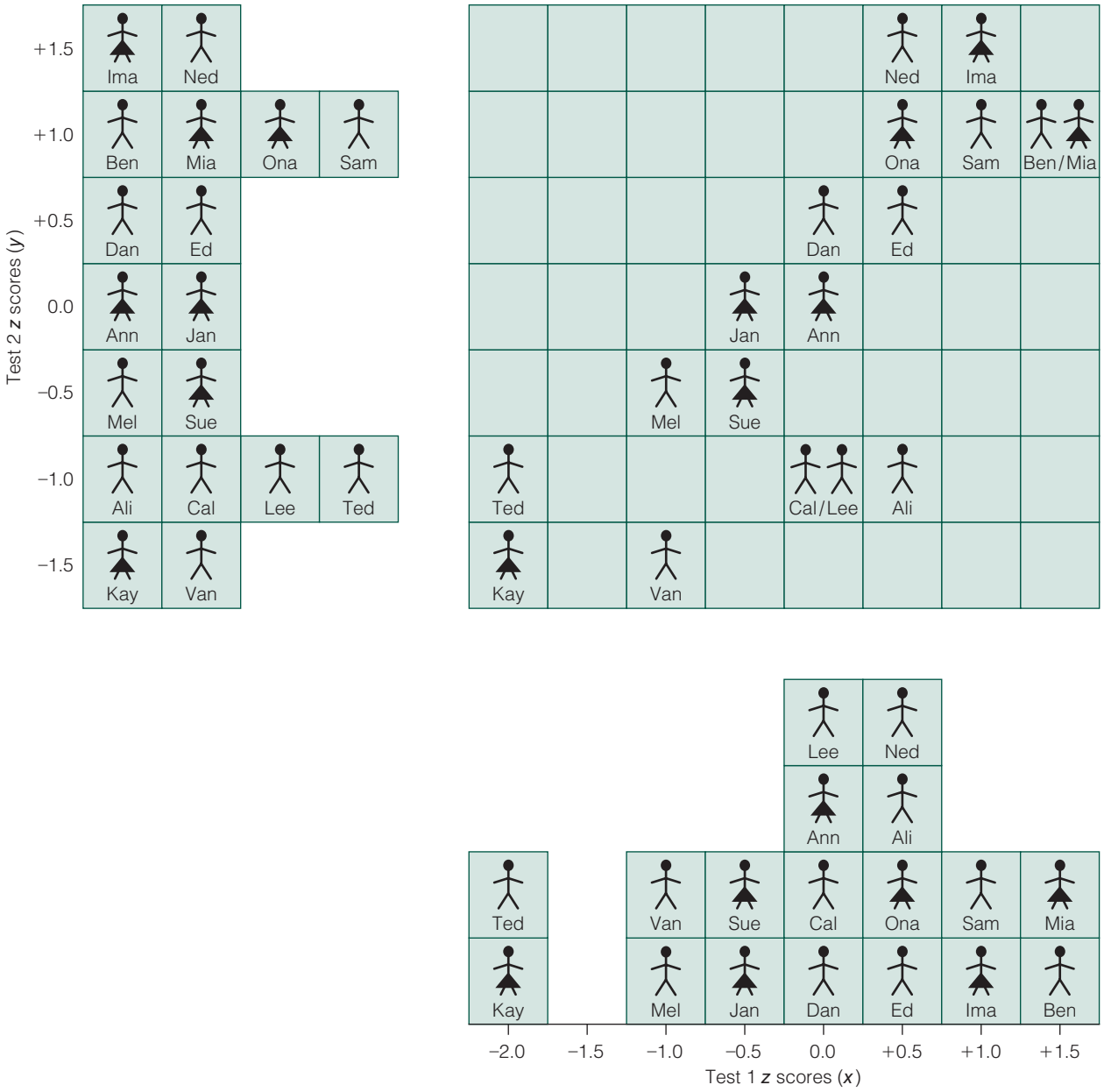



**Figure 6.11** Mr. Li's Students' *z* Scores on First and Second Exam Scores

Correlations indicate the relationship between paired scores. The correlation indicates whether the relationship between paired scores is positive or negative and the strength of this relationship. The pairs may be two scores for the same individual, natural pairs such as husbands and wives, or two individuals matched on some measure such as reading test scores. In addition to looking at correlation through visual means, the researcher can calculate a **correlation coefficient** that represents the correlation.

## PEARSON PRODUCT MOMENT CORRELATION COEFFICIENT

A very useful statistic, the **Pearson product moment correlation coefficient** (Pearson *r*), indicates both the direction and the magnitude of the relationship between two variables without needing a scatterplot to show it.

Start with the knowledge that because of the way *z* scores are defined, the sum of the *z* scores squared in any distribution will always equal the number in that distribution. (To convince yourself of this, you could square the *z* scores in column 1 or column 2, and find that the sum of the *z* scores squared in each column is 18.) Therefore, the mean of the squared deviations is always 1.

For example, if you measured precisely the square feet in each room in your building ($x$), and then precisely measured each room in square meters ($y$), the *z* scores would be identical and the Pearson *r* would be +1.0.

If you look at a schedule of Amtrak trains traveling between Chicago and Seattle, showing the miles from Seattle ($x$) and miles from Chicago ($y$), *z* scores of each station in terms of miles from Seattle ($x$) would be the exact opposite of *z* scores on $y$ (miles from Chicago). Each mile you get farther from Chicago, you get closer to Seattle. The Pearson *r* would be −1.0.

We have seen that the *z* scores on Mr. Li's two tests in Table 6.5 are similar but not perfectly aligned, so we know that the *z* score product averages for $x$ and $y$ will be less than +1.0 but approaching +1.0.

The definition of the Pearson *r* is simplicity itself. It is the mean $z_x z_y$ product:

$$r = \frac{\Sigma z_x z_y}{N} \tag{6.16}$$

where

$r$ = Pearson product moment coefficient of correlation
$\Sigma z_x z_y$ = sum of the *z* score products
$N$ = number of paired scores

The sum of the *z* score products in Table 6.5, column 4, is 13.75. Therefore, $r = 13.75/18 = .76$. This confirms what we had already concluded: The *z* scores on the two tests *are positively related, and this relationship is strong* but not perfect.

Whenever individuals tend to have *z* scores of the same sign but do not have exactly the same *z* score on *X* that they have on *Y*, the sum of the *z* scores will be positive but less than *N*. Therefore, the mean *z* score product, the Pearson *r*, will be less than +1. If positive $z_x$ scores tend to be paired with negative $z_y$ scores, but they are not perfect mirror images of each other, the sum of the $z_x z_y$ products will be a negative number, nearer to zero than to negative *N*. Therefore, the mean will be between −1 and 0. If there is a strong but not perfectly negative relationship, the *r* will be near to −1.00. If there is no overall relationship between the paired *z* scores, their product will be zero, and their mean will be zero.

Here we have an index that indicates not only the direction of relationships between variables but also the strength of the relationships. This index is never greater that +1.00 or less than −1.00. It has universal meaning.

Means and standard deviations that are whole numbers almost never occur except in textbooks. Typical *z* scores are awkward decimal or mixed numbers,

such as .4716 or 1.6667. Formula 6.16 is useful for understanding what the Pearson *r* means, but it is hopeless for calculation. Formula 6.17 avoids the need for calculating *z* scores and multiplying awkward decimals and mixed numbers. It also avoids rounding errors. Its result is the same as that of Formula 6.16:

$$r = \frac{\Sigma XY - \dfrac{(\Sigma X)(\Sigma Y)}{N}}{\sqrt{\left(\Sigma X^2 - \dfrac{(\Sigma X)^2}{N}\right)\left(\Sigma Y^2 - \dfrac{(\Sigma Y)^2}{N}\right)}} \tag{6.17}$$

where

$r$ = Pearson $r$
$\Sigma X$ = sum of scores in $X$ distribution
$\Sigma Y$ = sum of scores in $Y$ distribution
$\Sigma X^2$ = sum of the squared scores in $X$ distribution
$\Sigma Y^2$ = sum of the squared scores in $Y$ distribution
$\Sigma XY$ = sum of products of paired $X$ and $Y$ scores
$N$ = number of paired $X$ and $Y$ scores (subjects)

Table 6.6 provides the data needed to calculate the Pearson *r* for Mr. Li's tests 1 and 2 using Formula 6.17. Column 1 lists the students. Column 2 shows each student's raw score on test 1 ($X$), and column 3 shows these raw scores squared ($X^2$). Column 4 shows each student's raw score on test 2 ($Y$), and column 5 shows these scores squared ($Y^2$). Column 6 shows the product of each student's $X$ raw score multiplied by his or her $Y$ raw score ($XY$).

**Table 6.6** Mr. Li's Physics Class Raw Scores Illustrating the Calculation of the Pearson *r*

| (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|
| **Name** | $X$ | $X^2$ | $Y$ | $Y^2$ | $XY$ |
| Ali | 21 | 441 | 22 | 484 | 462 |
| Ann | 20 | 400 | 26 | 676 | 520 |
| Ben | 23 | 529 | 30 | 900 | 690 |
| Cal | 20 | 400 | 22 | 484 | 440 |
| Dan | 20 | 400 | 28 | 784 | 560 |
| Ed | 21 | 441 | 28 | 784 | 588 |
| Ima | 22 | 484 | 32 | 1024 | 704 |
| Jan | 19 | 361 | 26 | 676 | 494 |
| Kay | 16 | 256 | 20 | 400 | 320 |
| Lee | 20 | 400 | 22 | 484 | 440 |
| Mel | 18 | 824 | 24 | 576 | 432 |
| Mia | 23 | 529 | 30 | 900 | 690 |
| Ned | 21 | 361 | 32 | 1024 | 672 |
| Ona | 21 | 331 | 30 | 900 | 630 |
| Sam | 22 | 484 | 30 | 900 | 660 |
| Sue | 19 | 361 | 32 | 1024 | 672 |
| Ted | 16 | 256 | 22 | 484 | 352 |
| Van | 18 | 324 | 20 | 400 | 360 |
| | $\Sigma X = 360$ | $\Sigma X^2 = 7272$ | $\Sigma Y = 468$ | $\Sigma Y^2 = 12456$ | $\Sigma XY = 9470$ |

Using Formula 6.17, we get

$$r = \frac{9470 - \frac{(360)(468)}{18}}{\sqrt{\left(7272 - \frac{(360)^2}{18}\right)\left(12456 - \frac{(468)^2}{18}\right)}}$$

$$= \frac{9470 - \frac{168480}{18}}{\sqrt{\left(7272 - \frac{129600}{18}\right)\left(12456 - \frac{219024}{18}\right)}}$$

$$= \frac{9470 - 9360}{\sqrt{(7272 - 7200)(12456 - 12168)}}$$

$$= \frac{9470 - 9360}{\sqrt{(72)(288)}} = \frac{9470 - 9360}{\sqrt{20736}} = \frac{9470 - 9360}{144} = \frac{110}{144} = .76$$

## SCATTERPLOTS

The upper right part of Figure 6.11 shows people with their names to illustrate the concept that each figure represents an individual's *z* scores on both dimensions, test 1 and test 2. It is easier in practice to use a dot to represent individuals' pairs of scores and plot them on a graph called a **scatterplot,** as shown in Figure 6.12.

In a research situation, the *z* scores on the horizontal axis will be those of the independent variable, with the lowest *z* score on the left and the highest *z* score

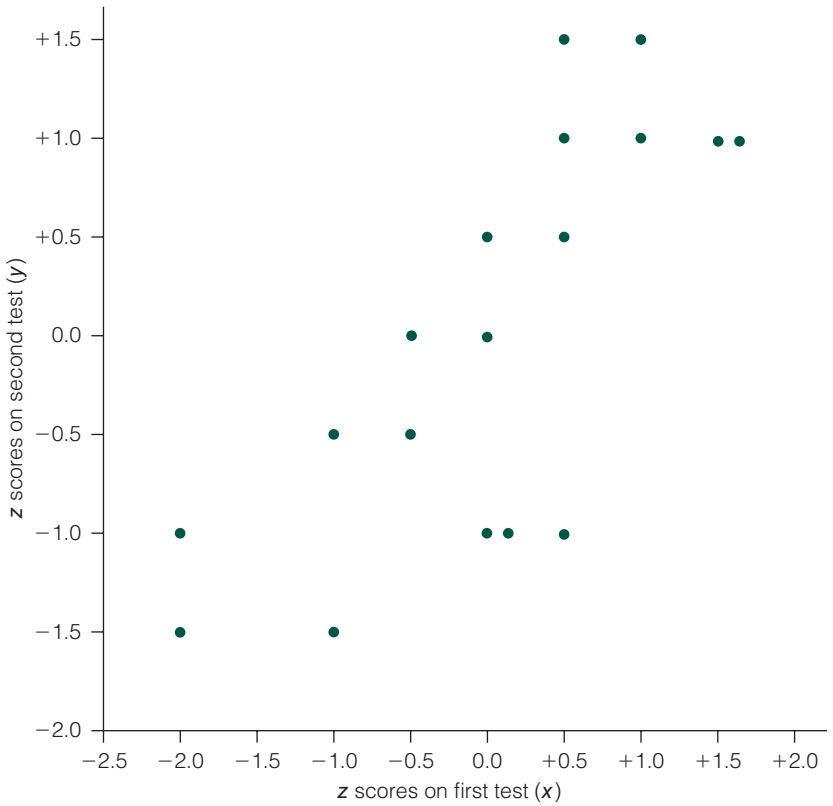



**Figure 6.12** Scatterplot of Mr. Li's Students' First and Second Test Scores

on the right. The $z$ scores on the vertical axis will be those of the dependent variable ($y$), with the lowest $z$ score at the bottom and the highest $z$ score at the top.

A scatterplot illustrates the direction of the relationship between the variables. A scatterplot with dots going from lower left to upper right indicates a **positive correlation** (as variable $x$ goes up, variable $y$ also goes up). One with dots going from upper left to lower right indicates a **negative correlation** (as variable $x$ goes up, variable $y$ goes down).

A scatterplot of $z$ scores also reveals the strength of the relationship between variables. If the dots in the scatterplot form a narrow band so that when a straight line is drawn through the band the dots will be near the line, there is a strong **linear relationship** between the variables. However, if the dots in the $z$ score scatterplot scatter widely, the relationship between variables is relatively weak. The scatterplots in Figure 6.13 show various positive and negative and strong and weak relationships, expressed mathematically by $r$.

You can see in comparing these scatterplots that the tilt of the "cloud" of dots becomes increasingly less as $r$ moves from a 45° angle for $r = +1.00$ or $r = -1.00$ until it reaches $r = 0$, where it is flat and matches the line for the mean of the $xy$ scores.

Once you get used to correlations, you will be able to picture a scatterplot for any Pearson correlation you encounter.

Like the mean and standard deviation, the Pearson $r$ is an interval statistic that can also be used with ratio data. An assumption underlying the product moment coefficient of correlation is that the relationship between the two variables ($X$ and $Y$) is linear—that is, that a straight line provides a reasonable expression of the relationship of one variable to the other. If a curved line is needed to express this relationship, it is said to be a **curvilinear relationship.** In a curvilinear relationship, as the values of $X$ increase, the values of $Y$ increase up to a point, at which further increases in $X$ are associated with decreases in $Y$. An example is the relationship between anxiety and performance. As individuals' anxiety level increases, so does their performance, but only up to a point. With further increases in anxiety, performance decreases. Another example is the amount of care people require during the course of a lifetime: It is high in the early and late years and usually relatively low in the middle years. A scatterplot of this relationship would produce a curve.

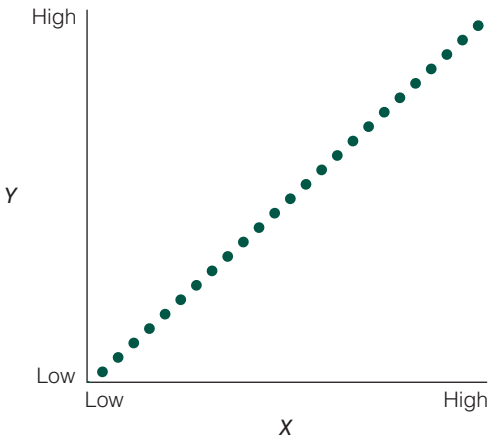
A practical way of determining whether the relationship between two variables is linear or curvilinear is to examine a scatterplot of the data. Figure 6.14 shows two diagrams, one of which (A) indicates a linear relationship and the other (B) a curvilinear one.

## THINK ABOUT IT 6.6

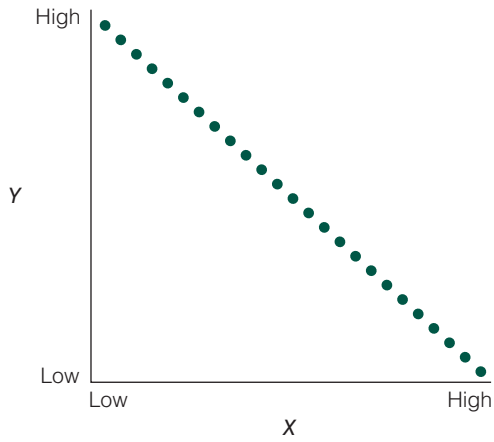
After scoring test 2 and entering these exam scores in his computer, Mr. Li also entered each student's number of days absent during the second unit. He instructed the computer to calculate the Pearson $r$ for these two variables (days absent and test 2 scores). The $r$ was $-.4$. What would he conclude?

**Answer**

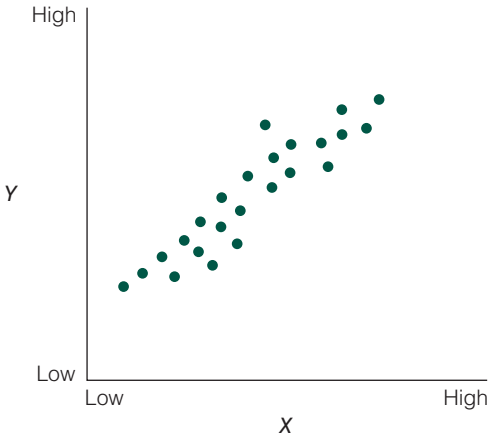
There was a moderate tendency for those with high days absent to have lower test 2 scores and those with low days absent to have higher scores on test 2.



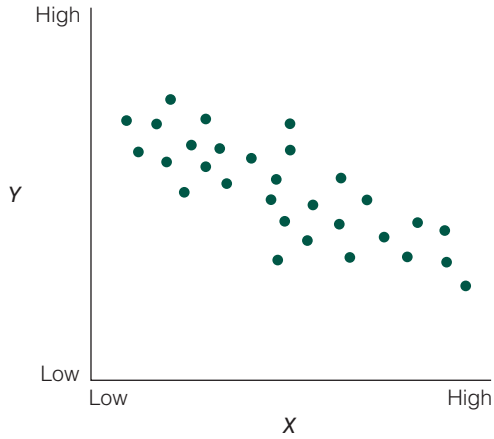

A. Perfect positive correlation (+1.00)



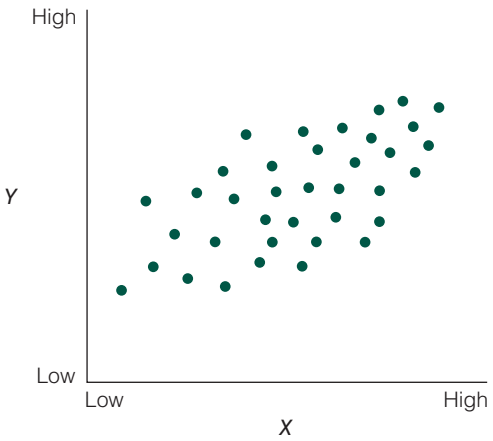

B. Perfect negative correlation (−1.00)



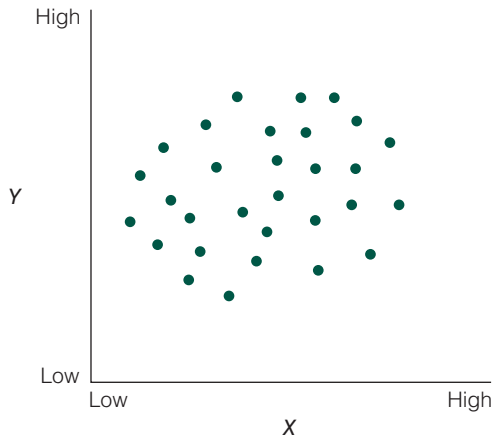

C. High positive correlation (+.93)




D. High negative correlation (−.76)




E. Moderate positive correlation (+.30)




F. Zero correlation

**Figure 6.13** Scatterplots of Selected Values of *r*

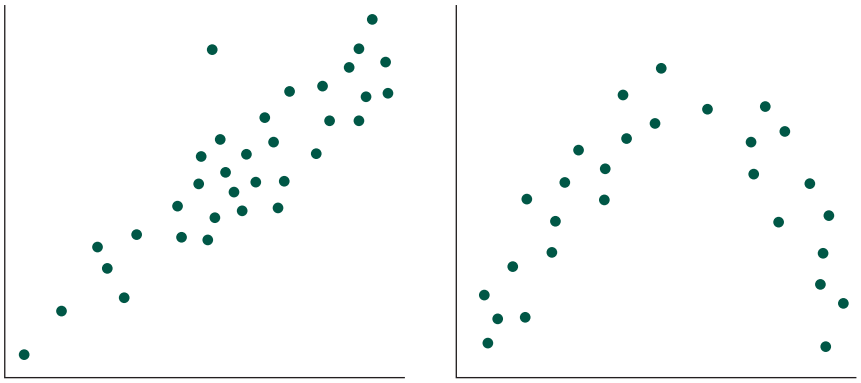



**Figure 6.14** Linear and Curvilinear Relationships

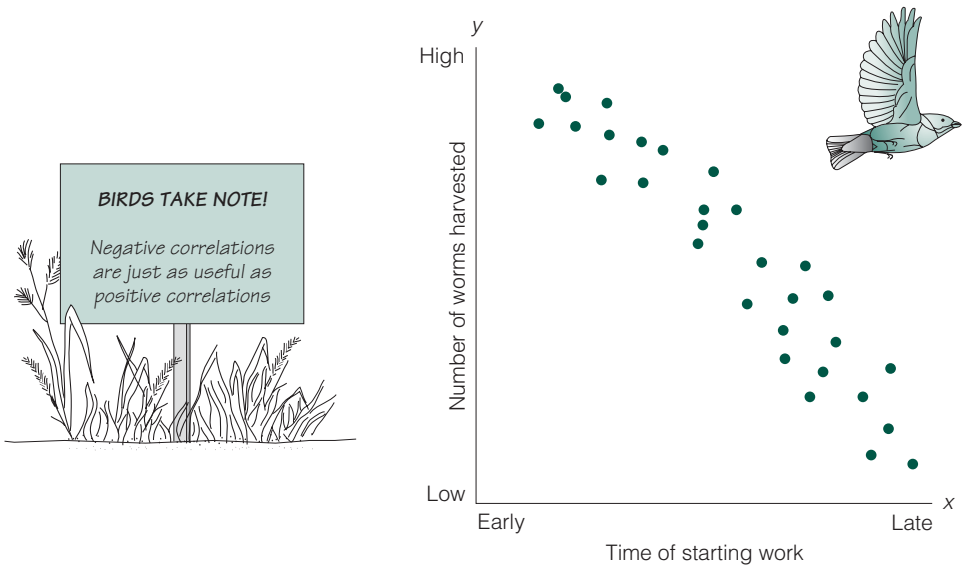



**Figure 6.15** A Useful Negative Correlation

If the relationship between variables is curvilinear, the computation of the Pearson *r* will result in a misleading underestimation of the degree of relationship. In this case, another index, such as the correlation ratio (Δ), should be applied. A discussion of the correlation ratio is presented in Chapter 13.

## THINK ABOUT IT 6.7

1. What is the best estimate of the Pearson *r* in Figure 6.15? (a) .8, (b) .6, (c) 0, (d) −.6, (e) −.9.
2. What old saying does it represent?

**Answers**

1. e
2. The early bird gets the worm.

## INTERPRETATION OF PEARSON *r*

You have seen that when two variables are highly related in a positive way, the correlation between them approaches +1.00. When they are highly related in a negative way, the correlation approaches −1.00. When there is little relation between variables, the correlation will be near 0. Pearson *r* provides a meaningful index for indicating relationship, with the sign of the coefficient indicating the direction of the relationship, and the difference between the coefficient and 0 indicating the degree of the relationship.

However, in interpreting the correlation coefficient, keep the following points in mind:

1. *Correlation does not necessarily indicate causation*. When two variables are found to be correlated, this indicates that relative positions in one variable are *associated* with relative positions in the other variable. It does not necessarily mean that changes in one variable are *caused* by changes in the other variable.

   We may find a correlation between two variables not because there is an intrinsic relationship between these variables but because they are both related to a third variable. For example, if we correlate the average teachers' salary for each of the past 20 years and the dollar value of hard liquor sold during each of these years, we get a high correlation. This does not mean that as soon as teachers' salaries are raised, they spend the extra money on booze. We observe a correlation between the two variables because each of them is highly correlated with a third variable, general inflation.

2. *The size of a correlation is in part a function of the variability of the two distributions to be correlated*. Restricting the range of the scores to be correlated reduces the observed degree of relationship between two variables. For example, people have observed that success in playing basketball is related to height: The taller an individual is, the more probable that he or she will do well in this sport. This statement is true about the population at large, where there is a wide range of heights. However, within a basketball team whose members are all tall, there may be little or no correlation between height and success because the range of heights is restricted. For a college that accepts students with a wide range of scores on a scholastic aptitude test, you would expect a correlation between the test scores and college grades. For a college that accepts only students with very high scholastic aptitude scores, you would expect very little correlation between the test scores and grades because of the restricted range of the test scores in this situation.

   If we correlate shoe size and reading vocabulary scores for a single grade level, we would expect a correlation of approximately zero. However, if we correlated this variable for all elementary students, we would get a high correlation because as children mature their feet get larger and their vocabulary increases.

3. *Correlation coefficients should not be interpreted in terms of percentage of perfect correlations*. Because correlation coefficients are expressed as decimal fractions, people who are not trained in statistics sometimes interpret correlation coefficients as a percentage of perfect correlation. An *r* of .80

does not indicate 80 percent of a perfect relationship between two variables. This interpretation is erroneous because, for example, an *r* of .80 does not express a relationship that is twice as great as an *r* of .40. A way of determining the degree to which you can predict one variable from the other is to calculate an index called the **coefficient of determination.** The coefficient of determination is the square of the correlation coefficient. It gives the percentage of variance in one variable that is associated with the variance in the other. For example, if you find a correlation of +.80 between achievement and intelligence, 64 percent of the variance in achievement is associated with variance in intelligence test scores. Probably the best way to give meaning to the size of the correlation coefficient is to picture the degree of scatter implied by correlations of different sizes (as illustrated in Figure 6.12) and to become familiar with the size of correlations commonly observed between variables of interest.

4. *Avoid interpreting the coefficients of correlation in an absolute sense.* In interpreting the degree of correlation, keep in mind the purpose for which it is being used. For example, it may not be wise to use a correlation of .5 for predicting the future performance of an individual. However, if you could develop a measure that you could administer to high school seniors that correlated with their subsequent college freshman grade point average (GPA), you could make a fortune because both ACT and SAT scores correlate approximately .4 with subsequent freshman GPAs. Correlations and their use in research are discussed further in Chapter 13.

## EFFECT SIZE

We have seen that the Pearson *r* indicates both the direction and the strength of a relationship between variables. The Pearson *r* has universal meaning in that an *r* near +1.00 always indicates a strong positive relationship no matter what the variables we are considering. An *r* near −1.00 always means a strong negative relationship, and an *r* near 0 always means a weak relationship. Smith and Glass (1977) originated the concept of **effect size,** a statistic that also has universal meaning to assess both the direction and the strength of a difference between two means. They subtract the mean of the control group from the mean of the experimental group and then divide this difference by the standard deviation of the control group, as seen in Formula 6.18:

$$\Delta = \frac{\overline{X}_e - \overline{X}_c}{s_c} \tag{6.18}$$

where

$\Delta$ = effect size for a difference between means
$\overline{X}_e$ = mean of the experimental group
$\overline{X}_c$ = mean of the control group
$s_c$ = standard deviation of the control group

In experimental studies, effect size can be used to compare the direction and the relative strength of different independent variables (intervention) on the same dependent variable.

Consider an experiment in which on-task behavior is the dependent variable. The experimental group receives contingent reinforcement, whereas the control group does not. (We explain this contrast more fully in Chapter 11.) The control group has a mean of 90 and a standard deviation of 10. The experimental group has a mean of 96. The effect size is $96 - 90/10 = .6$. Consider another experiment with the same dependent variable (on-task behavior) but with a drug treatment versus a placebo as the independent variable. The control group (which received the placebo) has a mean of 95 and a standard deviation of 8. The experimental group (which received a drug) has a mean of 97. The effect size here is $97 - 95/8 = .25$. The evidence suggests that contingent reinforcement has had a greater effect on on-task behavior than has the drug.

Effect sizes are interpreted in the same way that *z* scores are interpreted. Effect size can be used to compare the direction and the relative magnitude of the relationships that various independent variables have with a common dependent variable. In addition, it can be used to help decide whether the difference an independent variable makes on the dependent variable is strong enough to recommend its implementation in practice.

One approach is to ask if a given effect size is larger or smaller than effect sizes found in other studies with the same dependent variable. Also, you can assess the utility of an effect size by relating the cost in time, money, and other resources needed to implement the independent variable in relation to the importance of the dependent variable. A school of nursing would be interested in a brief, inexpensive procedure that produced an effect size of .20 on state nursing licensure exam scores. A researcher would hesitate to recommend an expensive, time-consuming independent variable with an effect size of .50 on a relatively unimportant dependent variable. Effect size is a useful statistic for assessing the strength and utility of a treatment or other independent variable. Cohen (1988) suggested the following interpretations:

An effect size of .20 is small.
An effect size of .50 is medium.
An effect size of .80 is large.

Cohen also developed an alternate index of effect size symbolized by lowercase *d*. **Cohen's *d*** is computed by subtracting the mean of the control group from the mean of the experimental group and dividing by the pooled standard deviation of the two groups:

$$d = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\Sigma x_1^2 + \Sigma x_2^2}{n_1 + n_2 - 2}}} \tag{6.19}$$

where

$d$ = effect size
$X_1$ = mean of one group
$X_2$ = mean of the other group
$\Sigma x_1$ = sum of deviation scores squared in the first group
$\Sigma x_2$ = sum of deviation scores squared in the second group
$n_1$ = number in first group
$n_2$ = number in second group

For example, we compare the scores of 28 students taught method A (group 1) and 22 students taught method B (group 2) with the following statistics. The $d$ is .56, indicating that the group 1 mean is higher than the group 2 mean by an effect size of .56, a medium effect size.

$$d = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\Sigma x_1^2 + \Sigma x_2^2}{n_1 + n_2 - 2}}} = \frac{82 - 79}{\sqrt{\frac{1390}{28 + 22 - 2}}}$$

$$= \frac{3}{\sqrt{\frac{1390}{48}}} = \frac{3}{\sqrt{28.9583}} = \frac{3}{5.3513} = .56$$

Many researchers prefer Cohen's $d$ to Smith and Glass's $\Delta$. Cohen's $d$ does not require designating one group as the control group in the numerator. Also, the denominator is an estimate of the population standard deviation based on the variance within both groups and the number in both groups.

As a form of $z$ scores, effect sizes have universal meaning. An effect size of $-.50$ always means that the group 1 mean was half a standard deviation below the group 2 mean. Effect size is important for evaluating the results of any quantitative study. The concept of effect size that originated in education is now widely used in other disciplines. The *Publication Manual of the American Psychological Association* (2001, p. 5) asserts, "For the reader to fully understand the importance of your findings, it is almost always necessary to include some index of effect size or strength of relationship in your results section."

Effect size can be calculated for various statistics other than $\overline{X}_1 - \overline{X}_2$. The coefficient of determination referred to previously is considered an effect size for correlations. For further discussion of effect size, see Cohen (1992), Kline (2004), or Vacha-Haase and Thompson (2004). Onwuegbuzie (2003) has proposed ways to apply effect size in qualitative research. The reporting of effect sizes is becoming more frequent in education and other fields. We hope it will soon be universal.

**A note of caution:** Effect size is independent of sample size. Therefore, large effect sizes can easily be observed through chance alone with very small samples. For example, an effect size of $d = .70$ between two samples of 4 each is essentially meaningless. A rule of thumb is that samples of less than 30 are considered small. (Yes, we know that most of our examples are less than 30. We did that so you would not get bogged down by the math.) In Chapter 7, we present ways of taking into account size of sample as well as effect size in evaluating results.

## META-ANALYSIS

Having formulated the concept of effect size, Smith and Glass (1977) proceeded to develop **meta-analysis,** a statistical technique that combines the effect sizes reported in the results of studies with the same (or similar) independent and dependent variables. In 1981, Glass, McGaw, and Smith wrote *Meta-Analysis in Social Research*, which is considered the classic work on meta-analysis. The result of a meta-analysis provides an overall summary of the outcomes of a number of studies by calculating a weighted average of their effect sizes. Meta-analysis gives a better estimate of the relationship among variables than do single studies alone. It is

important that the group of studies included in a meta-analysis focus on the same hypothesis or research questions with the same variables. One would not conduct a meta-analysis on school achievement in general but, rather, would focus on a particular area such as the effect of specific science teaching strategies on student achievement in high school chemistry. The average effect size as a formula is

$$\bar{\Delta} = \frac{\Delta_1 n_1 + \Delta_1 n_2 + \cdots + \Delta_k n_k}{N} \tag{6.20}$$

where

$\bar{\Delta}$ = average effect size
$\Delta_1$ = effect size for group 1
$\Delta_k$ = effect size of the last group
$n_1$ = number in first group
$n_k$ = number in last group
$N$ = total number of subjects

Suppose we have four studies investigating the effect of phonics instruction on reading proficiency. Their statistics are as follows:

Study 1: effect size = .9, $n = 60$
Study 2: effect size = .4, $n = 40$
Study 3: effect size = −.2, $n = 30$
Study 4: effect size = .1, $n = 70$

The average effect size is

$$\bar{\Delta} = .9(60) + .4(40) + -.2(30) + .1(70) = \frac{54 + 16 + -6 + 7}{200} = \frac{71}{200} = .36$$

This is approximately halfway between what Cohen describes as a small effect size and a medium effect size. Another way of looking at it is to use the normal curve. Consulting Table A.1 in the Appendix, you see that a $z$ score of .36 has a percentile rank of 64. The mean in a treatment group is equivalent to a score with a percentile rank of approximately 64 in the control group. If treatment is relatively inexpensive, you would be inclined to recommend it in practice. If treatment is expensive and/or the dependent variable is relatively unimportant, you would not be inclined to recommend it. The following is an abstract of a meta-analysis by Graham and Perin (2007):

> There is considerable concern that the majority of adolescents do not develop the competence in writing they need to be successful in school, the workplace, or their personal lives. A common explanation for why youngsters do not write well is that schools do not do a good job of teaching this complex skill. In an effort to identify effective instructional practices for teaching writing to adolescents, the authors conducted a meta-analysis of the writing intervention literature (Grades 4–12), focusing their efforts on experimental and quasi-experimental studies. They located 123 studies that yielded 154 effect sizes for quality of writing. The authors calculated an average weighted effect size (presented in parentheses) for the following 11 interventions: strategy instruction (.82), summarization (.82), peer assistance (.75), setting product goals (.70), word processing (.55), sentence combining (.50), inquiry (.32), prewriting activities (.32), process writing approach (.32), study of models (.25), grammar instruction (−.32).

> Based on the average weighted effect size of the interventions, the authors made a number of recommendations: (1) Teach adolescents strategies for planning, revision, and editing their compositions; (2) Teach strategies and procedures for summarizing reading material; (3) Develop instructional arrangements in which adolescents work together to plan, draft, revise, and edit their work; (4) Set clear and specific goals for what students are to accomplish with their writing; (5) Have students use word processing as a primary tool for writing; (6) Teach them how to combine sentences into increasingly more complex sentences; (7) Provide teachers with professional development in how to use the process writing approach; (8) Involve students with writing activities designed to sharpen their skills of inquiry; (9) Engage students in prewriting activities that help them to gather and organize ideas; (10) Provide students with good models for each type of writing that is the focus of instruction. (p. 445)

Meta-analysis is also used to integrate the findings of nonexperimental studies. For example, one might conduct a meta-analysis of studies that have examined gender differences in mathematics performance on standardized tests, or a meta-analysis might involve studies investigating the correlation of certain noncognitive variables to achievement in graduate education courses.

Meta-analysis has sometimes been criticized for including the results of poorly designed, poorly conducted studies along with the results of more credible studies (criteria for evaluating research designs are presented in Chapter 10). This problem can be resolved by first calculating the average effect size of all studies and then calculating the average effect size of well-designed studies to determine if the latter agree with the former. Qin, Johnson, and Johnson (1995) did this in a meta-analysis of studies comparing problem-solving performance of subjects under cooperative versus competitive conditions. In 55 cases, cooperation outperformed competition, whereas in 8 competition outperformed cooperation. The average effect size was .55 in favor of cooperation. The average effect size of the 33 studies judged as being of high methodological quality was .68.

The statistical computations involved in a meta-analysis are beyond the scope of this text. A number of computer packages are available for performing the two main computations: (1) calculating the effect size estimates and (2) analyzing the estimates obtained. The reader is referred to Hedges (1998) and Lipsey and Wilson (2000) for further discussion of meta-analysis approaches and applications. Effect size and meta-analysis, originally developed in education, are now widely used in medicine and many other fields.

## USING THE COMPUTER FOR DATA ANALYSIS

We have provided formulas and procedures for calculating the descriptive statistics covered in this chapter. Although we recognize that you will use the computer to calculate the statistics from your research, we believe that showing a statistic calculated "by hand" facilitates your understanding of what the statistic is and what it tells you about your data.

The best known and most widely used statistical package for data analysis in educational research is SPSS. This was initially developed in the 1960s for a mainframe platform, but it has long been available for PC and MacIntosh

environments. It can import data from most spreadsheet software and from databases. SPSS provides descriptive and inferential statistics and graphical presentations of data as well as more sophisticated statistical procedures. An SPSS 16.0 Student Version appropriate for use on home computers allows for up to 50 variables and 1500 cases. It is available in university or college bookstores or can be ordered online. After SPSS is loaded on your hard drive, you enter your data, name your variables, and then select the statistics from the menu. Visit the website at www.spss.com for further information about various options. The reader is also referred to Norusis (2006) or to Weinberg (2008) for an introduction to using SPSS for data analysis.

Another statistical package used in the social sciences is SAS (Statistical Analysis System). The SAS/STAT 9.2 version is currently available. It can be used to analyze most types of data and can produce linear models as well as multivariate and categorical statistics for different statistical approaches.

## SUMMARY

Descriptive statistics serve to describe and summarize observations. The descriptive technique to be employed is selected according to the purpose the statistic is to serve and the scale of measurement used.

Scales of measurement are means of quantifying observations and are of four types: (1) Nominal scales classify observations into mutually exclusive categories, (2) ordinal scales sort objects or classes of objects on the basis of their relative standing, (3) interval scales use equal intervals for measurement and indicate the degree to which a person or an object possesses a certain quality, and (4) ratio scales use equal intervals for measurement and measure from an absolute zero point. Once observations are quantified, the data can be arranged into frequency distributions and shown graphically in histograms or polygons.

Measures of central tendency—the mode, the median, and the mean—provide a single index to represent the average value of a whole set of measures. The mode, which is a nominal statistic, is the least stable and least useful measure in educational research. The median is an ordinal statistic that takes into account the ranks of scores within a distribution but not the size of the individual scores. The mean, which is an interval (or ratio) statistic, is the most stable and most widely used index of central tendency. Another way of describing observations is to indicate the variation, or spread, of the values within a distribution. The range, the variance, and the standard deviation are three indexes used for this purpose. The range, a nominal statistic, is the distance between the highest and the lowest values in a distribution, plus 1. Variance is the mean of the squared deviations of scores from the mean. It is an interval (or ratio) statistic. Standard deviation—the square root of the variance—is the most widely used index of variability.

Standard scores are used to indicate the position of a single score in a distribution. The most widely used is the $z$ score, which converts values into standard deviation units. The $z$ scores are often converted into stanines, $T$ scores, or other standard scores. An ordinal index of location shows a score's position in percentile rank, which indicates what percentage of scores fall below the midpoint of the score's interval. Using the characteristics and the areas of the normal curve, you can approximate the percentage of cases below and above each $z$ score in a normal distribution.

Correlation techniques enable researchers to describe the relationship between two sets of measures. Product moment correlation (Pearson $r$), the most widely used index of relationships, is used with interval or ratio data. Table 6.7 summarizes correlation indexes appropriate for interval, ordinal, and nominal data.

**Table 6.7** Summary of Descriptive Statistics Presented in This Chapter

| | Nominal | Ordinal | Interval |
|---|---|---|---|
| Indexes of central tendency | Mode | Median | Mean |
| Indexes of variability | Range | | Variance and standard deviation |
| Indexes of location | Label or classification | Percentile rank | *z* score, *T* scores, and other standard scores |
| Correlation indexes | | | Pearson *r* |

Effect size—the difference between the means of the experimental and control groups divided by the standard deviation of the control group—is a useful measure of the strength or magnitude of their relationship. A small effect size indicates a trivial relationship. A large effect size indicates a substantial relationship.

Meta-analysis is a widely used method for combining the statistical results of a group of studies on the same problem. It enables researchers to succinctly summarize results of many studies on a particular question. The most widely used index is the average effect size.

## KEY CONCEPTS

coefficient of determination
Cohen's *d*
correlation
correlation coefficient
curvilinear relationship
descriptive statistics
deviation scores
effect size
frequency distribution
frequency polygon
histogram
inferential statistics
interval scale
linear relationship
mean
measures of central tendency
median
meta-analysis
mode
negative correlation
negatively skewed distribution
nominal scale
normal curve
normal distribution
ordinal scale
Pearson product moment correlation coefficient
percentile rank
positive correlation
positively skewed distribution
range
scatterplot
skewed distribution ratio scale
standard deviation scatterplot
standard score
stanine score
symmetrical distribution
*T* score
variability
variance
*z* score

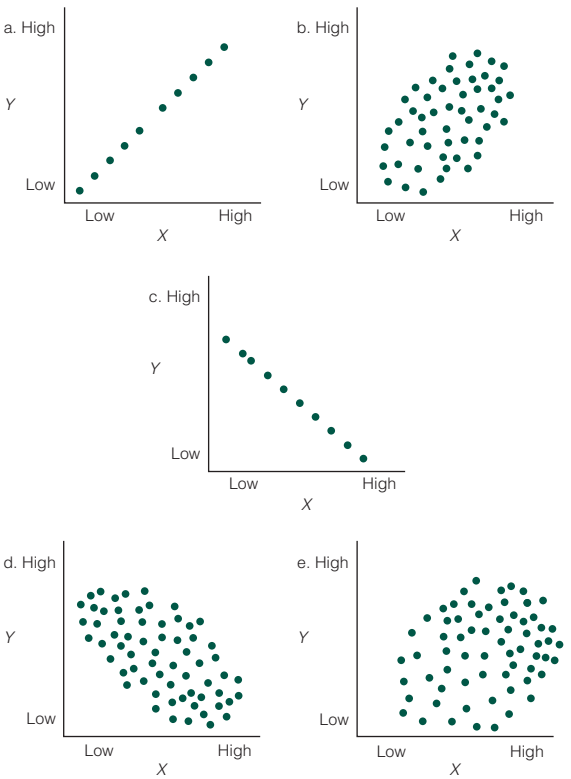
## EXERCISES

1. Identify the type of measurement scale—nominal, ordinal, interval, or ratio—suggested by each statement:
   - **a.** John finished the math test in 35 minutes, whereas Jack finished the same test in 25 minutes.
   - **b.** Jack speaks French, but John does not.
   - **c.** Jack is taller than John.
   - **d.** John is 6 feet 2 inches tall.
   - **e.** John's IQ is 120, whereas Jack's IQ is 110.

2. Draw a histogram and a frequency polygon for the following frequency distribution:

| *x* | *f* | *x* | *f* | *x* | *f* | *x* | *f* |
|---|---|---|---|---|---|---|---|
| 80 | 1 | 76 | 6 | 73 | 20 | 70 | 7 |
| 79 | 2 | 75 | 15 | 72 | 17 | 69 | 3 |
| 78 | 3 | 74 | 22 | 71 | 9 | | |
| 77 | 10 | | | | | | |

3. Provide answers as requested, given the following distribution: 15, 14, 14, 13, 11, 10, 10, 10, 8, 5.
   **a.** Calculate the mean.
   **b.** Determine the value of the median.
   **c.** Determine the value of the mode.
4. Briefly explain the relationship between the skewness of a distribution of scores and the resulting values of the mean, median, and mode.
5. Identify the measure—mode, mean, or median—that best suits each type of scale:
   **a.** Ordinal
   **b.** Nominal
   **c.** Interval
6. Identify the measure—mode, mean, or median—that each term defines:
   **a.** The middle score
   **b.** The arithmetic average
   **c.** The most frequently occurring score
7. Discuss the advantages and disadvantages of range and standard deviation as measures of variability of the scores.
8. **a.** Calculate the $z$ score for a score of 5 in a distribution with a mean of 7 and a standard deviation of 0.5.
   **b.** What would the stanine score for this score be?
9. Using Table A.1, what is the estimated percentile rank for a $z$ score of +1.20?
10. Why do you think the U.S. Census Bureau reports median income instead of mean income?
11. On an analysis of achievement test, male scores were more heterogeneous, more spread from low to high, than female scores.
    **a.** Which statistics would be greater for males?
    **b.** Which gender had more stanine scores of 9?
    **c.** Which gender had more stanine scores of 1?
    **d.** Which gender had more stanine scores of 5?
12. The mean score on a test is 40, and the standard deviation is 4. Express each of the following raw scores as a $z$ score:
    **a.** 41
    **b.** 30
    **c.** 48
    **d.** 36
    **e.** 46
13. **a.** What would be the $T$ score for the raw score of 46 in Exercise 12?
    **b.** What would the stanine score for the raw score of 46 be?
14. In a normal distribution, what percentage of the scores would fall below a $z$ score of −1.0? A $z$ score of 0? A $z$ score of +.67?
15. Describe the relationship shown by these scatterplots. Then estimate the correlation coefficients.




16. Each dot in a scatterplot represents _____.
17. Five girls took a history test and a geography test, with the following results:

| | History | | Geography | |
|---|---|---|---|---|
| | *Raw Score* | *z Score* | *Raw Score* | *z Score* |
| Ann | 28 | .5 | 85 | 1.5 |
| Nesa | 32 | 1.5 | 65 | .5 |
| María | 26 | 0 | 55 | 0 |
| Benazir | 20 | −1.5 | 45 | −.5 |
| Yoko | 24 | −.5 | 25 | −1.5 |

| History | Geography |
|---|---|
| $\Sigma = 310$ | $\Sigma X = 275$ |
| $\sigma = 4$ | $\sigma = 20$ |

**a.** What is the mean of the history test?
**b.** Whose performance in history is most in agreement with her performance in geography?
**c.** What is the correlation between the history and geography scores?

**18.** Given that the history test mean in Exercise 17 is lower than the geography test mean, which of the following conclusions would be correct?
**a.** These girls are better in history than in geography.
**b.** These girls are better in geography than in history.
**c.** Their teacher has probably spent twice as much time on geography as on history.
**d.** Their teacher knows more geography than history.
**e.** None of the above.

**19.** If the coefficient of correlation between variable $X$ and variable $Y$ is found to be $-.98$, which of the following would be indicated?
**a.** Variable $X$ and variable $Y$ are closely related.
**b.** Variable $X$ and variable $Y$ are unrelated.
**c.** Variable $X$ and variable $Y$ are perfectly related.
**d.** Variable $Y$ is a result of variable $X$.

**20.** For each of the following cases, indicate which statistic should be used—mean, standard deviation, $z$ score, or Pearson $r$.
**a.** We want to know how spread out or heterogeneous the scores of a class are.
**b.** We want to determine how Joe's score compares with the scores of the rest of the class.
**c.** We want to know how well the class as a whole did on an examination.
**d.** We want to predict the future achievement of students from their IQ.

**21.** Interpret the following: "On the College Board exam ($\mu = 500$, $\sigma = 100$), the mean of this year's Central High School seniors was 490, and the standard deviation was 110. The correlation between the exam scores and high school grade point average was $+.40$."

**22.** Interpret the following: "Parents of Central High students were asked to rank 10 problems from 10 = most serious to 1 = less serious. The median for the problem 'physical safety' was 5.21."

**23.** Smith and Glass's $\Delta$ and Cohen's $d$ are two ways of defining what?

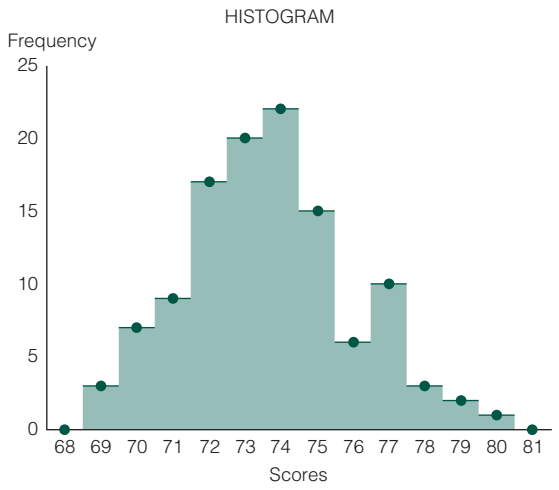
**24.** Define *effect size*, and tell how it is used.

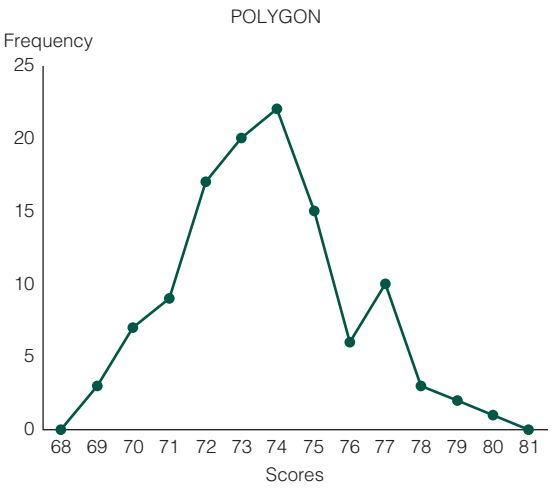
**25.** What is the purpose of meta-analysis?

## ANSWERS

**1. a.** Ratio
**b.** Nominal
**c.** Ordinal
**d.** Ratio
**e.** Interval

**2.** Figures may be expanded or contracted vertically or horizontally or both and be correct if the relationships between scores and frequencies are maintained.



3. **a.** Mean = 11
   **b.** Median = 10.5
   **c.** Mode = 10
4. The three measures are not equal in a skewed distribution. The mean is pulled in the direction of the skewed side. Thus, in a positively skewed distribution the mean is always higher than the median, and the mode is usually lowest in value. In a negatively skewed distribution, the mean is always lower than the median, and the mode is usually highest in value.
5. **a.** Median
   **b.** Mode
   **c.** Mean
6. **a.** Median
   **b.** Mean
   **c.** Mode
7. The range is easy to calculate and to explain. The standard deviation takes into account all the scores and is more stable.
8. **a.** $z = (5 - 7)/.5 = .4$
   **b.** Stanine = 1
9. 88 (rounded from 88.49)
10. Because the median is more typical of the incomes in the United States. It is not influenced by the extreme salaries of the billionaires and millionaires.
11. **a.** Male scores had higher variance, standard deviation, and range.
    **b.** Males
    **c.** Males
    **d.** Females
12. **a.** .25
    **b.** −2.5
    **c.** 2
    **d.** −1
    **e.** 1.5
13. **a.** $T = 10z + 50 = 10(1.5) + 50 = 65$
    **b.** Stanine score $2z + 5$ rounded = $(2)(1.5) + 5 = 8$
14. 16%; 50%; 75%
15. **a.** Perfect positive, +1 correlation
    **b.** Positive, +.75
    **c.** Perfect negative, −1
    **d.** Negative, −.75
    **e.** No correlation, 0
16. An individual's score on two dimensions or other paired *z* scores.
17. **a.** 26
    **b.** Maria; she had the same *z* score on both test.
    **c.** $r = \frac{\Sigma z_x z_y}{N} = \frac{3}{5} = .6$
18. e
19. a
20. **a.** Standard deviation
    **b.** *z* score
    **c.** Mean
    **d.** Pearson *r*
21. As a group, the Central High seniors were slightly below the national average. Their scores were slightly more heterogeneous than usual. Those with high scores tended to have high GPAs. Those with low scores tended to have low GPAs. The relationship between scores and GPAs was moderate positive.
22. Parents ranked "physical safety" about average.
23. Effect size
24. Effect size (a form of *z* score) is the difference between experimental and control groups divided by the standard deviation of the control group. It indicates the strength of the relationship between the independent and dependent variables.
25. It combines the result of studies with similar independent and dependent variables to produce an average effect size, a mathematical summary of the results.

## REFERENCES

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology*, *99*(3), 445–476.

Hedges, L. V. (1998). *Statistical methods for meta-analysis*. San Diego: Academic Press.

Kline, R. B. (2004). *Beyond significance testing; Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Lipsey, M. W., & Wilson, D. B. (2000). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Norusis, M. J. (2006). *SPSS 15.0 guide to data analysis*. Upper Saddle River, NJ: Prentice Hall.

Onwuegbuzie, A. D. (2003). Effect sizes in qualitative research: A prolegomenon. *Quality and Quantity*, *37*, 393–409.

Qin, Z., Johnson, D. W., & Johnson, R. T. (1995). Cooperative versus competitive efforts and problem solving. *Review of Educational Research*, *65*(2), 129–143.

Seashore, H. D. (Ed.). (1955). *Methods of expressing test scores*, Test Service Bulletin No. 48. New York: American Psychological Corporation.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, *32*, 752–760.

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (p. 1). New York: Wiley.

Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology*, *51*(4), 473–481.

Weinberg, S. (2008). *Statistics using SPSS: An integrative approach* (2nd ed.). Cambridge, UK: Cambridge University Press.

CHAPTER 7

# Sampling and Inferential Statistics

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.

H. G. Wells (1866–1946)

## INSTRUCTIONAL OBJECTIVES

After studying this chapter, the student will be able to:

1. Describe the meaning, rationale, and steps involved in sampling and distinguish between probability and nonprobability sampling.
2. List the characteristics, uses, and limitations of each kind of probability and nonprobability sampling.
3. Explain the meaning of sampling error and its relationship to making statistical inferences.
4. Describe random assignment and its use in research.
5. Explain the meaning of statistical significance.
6. Explain the general strategy of statistical significance testing.
7. Explain how a null hypothesis is used in scientific research.
8. Describe Type I and Type II errors.
9. Define alpha and beta.
10. Explain the difference between directional and nondirectional tests of significance and the appropriate use of each of these tests.
11. Describe the process of determining sample size.
12. Apply the *t* test to find the significance of a difference between independent means and a difference between dependent means.
13. Understand the concept of degrees of freedom.
14. Know what statistical tests are available for ranked data.
15. Describe the purpose of the *t* test for the Pearson *r* and determine if a given correlation coefficient differs significantly from zero.
16. Describe the uses of *F* tests.
17. Apply the *F* test for finding the significance of the differences between groups in one-way and two-way analysis of variance.

18 Apply the chi-square test for finding the significance of the differences between proportions in one-way and two-way classifications.

19 Select the kind of inferential statistical procedures appropriate for use in testing a given research hypothesis.

20 Demonstrate comprehension of the basic technical statistical terms used in reporting research results.

The statistics discussed in the previous chapter are used for organizing, summarizing, and describing data. In research, however, we often need to go further than describing data. After making observations of a sample, researchers employ induction or inference to generalize findings to the entire population from which the sample was drawn. To do this, they need techniques that enable them to make credible inferences from samples to whole populations.

## SAMPLING

An important characteristic of inferential statistics is the process of going from the part to the whole. For example, you might study a randomly selected group of 500 students attending a university in order to make generalizations about the entire student body of that university.

The small group that is observed is called a *sample*, and the larger group about which the generalization is made is called a *population*. A **population** is defined as all members of any well-defined class of people, events, or objects. For example, in a study in which students in American high schools constitute the population of interest, you could define this population as all boys and girls attending high school in the United States. A **sample** is a portion of a population. For example, the students of Washington High School in Indianapolis constitute a sample of American high school students.

Statistical inference is a procedure by means of which you estimate **parameters** (characteristics of populations) from **statistics** (characteristics of samples). Such estimations are based on the laws of probability and are best estimates rather than absolute facts. In making any such inferences, a certain degree of error is involved. Inferential statistics can be used to test hypotheses about populations on the basis of observations of a sample drawn from the population.

### RATIONALE OF SAMPLING

Inductive reasoning is an essential part of the scientific approach. The inductive method involves making observations and then drawing conclusions from these observations. If you can observe all instances of a population, you can, with confidence, base conclusions about the population on these observations (perfect induction). In Chapter 6, we treated the 18 students in Mr. Li's physics class as a population. Therefore, we could be confident that we had the true means, standard deviations, and so forth (the parameters). However, if you observe only some instances of a population, then you can do no more than infer that these observations will be true of the population as a whole (imperfect induction). This is

the concept of sampling, which involves taking a portion of the population, making observations on this smaller group, and then generalizing the findings to the parent population—the larger population from which the sample was drawn.

Sampling is indispensable to the researcher. Usually, the time, money, and effort involved do not permit a researcher to study all possible members of a population. Furthermore, it is generally not necessary to study all possible cases to understand the phenomenon under consideration. Sampling comes to your aid by enabling you to study a portion of the population rather than the entire population.

Because the purpose of drawing a sample from a population is to obtain information concerning that population, it is extremely important that the individuals included in a sample constitute a representative cross section of individuals in the population. Samples must be representative if you are to be able to generalize with reasonable confidence from the sample to the population. For example, the researcher may assume that the students at Washington High School are representative of American adolescents. However, this sample may not be representative if the individuals who are included have some characteristics that differ from the target population. The location of their school, their socioeconomic backgrounds, their family situations, their prior experiences, and many other characteristics of this group may make them unrepresentative of American adolescents. An unrepresentative sample is termed a **biased sample.** The findings on a biased sample in a research study cannot legitimately be generalized to the population from which it is taken. For example, if the population of interest is all students in a particular urban school district but the researchers sampled only students from the district's two magnet schools, the sample would be biased.

## STEPS IN SAMPLING

The first step in sampling is the identification of the **target population,** the large group to which the researcher wishes to generalize the results of the study. If the researcher is interested in learning about the teachers in the St. Louis public school system, all those who teach within that system constitute the target population. In a study of the attitudes and values of American adolescents, the target population would be all American boys and girls in the age range of 12 to 21 years, given that adolescence is operationally defined as the period between ages 12 and 21 years. We make a distinction between the target population and the **accessible population,** which is the population of subjects accessible to the researcher for drawing a sample. In most research, we deal with accessible populations. It would be expensive and time-consuming to sample from the total population of American adolescents, but we could draw a sample of adolescents from one state. Of course, we could only generalize results to adolescents in the chosen state, not to all American adolescents.

Once we have identified the population, the next step is to select the sample. Two major types of sampling procedures are available to researchers: probability and nonprobability sampling. **Probability sampling** involves sample selection in which the elements are drawn by chance procedures. The main characteristic of probability sampling is that every member or element of the population has a known probability of being chosen in the sample.

**Nonprobability sampling** includes methods of selection in which elements are not chosen by chance procedures. Its success depends on the knowledge,

expertise, and judgment of the researcher. Nonprobability sampling is used when the application of probability sampling is not feasible. Its advantages are convenience and economy.

## PROBABILITY SAMPLING

Probability sampling is defined as the kind of sampling in which every element in the population has an equal chance of being selected. The possible inclusion of each population element in this kind of sampling takes place by chance and is attained through random selection. When probability sampling is used, inferential statistics enable researchers to estimate the extent to which the findings based on the sample are likely to differ from what they would have found by studying the whole population. The four types of probability sampling most frequently used in educational research are simple random sampling, stratified sampling, cluster sampling, and systematic sampling.

### Simple Random Sampling

The best known of the probability sampling procedures is **simple random sampling.** The basic characteristic of simple random sampling is that all members of the population have an equal and independent chance of being included in the **random sample.** The steps in simple random sampling comprise the following:

1. Define the population.
2. List all members of the population.
3. Select the sample by employing a procedure where sheer chance determines which members on the list are drawn for the sample.

The first step in drawing a random sample from a population is to assign each member of the population a distinct identification number. Let us illustrate this procedure by showing how to obtain a sample of 50 students from the population attending Washington High School. First, you need to enumerate all the individuals in the population. The principal's office could supply a list of all students enrolled in the school. For identification purposes, you would then assign a number to each individual in the population. If there are 800 students in the school, you use the numbers 000, 001, 002, 003, . . . , 799 for this purpose. Each individual must have an identification value with the same number of digits as every other individual. Many schools have already assigned identification numbers to all their students. One way to draw a random sample would be to write the student numbers on separate slips of paper, place the pieces of paper in a container, shake the container, and draw out a slip of paper. Shake the container again, draw out another paper, and continue the process until 50 slips of paper have been picked. This process would be very tedious. A more systematic way to obtain a random sample is to use a table of random numbers, which includes a series of numbers, typically four to six digits in length, arranged in columns and rows (see Table 7.1 for a small segment of a table). A table of random numbers is produced by a computer program that guarantees that all the digits (0–9) have an equal chance of occurring each time a digit is printed. Most statistics books include a table of random numbers in the appendix. In previous editions of this book, we included a five-page table of random numbers. We decided this is no

longer needed because there are so many tables available on the Internet, in statistics texts, and from other sources.

Let us illustrate how to use a table of random numbers. With our list of the 800 students in the population, we will use a table to obtain numbers of three digits each, using only those numbers that are less than or equal to 799. For each number chosen, the corresponding member of the population falls into the sample. Continue the process until the desired number for the sample has been chosen—in this case, the first 50 numbers that meet the criterion.

We begin by randomly selecting a starting point in the table. You can do this by closing your eyes and putting your finger on the page, or you can use a procedure that is an absolutely random way to enter the table. First, roll a die to determine which page to use. We roll a 3, so we pull the third page from a table of random numbers (Table 7.1). Then we note the last two digits from the serial number on a dollar bill. They are 03, so we go to row 3. Then we take the last two digits from a second dollar bill, which are 22, taking us to the intersection of row 3 and column 22. The intersection of the row and column is the location of the first random number. Because our population is 800, we will only look at the first three digits of the numbers in the table. If the population were 1500, we would look at the first four digits. In our example, we could use either the first three digits or the last three; we have chosen to use the first three. The first three digits from that intersection are 403, so the individual with number 403 is in the sample. Because the digits in a table are random, the numbers can be used vertically in both directions or horizontally in both directions. You should specify the direction you will use prior to entering the table and use it consistently. The remaining numbers would be located by moving in the specified direction. If we have decided to move vertically, the next three digits are 497, 243, 262, 782, and on down the column through 351. The next number is 995, which is larger than 799 (the size of the sample) so we would skip it and move on down, selecting the numbers smaller than 799. We have highlighted the numbers in that column that would be selected. You would then move to the next column and continue the process until you have 50 random numbers less than 799.

You probably will not actually have to do all this. However, we wanted to show you a way in which the numbers drawn from a table of random numbers can be absolutely without bias. You most likely will have access to web-based random number generators such as Research Randomizer (www.randomizer.org). If you access this website, you will find information about the Research Randomizer and a tutorial on how to use it to generate random numbers quickly. It is part of the Social Psychology Network and is free. Or, you may be lucky and conduct your research in a school whose record-keeping system allows for drawing a random sample using the school's computer.

The generally understood meaning of the word *random* is "without purpose or by accident." However, random sampling is purposeful and methodical. It is apparent that a sample selected randomly is not subject to the biases of the researcher. Rather, researchers commit themselves to selecting a sample in such a way that their biases are not permitted to operate; chance alone determines which elements in the population will be in the sample. They are pledging to avoid a deliberate selection of subjects who will confirm the hypothesis.

**Table 7.1** Page from a Table of Random Numbers

| | Column Number | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 00000 | 00000 | 11111 | 11111 | 22222 | 22222 | 33333 | 33333 |
| Row | 01234 | 56789 | 01234 | 56789 | 01234 | 56789 | 01234 | 56789 |
| | | | | | *3rd Thousand* | | | |
| 00 | 89221 | 02362 | 65787 | 74733 | 51272 | 30213 | 92441 | 39651 |
| 01 | 04005 | 99818 | 63918 | 29032 | 94012 | 42363 | 01261 | 10650 |
| 02 | 98546 | 38066 | 50856 | 75045 | 40645 | 22841 | 53254 | 44125 |
| 03 | 41719 | 84401 | 59226 | 01314 | 54581 | 40398 | 49988 | 65579 |
| 04 | 28733 | 72489 | 00785 | 25843 | 24613 | 49797 | 85567 | 84471 |
| 05 | 65213 | 83927 | 77762 | 03086 | 80742 | 24395 | 68476 | 83792 |
| 06 | 65553 | 12678 | 90906 | 90466 | 43670 | 26217 | 69900 | 31205 |
| 07 | 05668 | 69080 | 73029 | 85746 | 58332 | 78231 | 45986 | 92998 |
| 08 | 39202 | 99718 | 49757 | 79519 | 27387 | 76373 | 47262 | 91612 |
| 09 | 64592 | 32254 | 45879 | 29431 | 38320 | 05981 | 18067 | 87137 |
| 10 | 07513 | 48792 | 47314 | 83660 | 68907 | 05336 | 82579 | 91582 |
| 11 | 86593 | 68501 | 56638 | 99800 | 82839 | 35148 | 56541 | 07232 |
| 12 | 83735 | 22599 | 97977 | 81248 | 36838 | 99560 | 32410 | 67614 |
| 13 | 08595 | 21826 | 54655 | 08204 | 87990 | 17033 | 56258 | 05384 |
| 14 | 41273 | 27149 | 44293 | 69458 | 16828 | 63962 | 15864 | 35431 |
| 15 | 00473 | 75908 | 56238 | 12242 | 72631 | 76314 | 47252 | 06347 |
| 16 | 86131 | 53789 | 81383 | 07868 | 89132 | 96182 | 07009 | 86432 |
| 17 | 33849 | 78359 | 08402 | 03586 | 03176 | 88663 | 08018 | 22546 |
| 18 | 61870 | 41657 | 07468 | 08612 | 98083 | 97349 | 20775 | 45091 |
| 19 | 43898 | 65923 | 25078 | 86129 | 78491 | 97653 | 91500 | 80786 |
| 20 | 29939 | 39123 | 04548 | 45985 | 60952 | 06641 | 28726 | 46473 |
| 21 | 38505 | 85555 | 14388 | 55077 | 18657 | 94887 | 67831 | 70819 |
| 22 | 31824 | 38431 | 67125 | 25511 | 72044 | 11562 | 53279 | 82268 |
| 23 | 91430 | 03767 | 13561 | 15597 | 06750 | 92552 | 02391 | 38753 |
| 24 | 38635 | 68976 | 25498 | 97526 | 96458 | 03805 | 04116 | 63514 |

You would expect a random sample to be representative of the target population sampled. However, a random selection, especially with small samples, does not absolutely guarantee a sample that will represent the population well. Random selection does guarantee that any differences between the sample and the parent population are only a function of chance and not a result of the researcher's bias. The differences between random samples and their parent population are not systematic. For example, the mean reading achievement of a random sample of sixth-graders may be higher than the mean reading achievement of the target population, but it is equally likely that the mean for the sample will be lower than the mean for the target population. In other words, with random sampling the sampling errors are just as likely to be negative as they are to be positive.

Furthermore, statistical theorists have shown, through deductive reasoning, how much a researcher can expect the observations derived from random samples to differ from what would be observed in the population when the null

hypothesis is true. All inferential statistical procedures have this aim in mind. When random sampling is used, the researcher can employ inferential statistics to estimate how much the population is likely to differ from the sample. The inferential statistics in this chapter are all based on random sampling and apply directly only to those cases in which the sampling has been random.

Unfortunately, simple random sampling requires enumeration of all individuals in a finite population before the sample can be drawn—a requirement that often presents a serious obstacle to the practical use of this method. Now let us look at other probability sampling methods that approximate simple random sampling and may be used as alternatives in certain situations.

### Stratified Sampling

When the population consists of a number of subgroups, or strata, that may differ in the characteristics being studied, it is often desirable to use a form of probability sampling called **stratified sampling.** For example, if you were conducting a poll designed to assess opinions on a certain political issue, it might be advisable to subdivide the population into subgroups on the basis of age, neighborhood, and occupation because you would expect opinions to differ systematically among various ages, neighborhoods, and occupational groups. In stratified sampling, you first identify the strata of interest and then randomly draw a specified number of subjects from each stratum. The basis for stratification may be geographic or may involve characteristics of the population such as income, occupation, gender, age, year in college, or teaching level. In studying adolescents, for example, you might be interested not merely in surveying the attitudes of adolescents toward certain phenomena but also in comparing the attitudes of adolescents who reside in small towns with those who live in medium-size and large cities. In such a case, you would divide the adolescent population into three groups based on the size of the towns or cities in which they reside and then randomly select independent samples from each stratum.

An advantage of stratified sampling is that it enables the researcher to also study the differences that might exist between various subgroups of a population. In this kind of sampling, you may either take equal numbers from each stratum or select in proportion to the size of the stratum in the population. The latter procedure is known as **proportional stratified sampling,** which is applied when the characteristics of the entire population are the main concern in the study. Each stratum is represented in the sample in exact proportion to its frequency in the total population. For example, if 10 percent of the voting population are college students, then 10 percent of a sample of voters to be polled would be taken from this stratum. If a superintendent wants to survey the teachers in a school district regarding some policy and believes that teachers at different levels may feel differently, he or she could stratify on teaching level and then select a number from each level in proportion to its size in the total population of teachers. If 43 percent of the teachers are high school teachers, then 43 percent of the sample would be high school teachers.

In some research studies, however, the main concern is with differences among various strata. In these cases, the researcher chooses samples of equal size from each stratum. For example, if you are investigating the difference between the attitudes of graduate and undergraduate students toward an issue, you include

equal numbers in both groups and then study the differences that might exist between them. You choose the procedure according to the nature of the research question. If your emphasis is on the types of differences among the strata, you select equal numbers of cases from each. If the characteristics of the entire population are your main concern, proportional sampling is more appropriate. When the population to be sampled is not homogeneous but consists of several subgroups, stratified sampling may give a more representative sample than simple random sampling. In simple random sampling, certain strata may by chance be over- or underrepresented in the sample. For example, in the simple random sample of high school students it would be theoretically possible (although highly unlikely) to obtain female subjects only. This could not happen, however, if males and females were listed separately and a random sample were then chosen from each group. The major advantage of stratified sampling is that it guarantees representation of defined groups in the population.

### Cluster Sampling

As mentioned previously, it is very difficult, if not impossible, to list all the members of a target population and select the sample from among them. The population of American high school students, for example, is so large that you cannot list all its members for the purpose of drawing a sample. In addition, it would be very expensive to study a sample that is scattered throughout the United States. In this case, it would be more convenient to study subjects in naturally occurring groups, or clusters. For example, a researcher might choose a number of schools randomly from a list of schools and then include all the students in those schools in the sample. This kind of probability sampling is referred to as **cluster sampling** because the unit chosen is not an individual but, rather, a group of individuals who are naturally together. These individuals constitute a cluster insofar as they are alike with respect to characteristics relevant to the variables of the study. To illustrate, let us assume a public opinion poll is being conducted in Atlanta. The investigator would probably not have access to a list of the entire adult population; thus, it would be impossible to draw a simple random sample. A more feasible approach would involve the selection of a random sample of, for example, 50 blocks from a city map and then the polling of all the adults living on those blocks. Each block represents a cluster of subjects, similar in certain characteristics associated with living in proximity. A common application of cluster sampling in education is the use of intact classrooms as clusters.

It is essential that the clusters actually included in your study be chosen at random from a population of clusters. Another procedural requirement is that once a cluster is selected, *all* the members of the cluster must be included in the sample. The sampling error (discussed later) in a cluster sample is much greater than in true random sampling. It is also important to remember that if the number of clusters is small, the likelihood of sampling error is great—even if the total number of subjects is large.

### Systematic Sampling

Still another form of probability sampling is called **systematic sampling.** This procedure involves drawing a sample by taking every *K*th case from a list of the population.

First, you decide how many subjects you want in the sample ($n$). Because you know the total number of members in the population ($N$), you simply divide $N$ by $n$ and determine the sampling interval ($K$) to apply to the list. Select the first member randomly from the first $K$ members of the list and then select every $K$th member of the population for the sample. For example, let us assume a total population of 500 subjects and a desired sample size of 50: $K = N/n = 500/50 = 10$.

Start near the top of the list so that the first case can be randomly selected from the first 10 cases, and then select every tenth case thereafter. Suppose the third name or number on the list was the first selected. You would then add the sampling interval, or 10, to 3—and thus the 13th person falls in the sample, as does the 23rd, and so on—and would continue adding the constant sampling interval until you reached the end of the list.

Systematic sampling differs from simple random sampling in that the various choices are not independent. Once the first case is chosen, all subsequent cases to be included in the sample are automatically determined. If the original population list is in random order, systematic sampling would yield a sample that could be statistically considered a reasonable substitute for a random sample. However, if the list is not random, it is possible that every $K$th member of the population might have some unique characteristic that would affect the dependent variable of the study and thus yield a biased sample. Systematic sampling from an alphabetical list, for example, would probably not give a representative sample of various national groups because certain national groups tend to cluster under certain letters, and the sampling interval could omit them entirely or at least not include them to an adequate extent.

Note that the various types of probability sampling that have been discussed are not mutually exclusive. Various combinations may be used. For example, you could use cluster sampling if you were studying a very large and widely dispersed population. At the same time, you might be interested in stratifying the sample to answer questions regarding its different strata. In this case, you would stratify the population according to the predetermined criteria and then randomly select the cluster of subjects from among each stratum.

## NONPROBABILITY SAMPLING

In many research situations, the enumeration of the population elements—a basic requirement in probability sampling—is difficult, if not impossible. Or a school principal might not permit a researcher to draw a random sample of students for a study but would permit use of certain classes. In these instances, the researcher would use nonprobability sampling, which involves nonrandom procedures for selecting the members of the sample. In nonprobability sampling, there is no assurance that every element in the population has a chance of being included. Its main advantages are convenience and economy. The major forms of nonprobability sampling are convenience sampling, purposive sampling, and quota sampling.

### Convenience Sampling

**Convenience sampling,** which is regarded as the weakest of all sampling procedures, involves using available cases for a study. Interviewing the first individuals

you encounter on campus, using a large undergraduate class, using the students in your own classroom as a sample, or taking volunteers to be interviewed in survey research are various examples of convenience sampling. There is no way (except by repeating the study using probability sampling) of estimating the error introduced by the convenience sampling procedures. Probability sampling is the ideal, but in practice, convenience sampling may be all that is available to a researcher. In this case, a convenience sample is perhaps better than nothing at all. If you do use convenience sampling, be extremely cautious in interpreting the findings and know that you cannot generalize the findings.

### Purposive Sampling

In **purposive sampling**—also referred to as **judgment sampling**—sample elements judged to be typical, or representative, are chosen from the population. The assumption is that errors of judgment in the selection will counterbalance one another. Researchers often use purposive sampling for forecasting national elections. In each state, they choose a number of small districts whose returns in previous elections have been typical of the entire state. They interview all the eligible voters in these districts and use the results to predict the voting patterns of the state. Using similar procedures in all states, the pollsters forecast the national results.

The critical question in purposive sampling is the extent to which judgment can be relied on to arrive at a typical sample. There is no reason to assume that the units judged to be typical of the population will continue to be typical over a period of time. Consequently, the results of a study using purposive sampling may be misleading. Because of its low cost and convenience, purposive sampling has been useful in attitude and opinion surveys. Be aware of the limitations, however, and use the method with extreme caution.

### Quota Sampling

**Quota sampling** involves selecting typical cases from diverse strata of a population. The quotas are based on known characteristics of the population to which you wish to generalize. Elements are drawn so that the resulting sample is a miniature approximation of the population with respect to the selected characteristics. For example, if census results show that 25 percent of the population of an urban area lives in the suburbs, then 25 percent of the sample should come from the suburbs.

Here are the steps in quota sampling:

1. Determine a number of variables, strongly related to the question under investigation, to be used as bases for stratification. Variables such as gender, age, education, and social class are frequently used.
2. Using census or other available data, determine the size of each segment of the population.
3. Compute quotas for each segment of the population that are proportional to the size of each segment.
4. Select typical cases from each segment, or stratum, of the population to fill the quotas.

The major weakness of quota sampling lies in step 4, the selection of individuals from each stratum. You simply do not know whether the individuals chosen are representative of the given stratum. The selection of elements is likely to be based on accessibility and convenience. If you are selecting 25 percent of the households in the inner city for a survey, you are more likely to go to houses that are attractive rather than dilapidated, to those that are more accessible, to those where people are at home during the day, and so on. Such procedures automatically result in a systematic bias in the sample because certain elements are going to be misrepresented. Furthermore, there is no basis for calculating the error involved in quota sampling.

Despite these shortcomings, researchers have used quota sampling in many projects that might otherwise not have been possible. Many believe that speed of data collection outweighs the disadvantages. Moreover, years of experience with quota samples have made it possible to identify some of the pitfalls and to take steps to avoid them.

## RANDOM ASSIGNMENT

We distinguish random sampling from random assignment. **Random assignment** is a procedure used after we have a sample of participants and before we expose them to a treatment. For example, if we wish to compare the effects of two treatments on the same dependent variable, we use random assignment to put our available participants into groups. Random assignment requires a chance procedure such as a table of random numbers to divide the available subjects into groups. Then a chance procedure such as tossing a coin is used to decide which group gets which treatment.

As with random sampling, any bias the researcher has will not influence who gets what treatment, and the groups will be statistically equivalent before treatment. Group 1 may have more highly motivated subjects than group 2, but it is just as likely that group 2 will have more highly motivated subjects than group 1. The same is true of all possible known or unknown variables that might influence the dependent variable. Therefore, the same lawful nature of sampling errors that are true of random sampling are true of random assignment.

## THE SIZE OF THE SAMPLE (FUNDAMENTALS)

Laypeople are often inclined to criticize research (especially research whose results they do not like) by saying the sample was too small to justify the researchers' conclusions. How large should a sample be? Other things being equal, a larger sample is more likely to be a good representative of the population than a smaller sample. However, the most important characteristic of a sample is its representativeness, not its size. A random sample of 200 is better than a random sample of 100, but a random sample of 100 is better than a biased sample of 2.5 million.

Size alone will not guarantee accuracy. A sample may be large and still contain a bias. The latter situation is well illustrated by the *Literary Digest* magazine poll of 1936, which predicted the defeat of President Roosevelt. Although the sample included approximately 2.5 million respondents, it was not representative of the voters; thus, the pollsters reached an erroneous conclusion. The

bias resulted from selecting respondents for the poll from automobile registrations, telephone directories, and the magazine's subscription lists. These subjects would certainly not represent the total voting population in 1936, when many people could not afford automobiles, telephones, or magazines. Also, because the poll was conducted by mail, the results were biased by differences between those who responded and those who did not. We have since learned that with mailed questionnaires, those who are against the party in power are more likely to return their questionnaires than those who favor the party in power. The researcher must recognize that sample size will not compensate for any bias that faulty sampling techniques may introduce. Representativeness must remain the prime goal in sample selection.

Later in this chapter, we introduce a procedure for determining appropriate sample size, on the basis of how large an effect size is considered meaningful and on statistical considerations. Such procedures, known as **power calculations,** are the best way to determine needed sample sizes.

## THE CONCEPT OF SAMPLING ERROR

When an inference is made from a sample to a population, a certain amount of error is involved because even random samples can be expected to vary from one to another. The mean intelligence score of one random sample of fourth-graders will probably differ from the mean intelligence score of another random sample of fourth-graders from the same population. Such differences, called **sampling errors,** result from the fact that the researcher has observed only a sample and not the entire population.

Sampling error is "the difference between a population parameter and a sample statistic." For example, if you know the mean of the entire population (symbolized $\mu$) and also the mean of a random sample (symbolized $\overline{X}$) from that population, the difference between these two ($\overline{X} - \mu$) represents sampling error (symbolized e). Thus, $e = \overline{X} - \mu$. For example, if you know that the mean intelligence score for a population of 10,000 fourth-graders is $\mu = 100$ and a particular random sample of 200 has a mean of $\overline{X} = 99$, then the sampling error is $\overline{X} - \mu = 99 - 100 = -1$. Because we usually depend on sample statistics to estimate population parameters, the notion of how samples are expected to vary from populations is a basic element in inferential statistics. However, instead of trying to determine the discrepancy between a sample statistic and the population parameter (which is not often known), the approach in inferential statistics is to estimate the variability that could be expected in the statistics from a number of different random samples drawn from the same population. Because each of the sample statistics is considered to be an estimate of the same population parameter, any variation among sample statistics must be attributed to sampling error.

### The Lawful Nature of Sampling Errors

Given that random samples drawn from the same population will vary from one another, is using a sample to make inferences about a population really any better than just guessing? Yes, it is because sampling errors behave in a lawful and predictable manner. The laws concerning sampling error have been derived through deductive logic and have been confirmed through experience.

Although researchers cannot predict the nature and extent of the error in a single sample, they can predict the nature and extent of sampling errors in general. Let us illustrate with reference to sampling errors connected with the mean.

### Sampling Errors of the Mean

Some sampling error can always be expected when a sample mean is used to estimate a population mean $\mu$. Although, in practice, such an estimate is based on a single sample mean, assume that you drew several random samples from the same population and computed a mean for each sample. You would find that these sample means would differ from one another and would also differ from the population mean (if it were known). Statisticians have carefully studied sampling errors of the mean and found that they follow known laws.

1. *The expected mean of sampling errors is zero*. Given an infinite number of random samples drawn from a single population, the positive errors can be expected to balance the negative errors so that the mean of the sampling errors will be zero. For example, if the mean height of a population of college freshmen is 5 feet 9 inches and several random samples are drawn from that population, you would expect some samples to have mean heights greater than 5 feet 9 inches and some to have mean heights less than 5 feet 9 inches. In the long run, however, the positive and negative sampling errors will balance. If you had an infinite number of random samples of the same size, calculated the mean of each of these samples, and then computed the mean of all these means, this mean would be equal to the population mean.

   Because positive errors equal negative errors, a single sample mean is as likely to underestimate a population mean as to overestimate it. Therefore, we can justify stating that a sample mean is an unbiased estimate of the population mean and is a reasonable estimate of the population mean.

2. *Sampling error is an inverse function of sample size*. As the size of a random sample increases, there is less fluctuation from one sample to another in the value of the mean. In other words, as the size of a sample increases, the expected sampling error decreases. Small samples produce more sampling error than large ones. You would expect the means based on samples of 10 to fluctuate a great deal more than the means based on samples of 100. In the height example, it is much more likely that a random sample of 4 will include 3 above-average freshmen and 1 below-average freshman than that a random sample of 40 would include 30 above-average and 10 below-average freshman. As sample size increases, the likelihood that the mean of the sample is near the population mean also increases. There is a mathematical relationship between sample size and sampling error. This relationship has been incorporated into inferential formulas, which we discuss later.

3. *Sampling error is a direct function of the standard deviation of the population*. The more spread, or variation, there is among members of a population, the more spread there will be in sample means. For example, the mean weights of random samples of 25, each selected from a population of professional jockeys, would show relatively less sampling error than the mean weights of samples of 25 selected from a population of schoolteachers. The weights of professional jockeys fall within a narrow range; the weights

of schoolteachers do not. Therefore, for a given sample size, the expected sampling error for teachers' weights would be greater than the expected sampling error for jockeys' weights.

4. *Sampling errors are distributed in a normal or near-normal manner around the expected mean of zero*. Sample means near the population mean will occur more frequently than sample means far from the population mean. As you move farther and farther from the population mean, you find fewer and fewer sample means occurring. Both theory and experience have shown that the means of random samples are distributed in a normal or near-normal manner around the population mean. Because a sampling error in this case is the difference between a sample mean and the population mean, the distribution of sampling errors is also normal or near normal in shape.

The distribution of sample means will resemble a normal curve even when the population from which the samples are drawn is not normally distributed. For example, in a typical elementary school you will find approximately equal numbers of children of various ages included, so a polygon of the children's ages would be basically rectangular. If you took random samples of 40 each from a school with equal numbers of children aged 6 through 11 years, you would find many samples with a mean age near the population mean of 8.5 years, sample means of approximately 8 or 9 would be less common, and sample means as low as 7 or as high as 10 would be rare. Note that the word error in this context does not mean "mistake"—it refers to what is unaccounted for.

### Standard Error of the Mean

Because the extent and the distribution of sampling errors can be predicted, researchers can use sample means with predictable confidence to make inferences concerning population means. However, you need an estimate of the magnitude of the sampling error associated with the sample mean when using it as an estimate of the population mean. An important tool for this purpose is the standard error of the mean. Sampling error manifests itself in the variability of sample means. Thus, if you calculate the standard deviation of a collection of means of random samples from a single population, you would have an estimate of the amount of sampling error. It is possible, however, to obtain this estimate on the basis of only one sample. We have noted that two things affect the size of sampling error: the size of the sample and the standard deviation in the population. When both of these are known, you can predict the standard deviation of sampling errors. This expected standard deviation of sampling errors of the mean is called the **standard error of the mean** and is represented by the symbol $\sigma_{\bar{X}}$. Deductive logic shows that the standard error of the mean is equal to the standard deviation of the population ($\sigma$) divided by the square root of the number in each sample ($\sqrt{n}$). As a formula,

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \tag{7.1}$$

where

$\sigma_{\bar{X}}$ = standard error of the mean
$\sigma$ = standard deviation of the population
$n$ = number in each sample

In Chapter 6, we noted that standard deviation ($\sigma$) is an index of the degree of spread among individuals in a population. In the same way, standard error of the mean ($\sigma_{\overline{X}}$) is an index of the spread expected among the means of samples drawn randomly from a population. As you will see, the interpretation of $\sigma$ and $\sigma_{\overline{X}}$ is very similar.

Because the means of random samples have approximately normal distributions, you can also use the normal curve model to make inferences concerning population means. Given that the expected mean of sample means is equal to the population mean, that the standard deviation of these means is equal to the standard error of the mean, and that the means of random samples are distributed normally, you can compute a ***z* score** for a sample mean and refer that $z$ to the normal curve table to approximate the probability of a sample mean occurring through chance that far or farther from the population mean. The $z$ is derived by subtracting the population mean from the sample mean and then dividing this difference by the standard error of the mean:

$$z = \frac{\overline{X} - \mu}{\sigma_{\overline{x}}} \tag{7.2}$$

To illustrate, consider a college admissions officer who wonders if her population of applicants is above average on the verbal subtest of the College Board examination. The national mean for College Board verbal scores is 500, and the standard deviation is 100. She pulls a random sample of 64 from her population and finds the mean of the sample to be 530. She asks the question, How probable is it that a random sample of 64 with a mean of 530 would be drawn from a population with a mean of 500? Using Formula 7.1, the admissions officer calculates the standard error of the mean as 12.5:

$$\begin{aligned}\sigma_{\overline{x}} &= \frac{\sigma}{\sqrt{n}} \\ &= \frac{100}{\sqrt{64}} \\ &= 12.5\end{aligned}$$

Calculating the $z$ score for her sample mean with Formula 7.2, she obtains the following result:

$$\begin{aligned}z &= \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} \\ &= \frac{530 - 500}{12.5} \\ &= 2.4\end{aligned}$$

Thus, the sample mean deviates from the population mean by 2.4 standard error units. What is the probability of having a sample mean that deviates by this amount (2.4 $\sigma_{\overline{X}}$) $\overline{X}$ or more from the population mean? It is only necessary to refer to the normal curve table in order to express this deviation ($z$) in terms of probability. Referring to the normal curve table, the admissions officer finds that the probability of a $z = 2.4$ or higher is .0082. This means that a $z$ score that great or greater would occur by chance only approximately 8 times in 1000. Because the probability of getting a sample mean that far

from the population mean is remote, she concludes that the sample mean probably did not come from a population with a mean of 500, and therefore the mean of her population—applicants to her college—is very probably greater than 500.

## THE STRATEGY OF INFERENTIAL STATISTICS

Inferential statistics is the science of making reasonable decisions with limited information. Researchers use what they observe in samples and what is known about sampling error to reach fallible but reasonable decisions about populations. The statistical procedures performed before these decisions are made are called **tests of significance.** A basic tool of these statistical tests is the null hypothesis.

### THE NULL HYPOTHESIS

Suppose you have 100 fourth-graders available to participate in an experiment concerning the teaching of certain number concepts. Furthermore, suppose that your research hypothesis is that method B of teaching results in a greater mastery of these concepts than method A. You randomly assign 50 students to be taught these concepts by method A and the other 50 to be taught by method B. You arrange their environment in such a way that the treatment of the two groups differs only in method of instruction. At the end of the experiment, you administer a measure that is considered to be a suitable operational definition of mastery of the number concepts of interest. You find that the mean for the students taught by method B is higher than the mean for those taught by method A. How do you interpret this difference?

Assuming you have been careful to make the learning conditions of the two groups equivalent, except for the method of teaching, you could account for the difference by deciding that (1) the method of teaching caused the difference or (2) the difference occurred by chance. Although the subjects were randomly assigned to the treatments, it is possible that through chance the method B group had students who were more intelligent, more highly motivated, or for some other reason were more likely to learn the number concepts than the students in the method A group, no matter how they were taught.

The difference between the groups therefore could be a result of (1) a relationship between the variables, method of teaching, and mastery of the concepts or (2) chance alone (sampling error). How are you to know which explanation is correct? In the ultimate sense, you cannot positively prove that the method of teaching caused the difference. However, you can estimate the likelihood of chance alone being responsible for the observed difference and then determine which explanation to accept as a result of this estimate.

The chance explanation is known as the **null hypothesis,** which, as you recall from Chapter 5, is a statement that there is *no* actual relationship between the variables and that any observed relationship is only a function of chance. In the example, the null hypothesis would state that there is no relationship between teaching method and mastery of the number concepts.

Another way of stating the null hypothesis in the example is to declare that the mean for all fourth-graders taught by method A is equal to the mean for

all fourth-graders taught by method B. In formula form, using the symbol $\mu$ for population mean, this statement becomes

$$H_0: \mu_A = \mu_B \tag{7.3}$$

where

$H_0$ = null hypothesis
$\mu_A$ = mean of all fourth-graders taught by method A
$\mu_B$ = mean of all fourth-graders taught by method B

Note that the assumption is made that the 50 pupils taught by method A are a representative sample of the population of fourth-graders who might be taught by method A and the 50 pupils taught by method B are a representative sample of the population who might be taught by method B. The investigator hopes to use the data from the experiment to infer what would be expected when other fourth-graders are taught by method A or B.

In interpreting the observed difference between the groups, the investigator must choose between the chance explanation (null hypothesis) and the explanation that states there is a relationship between variables (research hypothesis)—and must do so without knowing the ultimate truth concerning the populations of interest. This choice is based on incomplete information and is therefore subject to possible error.

## TYPE I AND TYPE II ERRORS

The investigator will either retain or reject the null hypothesis. Either decision may be correct or wrong. If the null hypothesis is true, the investigator is correct in retaining it and in error in rejecting it. The rejection of a true null hypothesis is labeled a **Type I error.** If the null hypothesis is false, the investigator is in error in retaining it and correct in rejecting it. The retention of a false null hypothesis labeled a **Type II error.** The four possible states of affairs are summarized in Table 7.2.

Let us consider some possible consequences of the two types of errors, using the example we have been following.

### Type I Error

A Type I error, symbolized by a Greek lowercase letter alpha ($\alpha$), is a "false alarm"—the investigator thinks he or she has something when there is nothing there. For example, the investigator rejects the null hypothesis and declares that there is a relationship between teaching method and mastery of the numerical

**Table 7.2** Schematic Representation of Type I and Type II Error

| | | Real Situation in the Population | |
|---|---|---|---|
| | | $H_0$ is true | $H_0$ is false |
| **Investigator's Decision after Making Test of Significance** | Rejects $H_0$ | Type I error | Correct |
| | Retains $H_0$ | Correct | Type II error |

concepts and therefore recommends method B as the better method. Schools discard textbooks and other materials based on method A and purchase materials based on method B. In-service training is instituted to train teachers to teach by method B. After all this expenditure of time and money, the schools do not observe an increase in mastery of the numerical concepts. Subsequent experiments do not produce the results observed in the original investigation. Although the ultimate truth or falsity of the null hypothesis is still unknown, the evidence supporting it is overwhelming. The original investigator is embarrassed and humiliated.

### Type II Error

A Type II error, symbolized by a lowercase Greek letter beta (β), is a "miss"—the investigator concludes there is nothing when there really is something. In the example, the investigator concludes that the difference between the two groups may easily be attributed to chance and that the null hypothesis may well be true. She thus retains the null hypothesis and declares that there is insufficient evidence for concluding that one method is better than the other. Subsequent investigators conclude that method B is better than method A, and schools that change from method A to method B report impressive gains in student mastery. Although the ultimate truth still remains unknown, a mountain of evidence supports the research hypothesis. The original investigator is embarrassed (but probably not humiliated). Figure 7.1 provides a way of remembering the difference between Type I and Type II errors.

**A fable for the remembrance of Type I and Type II errors**

*by Donald Ary and John W. Sloan*

A beautiful maiden was married to Alpha, the handsome and favorite number one son of the king. Alas, she discovered her husband was impotent. The king's ugly, despised number two son, Beta, became her lover.

In order to escape being stoned to death, she had to declare to the world that with number one, Alpha, there was a relationship when in truth there was none. Concerning number two, Beta, she had to declare that there was no relationship when in truth there was.

The Gods took pity on her and forgave both Type I and Type II errors. They promised her immortality with these words: "Our revered statisticians shall sing of your tragedy to generation after generation throughout all time to come, and the understanding of your dilemma shall cause scholars of all nations to sigh."

*Source:* From *The CEDR Quarterly* (1977). 10. (4) p. 19

**Figure 7.1**

### Comparison of Type I and Type II Errors

Type I errors typically lead to changes that are unwarranted. Type II errors typically lead to a maintenance of the status quo when a change is warranted. The consequences of a Type I error are generally considered more serious than the consequences of a Type II error, although there are certainly exceptions.

## LEVEL OF SIGNIFICANCE

Recall that all scientific conclusions are statements that have a high probability of being correct rather than statements of absolute truth. How high must the probability be before an investigator is willing to declare that a relationship between variables exists? In other words, how unlikely must the null hypothesis be before a researcher rejects it? The consequences of rejecting a true null hypothesis, a Type I error, vary with the situation. Therefore, investigators usually weigh the relative consequences of Type I and Type II errors and decide, before conducting their experiments, how strong the evidence must be before they would reject the null hypothesis. This predetermined level at which a null hypothesis would be rejected is called the **level of significance.** The probability of a Type I error is directly under the control of the researcher, who sets the level of significance according to the type of error he or she wishes to guard against.

Of course, a researcher could avoid Type I errors by always retaining the null hypothesis or avoid Type II errors by always rejecting it. Neither of these alternatives is productive. If the consequences of a Type I error would be very serious but a Type II error would be of little consequence, the investigator might decide to risk the possibility of a Type I error only if the estimated probability of the observed relationship's being caused by mere luck is 1 chance in 1000 or less. This is testing the hypothesis at the .001 level of significance, which is considered to be a quite conservative level. In this case, the investigator is being very careful not to declare that a relationship exists when there is no relationship. However, this decision means accepting a high probability of a Type II error, declaring there is no relationship when in fact a relationship does exist.

If the consequences of a Type I error are judged to be not serious, the investigator might decide to declare that a relationship exists if the probability of the observed relationship's being caused by mere luck is 1 chance in 10 or less. This is called "testing the hypothesis at the .10 level of significance." Here, the investigator is taking only moderate precautions against a Type I error but is not taking a great risk of a Type II error.

The *level of significance* is the probability of a Type I error that an investigator is willing to risk in rejecting a null hypothesis. It is symbolized by the lowercase Greek alpha ($\alpha$). If an investigator sets the level of significance at .01, it means that the null hypothesis will be rejected if, under the assumption that chance alone is responsible, the obtained probability is equal to or less than 1 time in 100. The investigator is saying that he or she is willing to limit the probability of rejecting the null hypothesis when it is true (Type I error) to 1 in 100. If the level of significance is set at .0001, the probability of making a Type I error is 1 in

10,000 or less. The most commonly used levels of significance in the behavioral sciences are the .05 and the .01 levels.

Traditionally, investigators determine the level of significance after weighing the relative seriousness of Type I and Type II errors but before running the experiment. If the data derived from the completed experiment indicate that the probability of the null hypothesis being true is equal to or less than the predetermined acceptable probability, the investigators reject the null hypothesis and declare the results to be statistically significant. If the probability is greater than the predetermined acceptable probability, the results are described as nonsignificant—that is, the null hypothesis is retained.

The familiar meaning of the word *significant* is "important" or "meaningful." In statistics, this word means "less likely to be a function of chance than some predetermined probability." Results of investigations can be statistically significant without being inherently meaningful or important.

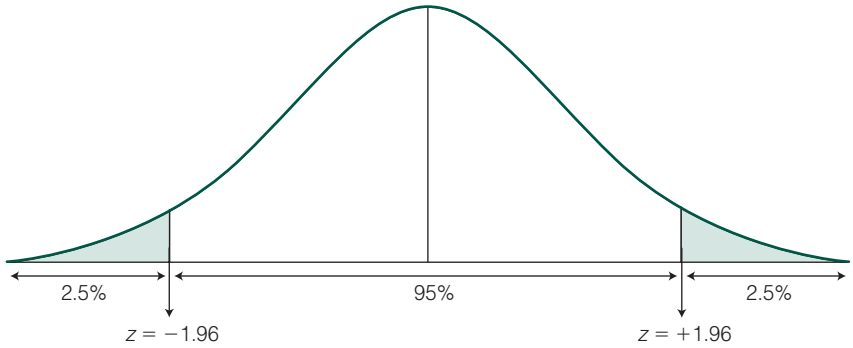
We are sorry the phrase "statistically significant" was chosen to indicate when results were less likely to be chance than a previously chosen $\alpha$. If a phrase such as "statistically not chance" had been chosen, people would not think "statistically significant" had anything to do with importance.

## DIRECTIONAL AND NONDIRECTIONAL TESTS

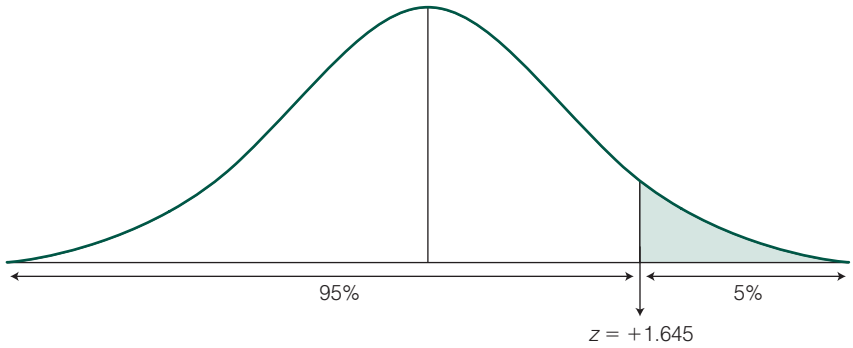
In testing a null hypothesis, researchers are not usually concerned with the direction of the differences. Rather, they are interested in knowing about the possible departure of sample statistics from population parameters. When comparing the effectiveness of competing treatments, an investigator usually wants to learn if treatment A is better than treatment B or if treatment B is better than treatment A. This kind of test is called **nondirectional** (two-tailed) because the investigator is interested in differences in either direction. The investigator states only that there will be a difference. Note in Figure 7.2 that the region of rejection is equally divided between the two tails of the distribution. Thus, if a sample mean is observed that is *either* sufficiently greater or sufficiently less than the hypothesized value, the null hypothesis would be rejected. The direction of the difference is not important.

However, if only one alternative to the null hypothesis is of interest, a **directional test** (one-tailed) is used. For example, an investigator studying the effects of a specific diet among obese people would only be interested in assessing the probability that the diet reduces weight. Or, if a basketball coach is investigating the effects of a practice regimen on the success rate for shooting from the foul line, he or she would only implement the regimen if there is reasonable evidence that it increases success. Results indicating that the regimen decreases success would lead to the same decision as would a retained null hypothesis: Forget the regimen.

If on the basis of experience, previous research, or theory the researcher chooses to state the direction of possible differences, then he or she would perform a directional test. A directional hypothesis would state *either* that the parameter is greater than *or* that the parameter is less than the hypothesized




A. Curve showing the critical region for a nondirectional test (two-tailed test)




B. Curve showing the critical region for a directional test (one-tailed test)

**Figure 7.2** Curves Showing the Critical Regions of Directional and Nondirectional Tests

value. Thus, in directional tests the critical region is located in only one of the two tails of the distribution. For a .05 level of significance, this region in a normal curve is the point equal to $z = 1.645$. You do not, as you would in nondirectional tests, divide the 5 percent between the two sides of the curve. Rather, you place the whole 5 percent of chance error on one side of the curve. This means that the null hypothesis will be retained unless the observed difference is in the hypothesized direction. It is obvious that for rejecting a null hypothesis at a given level, a directional test requires a smaller $z$ value than a nondirectional test (compare $z = 1.645$ with $z = 1.96$, the $z$ value required for a two-tailed test). Therefore, a directional test makes it easier to reject the null hypothesis and thus increases the probability that the null hypothesis will be rejected if the difference is in the hypothesized direction.

In statistical terminology, a nondirectional test is often referred to as a **two-tailed test of significance,** and a directional test is referred to as a **one-tailed test.** The decision to use either a one-tailed test or a two-tailed test should be made early in the study, before any statistical tests are performed. You do not wait to see what the data look like and then select a one-tailed or two-tailed test.

## DETERMINING THE APPROPRIATE SAMPLE SIZE

Now that we have considered Type I and Type II errors and level of significance, let us return to the question of the size of the sample needed in a research study. A scientific method of determining the sample size needed is to specify a meaningful **effect size** ($\Delta$ or $d$) and then determine the sample size needed to reach a desired probability of rejecting the null hypothesis at a given level of significance. Recall that effect size is the difference between experimental and control groups divided by the standard deviation of the control group ($\Delta$) or the difference between two groups divided by the estimated population standard deviation ($d$).

To illustrate, let us return to the college admissions officer who wonders if her population of applicants is above average. To determine the number of subjects needed to test her hypothesis, she first specifies what effect size would be meaningful. She decides that an effect size of .33 (one-third of a standard deviation) or more would be meaningful, but an effect size of less than .33 would not be meaningful.

The specification of what is a meaningful effect size is a judgment call. However, professionals in their fields are usually able to specify an effect size that serves as a reasonable dividing line between meaningful and trivial differences. Cohen's (1988) verbal labels for effect sizes (see Chapter 6) are useful guides for this purpose. In this example, an effect size of .33 on the verbal subtest of the College Board examination might be considered a reasonable dividing line. It is stating that a mean score of 533 or greater [500 + (.33)(100)] on the verbal subtest of the College Board examination is enough to matter, whereas a smaller effect size has little or no practical importance. Consulting the normal curve table (Table A.1 in the Appendix), we see that a $z$ score of .33 has 62.93 percent below it. Therefore, the admissions officer is stating that if approximately 63 percent of the general population have scores less than the average for her population of applicants, the difference is considered meaningful, whereas less difference would not be considered meaningful.

Next, the admissions officer determines her level of significance, designated as $\alpha$, and her desired probability of rejecting the false null hypothesis $(1 - \beta)$. Let us say that she decides she wants a 90 percent chance of rejecting the null hypothesis with a one-tailed .05 $\alpha$. Now she has all the ingredients needed to determine the needed sample size, as shown by the following formula:

$$N = \left(\frac{1}{\Delta}\right)^2 (z\alpha + z\beta)^2 \tag{7.4}$$

where

$N$ = number needed in the sample
$\Delta$ = specified effect size
$z\alpha$ = $z$ score for the level of significance
$z\beta$ = $z$ score for the desired probability of rejecting the null hypothesis $(1 - \beta)$

In Table A.1, you find that the $z$ score for a one-tailed .05 $\alpha$ is 1.645. The $z$ score for a one-tailed probability of 90 percent (the desired probability of rejecting the

null hypothesis) is 1.28. Substituting these $z$ values and the specified $\Delta$, you have the following:

$$\begin{aligned} N &= \left(\frac{1}{.33}\right)^2 (1.645 + 1.28)^2 \\ &= 3^2(2.925)^2 \\ &= 9(8.56) \\ &= 77.04 \end{aligned}$$

With a random sample of 78 subjects from her population of applicants, the admissions officer has a 90 percent chance of rejecting the null hypothesis at the one-tailed .05 level if the true effect size in her population is .33 or greater.

Note that as effect size gets smaller, the number needed gets larger. If the investigator had wanted an effect size of .10 to have a 90 percent probability of rejecting the null hypothesis at the one-tailed .05 level, she would need 856 subjects in her sample. Note also that as $(1 - \beta)$, the desired probability of rejecting the null hypothesis gets larger and as $\alpha$ gets smaller (more demanding), the number of subjects needed gets larger.

Determining the number needed in a sample is really a function of how precise you want to be—that is, how large or small an effect size you want to be statistically significant, how much chance of Type I error you are willing to live with, and how much probability of rejecting a false null hypothesis you want. These are all judgment calls, but they can all be made on a rational basis. We would defend the specifications of $\Delta - .33$, $\alpha = .05$, and $(1 - \beta)$ .90 in the example as reasonable. Therefore, we consider a sample size of 78 sufficient in this case. In our college admissions officer example, a one-sample test of a hypothesized mean is an easy to describe but seldom used statistical test. Each statistical test has its own formula for determining needed sample size.

Cohen's 1988 book is the standard reference for determining the minimum sample size for the different statistical tests. Once you have determined which statistical test you should use, consult Cohen's book to determine the minimum sample size you will need. You may also find Table 2 in Olejnik's 1984 article useful for finding minimum sample size in various situations.

In summary, four elements determine the minimum sample size needed: (1) the statistical test to be used, (2) the acceptable probability of a type I error, (3) the effect size separating meaningful from trivial effect size (e.g., $\Delta$), and (4) the desired probability $(1 - \beta)$ of rejecting the mull hypothesis if the specified population parameter is true.

In Chapter 12, we show the calculations required to determine needed sample size for a specified Pearson *r*. In Chapter 14, we show sample size calculations needed for specified proportions.

## POWER

**Power** is the ability to reject a null hypotheses when it is false. Formula 7.4 is an example of a **power formula.** It indicates the number of subjects needed to give the investigator the desired power to reject the null hypothesis for specified effect size and level of significance. Many statistics books describe calculations

for determining numbers needed when effect size is defined in terms of proportion, correlation, or other statistics. Cohen's *Statistical Power Analysis* (1988) provides extensive information on these procedures.

## THINK ABOUT IT 7.1

Complete each line with either (a) goes up, (b) goes down, or (c) stays the same.

I. As you choose to put α up from .01 to .05,
   1. Probability of Type I error _____.
   2. Probability of Type II error _____.
   3. Power _____.

II. As you increase the number in the samples,
   1. Probability of Type I error _____.
   2. Probability of Type II error _____
   3. Power _____.

III. As the true difference between means goes up from 3 to 7,
   1. Probability of Type I error _____.
   2. Probability of Type II error _____.
   3. Power _____.

IV. As effect size increases,
   1. Probability of Type I error _____.
   2. Probability of Type II error _____.
   3. Power _____.

V. As the heterogeneity (variance) within the samples increases,
   1. Probability of Type I error _____.
   2. Probability of Type II error _____.
   3. Power _____.

VI. If you do a one-tailed test instead of a two-tailed test and if you correctly predicted the direction of the difference
   1. Probability of Type I error _____.
   2. Probability of Type II error _____.
   3. Power _____.

**Answer**

I. 1. a
   2. b
   3. a

II. 1. c
   2. b
   3. a

III. 1. c
   2. b
   3. a

IV. 1. c
   2. b
   3. a

V. 1. c
   2. a
   3. b

IV. 1. c
   2. b
   3. a

# THE GENERAL STRATEGY OF STATISTICAL TESTS

A statistical test compares what is observed (a statistic) with what we would expect to observe through chance alone. What we would expect through chance alone is called the error term. A ratio is formed:

$$\frac{\text{observation}}{\text{chance expectation}} = \frac{\text{statistic}}{\text{error term}}$$

When the observed statistic is equal to or less than the average value expected through chance alone (the **error term**), the most plausible explanation for the statistic is that it was due to chance alone. If the statistic is greater than the error term, then the chance explanation becomes less and less plausible as this ratio becomes greater and greater than 1.

## THE *t* TEST FOR INDEPENDENT SAMPLES

In our math concepts example, the statistic is the difference between the mean of the group taught by method B and the group taught by method A ($\overline{X}_B - \overline{X}_A$). Through deductive logic statisticians have determined the average difference between the means of two randomly assigned groups that would be expected through chance alone. This expected value (the error term) is derived from the variance within each of the two groups and the number of subjects in each of the two groups. It is called the standard error of the difference between two independent means ($s_{\bar{x}_1 - \bar{x}_2}$). Its definition formula is

$$s_{\bar{x}^1 - \bar{x}^2} = \sqrt{\frac{\Sigma x_1^2 + \Sigma x_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{7.5}$$

where

$s_{\bar{x}_1 - \bar{x}_2}$ = standard error of the difference between two means
$n_1$ = number of cases in group 1
$n_2$ = number of cases in group 2
$\Sigma x_1^2$ = sum of the squared deviation scores in group 1
$\Sigma x_2^2$ = sum of the squared deviation scores in group 2

The standard error of the difference between two means is sometimes referred to as the "error term for the independent *t* test."

The ***t* test for independent samples** is a straightforward ratio that divides the observed difference between the means by the difference expected through chance alone. In formula form,

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\bar{x}_1 - \bar{x}_2}} \tag{7.6}$$

If this *t* ratio is equal to 1.00 or less, the observed difference between means is very probably due to chance alone. Recall the investigator doing the math concepts teaching study comparing a randomly assigned group that did method A with a randomly assigned group that did method B set his $\alpha$ at .05. Suppose he found that the group 1 mean was 2.05 higher than the group 2 mean and his error term was 2.85. The observed/chance expectation ratio is .83. The observed difference is less than the difference expected by chance. Therefore, the null

**Table 7.3** Computing the *t* Value for Two Independent Means

| Group 1 Frequent Feedback | | Group 2 Infrequent Feedback | |
|---|---|---|---|
| $X_1$ | $X_1^2$ | $X_2$ | $X_2^2$ |
| 18 | 324 | 15 | 225 |
| 17 | 289 | 13 | 169 |
| 16 | 256 | 12 | 144 |
| 16 | 256 | 12 | 144 |
| 16 | 256 | 11 | 121 |
| 15 | 225 | 11 | 121 |
| 15 | 225 | 10 | 100 |
| 15 | 225 | 10 | 100 |
| 14 | 196 | 10 | 100 |
| 14 | 196 | 10 | 100 |
| 13 | 169 | 9 | 81 |
| 12 | 144 | 8 | 64 |
| 11 | 121 | 7 | 49 |
| 10 | 100 | 6 | 36 |
| 8 | 64 | 6 | 36 |
| $\frac{\Sigma X_1}{n_1} = \frac{210}{15}$ | $\Sigma X_1^2 = 3046$ | $\frac{\Sigma X_2}{n_2} = \frac{150}{15}$ | $\Sigma X_2^2 = 1590$ |

hypothesis is retained. There is not sufficient evidence to draw a tentative conclusion.

A physical education teacher conducted an experiment to determine if archery students perform better if they get frequent feedback concerning their performance or do better with infrequent feedback. She randomly divided her class into two groups of 15 and flipped a coin to determine which group got frequent feedback and which group got infrequent feedback. She set her $\alpha$ at .05 for a two-tailed test. At the end of her study, she administered a measure of archery performance. The results are shown in Table 7.3.

The computation formula for the independent *t* test is

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\Sigma X_1^2 - (\Sigma X_1)^2/n_1 + \Sigma X_2^2 - (\Sigma X_2)^2/n_2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{7.7}$$

Inserting the numbers from Table 7.3 into this formula gives us

$$t = \frac{(210/15) - (150/10)}{\sqrt{\frac{3046 - \frac{(210)^2}{15} + 1590 - \frac{(150)^2}{15}}{15 + 15 - 2}\left(\frac{1}{15} + \frac{1}{15}\right)}}$$

$$= \frac{14 - 10}{\sqrt{\frac{3046 - \frac{44100}{15} + 1590 - \frac{22500}{15}}{28}(.0667 + .0667)}}$$

$$= \frac{4}{\sqrt{\frac{(3046-2940)+(1590-1500)}{28}(.1333)}}$$

$$= \frac{4}{\sqrt{\left(\frac{196}{28}\right)(.1333)}}$$

$$= \frac{4}{\sqrt{(7)(.1333)}} = \frac{4}{\sqrt{.9331}} = \frac{4}{.966} = 4.14$$

Here, we have an observed difference that is 4.14 times as large as the average difference expected by chance. Is it large enough to reject the null hypothesis? To answer this question, we must consider the *t curves* and *degrees of freedom*.

## THE *t* DISTRIBUTIONS

For generations, researchers used the normal curve to assess the probability of an observation/chance expectation ratio. Consulting Table A.1, we find that a ratio of 4.0 has only 3 chances in 100,000 of occurring by chance.

In the 18th century, a quality control officer for Guinness brewery who used the pen name "Student" showed that the normal curve was appropriate for assessing observation/chance expectation ratios only when the population standard deviation formula, $\sigma = \sqrt{\Sigma x^2/N}$, was known. In most research, the population standard deviation is not known and must be estimated by $s = \sqrt{\Sigma x^2/(N-1)}$. When $s$ is used instead of $\sigma$, each statistical test has its unique probability distribution based on the number of subjects and the number of population estimates used, its degrees of freedom.

## DEGREES OF FREEDOM

The number of **degrees of freedom (*df*)** is the number of observations free to vary around a constant parameter. To illustrate the general concept of degrees of freedom, suppose a teacher asks a student to name any five numbers that come into his or her mind. The student would be free to name any five numbers he or she chooses, so we would say that the student has 5 degrees of freedom. Now suppose the teacher tells the student to name five numbers but to make sure that the mean of these five numbers is 20. The student now is free to name any numbers for the first four, but for the last number he or she must name the number that will make the total for the five numbers 100 in order to arrive at a mean of 20. If the student names, as the first four numbers, 10, 16, 20, and 35, then the fifth number must be 19. The student has five numbers to name and one restriction, so his or her degrees of freedom are $5 - 1 = 4$. We can show this as a formula:

$$\begin{aligned} df &= n - 1 \\ &= 5 - 1 \\ &= 4 \end{aligned} \tag{7.8}$$

Now suppose the teacher asks the student to name seven numbers in such a way that the first three have a mean of 10 and all seven have a mean of 12. Here, we have seven numbers and two restrictions, so

$$\begin{aligned} df &= n - 2 \\ &= 7 - 2 \\ &= 5 \end{aligned}$$

When the unknown population standard deviation is estimated from the sample standard deviation, $s = \sqrt{\Sigma x^2/(N-1)}$, one degree of freedom is lost. The one degree of freedom is lost because the sample statistic ($s$) is derived from the deviations about the sample mean that must always sum to 0. Thus, all but one of the deviations is free to vary, or $df = N - 1$.

The concept of degrees of freedom is involved in most of the procedures in inferential statistics. An appropriate method of computing the degrees of freedom is associated with each procedure. Often, the number of degrees of freedom on which a sample statistic is based depends on the sample size ($N$) and the number of sample statistics used in its calculation. In the archery example, the error term was calculated using the deviation of the frequent feedback scores around the frequent feedback mean and the deviation of the individual infrequent feedback scores around the infrequent feedback mean. The degrees of freedom for the $t$ test for independent means are $n_1 + n_2 - 2$. In our example, $df = 15 + 15 + 2 = 28$. Each degree of freedom has its own probability distribution, its $t$ curve. As the degrees of freedom become larger, the $t$ distribution becomes more like the normal curve. The most frequently needed areas of the ***t* curves** are to be found in Table A.2 in the Appendix. The $t$ curve does not approach the baseline as rapidly as does the normal curve. Some of the $t$ curves are shown in Figure 7.3 along with the normal curve, the solid line labeled $df = \infty$. The $t$ curves are labeled according to their degrees of freedom.

Now we can use Table A.2 in the Appendix to determine the significance of the results. The first column in this table is labeled *df* (degrees of freedom). For the archery example, consult the row for 28 *df*. The remaining columns show the $t$ values associated with certain probabilities for directional and nondirectional tests. Because the independent variable in this example could affect archery performance in either a positive or a negative direction, we need to perform a nondirectional test. In the row for 28 *df*, we find 1.701 in the column labeled .10 for a nondirectional test, which tells you that with a true null

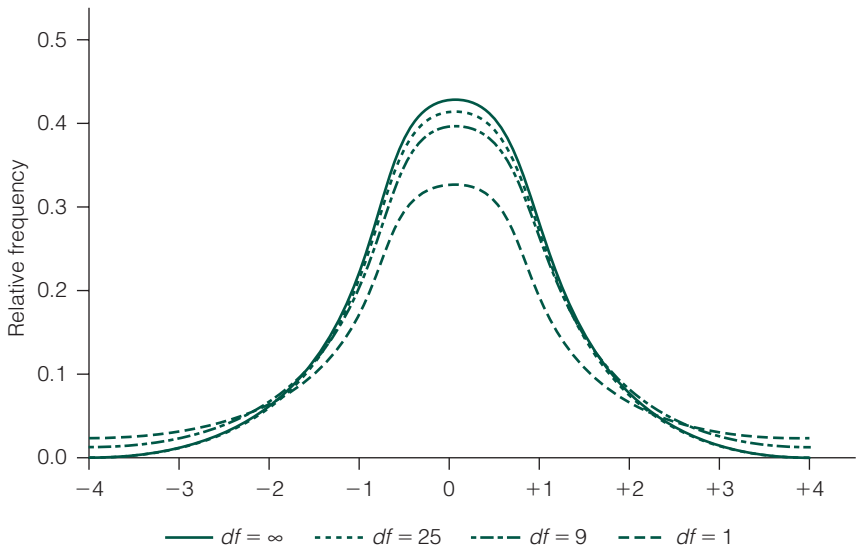



**Figure 7.3** *t* Curves for Various Degrees of Freedom

hypothesis and 28 *df*, a *t* ratio of 1.701 or more or 1.701 or less will occur by chance 1 time in 10. The number 2.048 in the column labeled .05 indicates that under a true null hypothesis and 28 *df*, a *t* ratio of ±2.048 or more will occur by chance 5 percent of the time.

The observed ratio of 4.14 is greater than 2.048, which means that the difference between the groups is greater than the value required to reject the null hypothesis at the .05 level of significance. The estimated probability of the null hypothesis being true is less than 5 percent ($p < .05$). Thus, we would reject the null hypothesis. Although we do not know for certain that the variables frequent feedback versus infrequent feedback and archery performance are related, the evidence is significant enough according to our previously set criteria to enable us to conclude that the observed relationship is probably not just a chance occurrence.

If the observed *t* ratio had been less than 2.048, we would have concluded that the evidence was not good enough to lead us to declare that a relationship exists between the variables. In other words, we would have retained the null hypothesis.

Notice that as we proceed from left to right in the *t* table we find the *t* values required for rejecting the null hypothesis at increasingly rigorous levels of significance. For 28 *df*, a value of 2.763 or greater would lead to the rejection of a null hypothesis at the .01 level. A value of 3.674 or greater would lead to the rejection of the null hypothesis at the .001 level. So the value of 4.14 is significant not only at the .05 level ($p < .05$) but also at the .01 level ($p < .01$) and the .001 level ($p < .001$). If the ratio of observed difference (numerator) divided by error term (denominator) equals or exceeds the value indicated in Table A.2, you can reject the null hypothesis at the indicated level of significance. If you want to report an effect size, first calculate Cohen's *d* and then multiply it by the square root of the sample size. With this approach, it is easy to see the relationship of effect size and number, on the one hand, and statistical significance, on the other hand. The larger the effect size, the larger the *t* and the more likely that the results are statistically significant.

With two independent samples when the dependent variable is ranked data, the Mann–Whitney test serves the same purpose as the *t* test for independent samples. For example, if archery proficiency had been assessed by having a judge rank the 30 subjects from 1st to 30th, the Mann–Whitney test (a nonparametric test) would be used to test the hypothesis.

Recall that much data that appear to be interval, such as grade point averages, actually do not meet the requirement of equal intervals between scores. This would probably be the case if a judge rated archery performance on a scale of 1 to 10. In such a case, it would be advisable to convert the 1 to 10 scores to ranks and do a Mann–Whitney test. Siegel and Castellan's (1988) book is an excellent, very easy to read, text about statistical tests for ordinal and nominal data, including the Mann–Whitney test.

## THE *t* TEST FOR DEPENDENT SAMPLES

So far, our discussion has centered on comparing the means obtained from two independent samples. In an independent sample, each member is chosen randomly from the population and the composition of one group has no bearing on

the composition of the other group. Sometimes, however, investigators may wish to match the subjects of their two groups on some qualities that are important to the purpose of their research, or they may wish to compare the means obtained from the same subjects under two different experimental conditions. In such cases, the groups are no longer independent, inasmuch as the composition of one group is related to the composition of the other group. You would expect the dependent variable scores to be correlated. Therefore, the ***t* test for dependent samples** must be used. This test is also known as the correlated, or nonindependent, or paired *t* test. The measure to be analyzed by the dependent *t* test is the mean difference between the paired scores. Pre- and posttest scores of the same individuals are an example of paired scores.

Let us consider an example. Suppose you wish to know whether taking a research course affects the attitudes of the students toward research. To investigate this, you select a research class and obtain attitude measures toward research from the students on the first and last days of class. Let us suppose you have collected such data, and the results are as presented in Table 7.4.

Columns 2 and 3 of Table 7.4 show the scores of each student in the first and second measures. Column 4 presents the difference between the first and second scores of each student. The sum of these differences amounts to +30. The mean of the differences, 2, is found by dividing +30 ($\Sigma D$) by the number of paired observations, or 15. Column 5 shows the squares of the differences.

**Table 7.4** Before and After Attitude Scores of 15 Students in an Introduction to Research Class

| (1) Subject Number | (2) Pretest | (3) Posttest | (4) $D$ | (5) $D^2$ |
|---|---|---|---|---|
| 1 | 10 | 12 | +2 | +4 |
| 2 | 9 | 13 | +4 | +16 |
| 3 | 8 | 12 | +4 | +16 |
| 4 | 11 | 9 | −2 | +4 |
| 5 | 10 | 8 | −2 | +4 |
| 6 | 7 | 9 | +2 | +4 |
| 7 | 10 | 12 | +2 | +4 |
| 8 | 9 | 11 | +2 | +4 |
| 9 | 8 | 10 | +2 | +4 |
| 10 | 6 | 10 | +4 | +16 |
| 11 | 10 | 12 | +2 | +4 |
| 12 | 7 | 13 | +6 | +36 |
| 13 | 10 | 6 | −4 | +16 |
| 14 | 9 | 13 | +4 | +16 |
| 15 | 10 | 14 | +4 | +16 |
| | | | $\Sigma D = +30$ | $\Sigma D^2 = 164$ |

The standard deviation of the difference scores $s_D$ is

$$s_D = \sqrt{\frac{\Sigma D^2 - \frac{(\Sigma D)^2}{n}}{N-1}} \tag{7.9}$$

In the example, this is

$$s_D = \sqrt{\frac{164 - \frac{(30)^2}{15}}{15-1}} = \sqrt{\frac{164-60}{14}} = \sqrt{\frac{104}{14}} = \sqrt{7.4286} = 2.7255$$

Effect size, the mean difference divided by the standard deviation ($d = \overline{D}/s_D$), is 2/2.7255 = .7338. Since there is only one sample standard deviation, $d$ and $d$ are equal. By Cohen's (1988) definition, this is nearer a "large" (.8) than a "moderate" (.6) effect size.

To determine the likelihood that the effect size is a function of chance, first calculate a $t$ ratio by multiplying effect size by the square root of the number of pairs:

$$t = \Delta\sqrt{N}$$

In this example, the $t$ is $.7338\sqrt{15} = (.7338)(3.87) = 2.84$. The $t$ ratio tells you that the observed difference is 2.84 times as great as the difference that would be expected under a true null hypothesis. Now consult the table of $t$ values (Appendix Table A.2) to determine the statistical significance of the observed ratio. A more commonly used formula for the dependent $t$ test is

$$t = \frac{\overline{D}}{\sqrt{\frac{\Sigma D^2 - \frac{(\Sigma D)^2}{N}}{N(N-1)}}} \tag{7.10}$$

where

$t$ = $t$ ratio
$\overline{D}$ = average difference
$\Sigma D^2$ = different scores squared, then summed
$(\Sigma D)^2$ = difference scores summed then squared
$N$ = number of pairs

The number of degrees of freedom for the dependent $t$ test equals $N-1$, with $N$ being the number of *pairs* of observations. In the example, you have $15-1=14$ degrees of freedom. In the table of $t$ values, you find that with 14 degrees of freedom a $t$ value of 2.145 is needed for the $t$ to be significant at the .05 level, and a $t$ value of 2.977 is needed for significance at the .01 level when a nondirectional test is performed. The obtained value of 2.84 exceeds the given value for the .05 level but does not reach the given value for the .01 level. This means that the difference between the two means is significant at the .05 level but not at the .01 level. If we had set our level of significance at .05, we could conclude that the attitude of the students toward research had changed.

When the dependent variable is rank data, the Wilcoxon paired sample test serves the same purpose as the dependent $t$ test. This test can also be found in Siegel and Castellan (1988).

## THE *t* TEST FOR PEARSON *r* CORRELATION COEFFICIENTS

Another important use for the *t* test is in testing hypotheses concerning a population correlation rho (ρ). The most common null hypothesis in such cases is that the population correlation is zero and that the correlation observed in the sample (*r*) is a function of chance. For example, an investigator might draw a sample of 27 college freshmen, administer vocabulary and spatial orientation tests to them, and find a **Pearson *r*** of .20 between the two measures. The next step is to decide whether this observed correlation coefficient could easily be a result of chance in a population where the true correlation in the population (ρ) is zero. To test the hypothesis that the population correlation is zero, you do not have to calculate a *t* test. These calculations have been done for various sample sizes and sample correlations and are shown in Table A.3 (critical values of Pearson's *r*).

Whereas with the dependent *t* test the degrees of freedom are the number of paired observations minus 1, it has been shown mathematically that with the Pearson *r* the degrees of freedom are the number of paired observations minus 2. A significant *r* is equal to or larger than the tabled value with $N-2$ degrees of freedom, where *N* is the number of *pairs* of scores. Table A.3 indicates that with $df = 25$ (the number in the sample minus 2), when a two-tailed test is performed, an observed Pearson *r* above +.3809 or less than −.3801 is required to reject the null hypothesis at the two-tailed .05 level. Thus, the correlation of .20 obtained in the study involving college freshmen is not significant at the .05 level.

With a reasonably large number of cases, a coefficient of correlation may be relatively low in value and yet be statistically significant (although not necessarily important). Assume that a study with a sample of 292 sixth-graders found a correlation of +.25, a correlation statistically significant at the .01 level, between the students' response on a questionnaire measuring students' satisfaction with school and their teachers' prediction of how these students would respond to the questionnaire. The statistically significant findings indicate that in the population represented by the sample, the correlation is not likely to be 0. However, recall that the coefficient of determination ($r^2$) indicates the extent of relationship between variables. Only $.25^2$, or 6.25 percent, of the variance of teachers' predictions is in common with the students' actual responses. Although teachers can predict student satisfaction at greater than chance level, the teachers' predictions of satisfaction have only a weak relationship with actual student satisfaction. Statistical significance does not necessarily mean important.

The *t* test can also be used to test hypotheses about population correlations other than 0. You can also use it to test the hypothesis that the correlations observed in two samples could have arisen from the same population. Because this is an introductory text, we have chosen not to include such tests here. A useful description of these tests may be found in Glass and Hopkins (1996) and in various other texts.

## ANALYSIS OF VARIANCE

**Simple** or **one-way analysis of variance (ANOVA)** is a statistical procedure used to analyze the data from a study with more than two groups. The null hypothesis is that there is no difference among the group means. It is called one-way ANOVA because there is only one independent variable and one dependent variable. In

analysis of variance, as in the *t* test, a ratio comparing observed differences to the error term is used to test hypotheses about differences among groups. This ratio, called the ***F* ratio,** employs the variance ($\sigma^2$) of group means as a measure of observed differences among groups. The *F* ratio is named for R. A. Fisher, the early statistician who developed it. Because ANOVA can be used with more than two groups, it is a more versatile technique than the *t* test. A *t* test can be used only to test a difference between *two* means. ANOVA can test the difference between *two or more* means.

The general rationale of ANOVA is that the *total variance* of all subjects in an experiment can be subdivided into two sources: *variance between groups* and *variance within groups*. Variance between groups is incorporated into the numerator in the *F* ratio. Variance within groups is incorporated into the error term or denominator, as it is in the *t* test. As variance between groups increases, the *F* ratio increases. As variance within groups increases, the *F* ratio decreases. The number of subjects influences the *F* ratio: The larger the number, the larger the numerator becomes. When the numerator and denominator are equal, the differences between group means are no greater than would be expected by chance alone. If the numerator is greater than the denominator, you consult the table of *F* values (Table A.4 in the Appendix) to determine whether the ratio is great enough to let you reject the null hypothesis at the predetermined level.

### Computing the *F* Ratio (simple analysis of variance)

Suppose you have the three experimental conditions of high stress, moderate stress, and no stress, and you wish to compare the performance on a simple problem-solving task of three groups of individuals, randomly assigned to these three conditions. Assume that the data presented in Table 7.5 summarize your

**Table 7.5** Measures Obtained in Three Random Samples after Performance of a Task under Conditions of High Stress, Moderate Stress, and No Stress

| Group 1 High Stress | | Group 2 Moderate Stress | | Group 3 No Stress | |
|---|---|---|---|---|---|
| $X_1$ | $X_1^2$ | $X_2$ | $X_2^2$ | $X_3$ | $X_3^2$ |
| 19 | 361 | 22 | 484 | 15 | 225 |
| 18 | 324 | 20 | 400 | 14 | 196 |
| 17 | 289 | 19 | 361 | 14 | 196 |
| 16 | 256 | 18 | 324 | 13 | 169 |
| 15 | 225 | 17 | 289 | 13 | 169 |
| 15 | 225 | 16 | 256 | 12 | 144 |
| 14 | 196 | 16 | 256 | 12 | 144 |
| 13 | 169 | 15 | 225 | 11 | 121 |
| 12 | 144 | 14 | 196 | 11 | 121 |
| 11 | 121 | 12 | 144 | 10 | 100 |
| $\Sigma X_1 = 150$ | $\Sigma X_1^2 = 2310$ | $\Sigma X_2 = 169$ | $\Sigma X_2^2 = 2935$ | $\Sigma X_3 = 125$ | $\Sigma X_3^2 = 1585$ |
| $\overline{X}_1 = 15.0$ | | $\overline{X}_2 = 16.9$ | | $\overline{X}_3 = 12.5$ | $\overline{X} = 14.8$ |
| $\Sigma \overline{\overline{X}}_t = 144$ | | | | $\Sigma X_t^2 = 6830$ | |

observations of the performance of these three groups and that you are now going to test the null hypothesis that there is no significant difference among these three groups at the .01 level.

The means can be seen to differ from one another and from the overall mean for all 30 subjects ($\overline{\overline{X}}$,the *grand mean* or overall mean). Are the differences among these means great enough to be statistically significant, or is it likely that they occurred by chance? To answer this, one can compute the *F* ratio by following the required steps.

The first step is to find the sum of the squared deviation of each of the individual scores from the grand mean. This index is called the "total sum of squares" and reflects all treatment effects plus sampling error. It is expressed by the following formula:

$$SS_t = \Sigma X_t^2 - \frac{(\Sigma X_t)^2}{N} \tag{7.11}$$

where

$SS_t$ = sum of squares total
$\Sigma X^2$ = each score squared, then summed
$(\Sigma X_t)^2$ = all the scores summed first, then this sum squared
$N$ = number of scores

(Note that *SS* is the same as $\Sigma X^2$, which you first encountered in Chapter 6. We do not know why *SS* is used instead of $\Sigma X^2$ in analyses of variance, but it is standard operating procedure.)

In this example, this value is

$$SS_t = 6830 - \frac{(444)^2}{30} = 258.8$$

Next, find the part of the total sum of squares that is due to the deviations of the group means from the grand mean. This index is called the "sum of the squares between groups." (To be grammatically correct, we should say "the sum of squares *among* groups" when more than two groups are involved. However, it is a long-standing tradition to use the term "sum of squares between groups," and to be consistent with other texts we are retaining this usage here.) This index is found by applying the formula

$$SS_b = \Sigma x_b^2 = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} \cdots - \frac{(\Sigma X_2)^2}{N} \tag{7.12}$$

In our example, this value is

$$SS_b = \Sigma X_b^2 = \frac{(150)^2}{10} + \frac{(169)^2}{10} + \frac{(125)^2}{10} - \frac{(444)^2}{30} = 97.4$$

Then, you find the part of the total sum of squares that is caused by the deviations of each individual score from its own group mean. This index, called the "sum of the squares within groups," is found by applying the raw score formula for the sum of squared deviations to each group and then summing across groups ($SS_w = SS_1 + SS_2 + \cdots$).

The computation formula is

$$SS_w = \Sigma x_w^2 = \left[\Sigma X_1^2 - \frac{(\Sigma X_1)^2}{n_1}\right] + \left[\Sigma X_2^2 - \frac{(\Sigma X_2)^2}{n_2}\right] + \cdots \quad (7.13)$$

In the example, this value is

$$SS_w = \Sigma x_w^2 = \left[2310 - \frac{(150)^2}{10}\right] + \left[2935 - \frac{(169)^2}{10}\right] + \left[1585 - \frac{(125)^2}{10}\right] = 161.4$$

A shortcut to finding the sum of the squares within groups is to subtract the sum of squares between groups from the total sum of the squares, expressed as follows:

$$SS_w = SS_t - SS_b \quad (7.14)$$

In the example,

$$SS_w = 258.8 - 97.4 = 161.4$$

The sum of squares between groups is a systematic variation, possibly due to effects of treatment. The sum of squares within groups is associated with spread within groups (error variance). If the treatment had an effect, the between-group variability will be larger than the within-group variability.

An approach to conceptualizing analysis of variance is to view $SS_t$ as reflecting what you observe, $SS_b$ (sum of squares between) as reflecting what you can explain, and $SS_w$ (sum of squares within) as what you cannot explain. The next step is to determine if our systematic knowledge is enough greater than what would be expected by chance to reject the null hypothesis.

### The *F* Test of Significance

Table 7.6 summarizes the results of the calculations so far, together with the results of further calculations. Column 1 of the table lists the three sources of variation: between-groups variance, within-groups variance, and total variance. Column 2 contains the sums of squares, which you have already calculated. Column 3 lists the number of degrees of freedom associated with each source of variance. The number of degrees of freedom for between-groups variance is equal to $(G - 1)$, where $G$ is the number of groups. In our example, this value is $3 - 1 = 2$. The number of degrees of freedom for within-groups variance is $[N - G]$, total number of scores $[N]$ minus the number of groups $[G]$ or $30 - 3 = 27$. The number of degrees of freedom for total variance equals $N - 1$; in our example, $30 - 1 = 29$. This last value could also be obtained by adding the between-groups and within-groups degrees of freedom.

**Table 7.6** Summary of the Analysis of Variance of the Three Groups

| (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|
| Source of Variance | *SS* | *df* | *MS* | *F* | Level of Significance |
| Between groups | 97.4 | 2 | 48.70 | 8.14 | 0.01 |
| Within groups | 161.4 | 27 | 5.98 | | |
| Total | 258.8 | 29 | | | |

The next step is to calculate the *between-groups mean square* ($MS_b$) and the *within-groups mean square* ($MS_w$) by dividing the between-groups and within-groups sums of squares by their respective degrees of freedom. The resulting mean squares are independent estimates of the parent population variance. It has been shown mathematically that if the null hypothesis is true, the population variance estimates mean square between and mean square within will be approximately equal, and the ratio $MS_b/MS_w$ will be approximately 1. As differences among group means increase, the value of $MS_b$ increases. As the number of subjects increases, the value of $MS_w$ decreases. Therefore, as the ratio $MS_b/MS_w$ (called the $F$ ratio) increases, the likelihood that group differences were a function of chance decreases. In the example, the mean square between groups is $97.4/2 = 48.7$ and the mean square within groups is $161.4/27 = 5.98$. The mean square within groups is the error term for your $F$ ratio. By applying the following formula, you finally arrive at the end product of the ANOVA procedure, the $F$ ratio:

$$F = \frac{MS_b}{MS_w} = \frac{SS_b/df_b}{SS_w/df_w} \tag{7.15}$$

In your example, this value is

$$F = \frac{48.70}{5.98} = 8.14$$

The $MS_b$ is 8.14 times what we would expect through chance alone.

## THINK ABOUT IT 7.2

Match $SS_w$, $SS_b$, $SS_t$, $MS_b$, and $MS_w$ to the following phrases:

1. The numerator in the $F$ ratio_____.
2. All the variation _____.
3. Systematic variation _____.
4. Unexplained variation _____.
5. A population variance estimate based on differences among group means _____.
6. The degrees of freedom is $N - G$ for _____.

**Answer**

1. $MS_b$, 2. $SS_t$, 3. $SS_b$, 4. $SS_w$, 5. $MS_b$, 6. $SS_w$

Now consult Table A.4 in the Appendix to determine whether the $F$ ratio is statistically significant. You find the column headed by the between-groups (numerator) degrees of freedom and go down this column to the row entry corresponding to the number of within-groups (denominator) degrees of freedom. At this point in the column you find two values, one in lightface type and one in boldface type. If the $F$ ratio is equal to or greater than the value given in lightface, it is significant at the .05 level. If the obtained $F$ ratio is equal to or greater than the value given in boldface, it is also significant at the .01 level. In the example,

with 2 and 27 degrees of freedom you need an $F$ ratio of 3.35 to reject the null hypothesis at the .05 level and an $F$ ratio of 5.49 to reject the null hypothesis at the .01 level. Because the obtained $F$ ratio is greater than both of these values, it is significant at the .01 level and the null hypothesis is rejected at that level.

A significant $F$ ratio does not necessarily mean that all groups differ significantly from all other groups. An overall significant $F$ indicates that at least one out of all possible comparisons between pairs of means or combinations of means will be significant. The significant $F$ may be a result of a difference between one group and the rest of the groups. For instance, in your problem it might be that group 3 is significantly different from group 1 and group 2, but groups 1 and 2 do not differ significantly from each other. In some cases, it is obvious which pairs differ just from looking at the data. However, in most cases, we need further statistical tests called post hoc comparisons to indicate the means that are significantly different from other means. Two of the most widely used are Tukey's and Scheffe's tests. Many statistics texts describe these tests and their applications.

If the dependent variable is in the form of ranks, the Kruskal–Wallis one-way analysis by ranks test can assess the statistical significance of differences among groups. This test may also be found in the Siegel and Castellan (1988) text.

### Effect Size

The effect size estimate to use in ANOVA is called $\eta^2$. Recall that sum of squares ($SS$) between groups is a systematic variation, believed to be due to the effects of the treatment. If we divide $SS$ between groups by the total sum of squares, we determine the proportion of the total variability of the dependent variable that is accounted for by the independent variable (the treatment). This index is called $\eta^2$.

$$\eta^2 = \frac{SS_b}{SS_t} = \frac{97.4}{258.8} = .38$$

If we multiply .38 by 100, we could say that 38 percent of the variation in the dependent variable is influenced by the independent variable.

The following is a guideline for interpreting $\eta^2$:

$\eta^2 = .01$ is a small effect.
$\eta^2 = .06$ is a medium effect.
$\eta^2 = .14$ is a large effect.

Thus, we have a large effect size in this study.

## MULTIFACTOR ANALYSIS OF VARIANCE

In the complex world we live in, it often is not the case that an independent variable has a consistent influence on a dependent variable. Often, $X$ influences $Y$ in certain circumstances but not in others. For instance, phonics instruction is useful in teaching reading with children with normal hearing but worthless with the deaf. The independent variable, method of teaching reading, does not have the same effect on the dependent variable, reading proficiency, when the variable normal hearing versus deaf is taken into account.

In the language of research we say, "Is there an *interaction* between the independent variable method of teaching reading ($X_1$) and the independent variable normal hearing vs. deaf ($X_2$) and the dependent variable reading proficiency ($Y$)?"

You may want to investigate the combined effect of stress level and need for achievement on performance in a problem-solving task. To investigate this problem, you will vary both the level of stress and the achievement need. The layout for an experiment investigating the combined effects of two or more independent variables is called a **factorial design,** and the results are analyzed by means of a **multifactor analysis of variance.** If two independent variables are investigated, we call the analysis a two-way analysis of variance. The null hypothesis for a two-way ANOVA is that the population means are equal.

Let us assume that you have carried out this experiment using five subjects in each group and that the data shown in Table 7.7 represent a summary of your observations of the performance of the subjects. Applying multifactor ANOVA will enable you to learn (1) whether there is a significant difference between the performance of all the subjects under a high-stress condition and all those under a low-stress condition, (2) whether there is a significant difference between the performance of all the subjects with high achievement need and all those with low achievement need, and (3) whether the difference between performance under high- and low-stress conditions is the same for both subjects with high need for achievement and subjects with low need for achievement. The effects of each independent variable (stress and motivation level) on the dependent variable are called **main effects,** whereas the effect of different combinations of the two independent variables on the dependent variable is referred to as an **interaction effect.** The end products of these analyses will be three $F$ ratios, two of which indicate the significance of the two main effects and the third indicates that of the interaction effect.

**Table 7.7** Measures on Problem-Solving Tasks of Subjects with Low and High Achievement Need under High and Low Conditions of Stress

| Achievement Need | Stress: High | Stress: Low | |
|---|---|---|---|
| High | 20<br>20 Group 1<br>19<br>19 $\overline{X} = 19$<br>17<br>$\Sigma X$ 95 | 23<br>22 Group 3<br>21<br>20 $\overline{X} = 21$<br>19<br>$\Sigma X$ 105 | $\Sigma X_{r1} = 200$<br>$\overline{X} = 20.0$ |
| Low | 22<br>21 Group 2<br>20<br>19 $\overline{X} = 20$<br>18<br>$\Sigma X$ 100 | 18<br>16 Group 4<br>15<br>14 $\overline{X} = 15$<br>12<br>$\Sigma X$ 75 | $\Sigma X_{r2} = 175$<br>$X_{r2} = 17.5$ |
| | $\Sigma X_{c_1} = 195$<br>$\overline{X}_{c_1} = 19.5$ | $\Sigma X_{c_2} = 180$<br>$\overline{X}_{c_2} = 18.0$ | $\Sigma X$ Total = 375<br>$\overline{\overline{X}}$ (grand mean) = 18.75 |

The computation of these $F$ ratios involves the following steps:

1. Find the total sum of squares, the sum of squares between groups, and the sum of squares within groups using the same procedures and formulas applied in simple ANOVA. These values, derived from the data in Table 7.7, are as follows:

$$SS_t = 7181 - \frac{(375)^2}{20} = 149.75$$
$$SS_b = \frac{(95)^2}{5} + \frac{(100)^2}{5} + \frac{(75)^2}{5} + \frac{(105)^2}{5} - \frac{(375)^2}{20} = 103.75$$
$$SS_w = 149.75 - 103.75 = 46.00$$

2. Partition the sum of the squares between groups into three separate sums of squares—(a) the sum of squares between columns, (b) the sum of squares between rows, and (c) the sum of squares for interaction between columns and rows—as shown next:

   a. The between-columns sum of squares represents the sum of the squared deviations caused by the difference between the column means and the grand mean. Find it by using the formula

$$SS_{bc} = \frac{(\Sigma X_{c_1})^2}{n_{c_1}} + \frac{(\Sigma X_{c_2})^2}{n_{c_2}} + \cdots - \frac{(\Sigma X)^2}{N} \tag{7.16}$$

   With this formula, the sum of squares between the columns for the data shown in Table 7.7 is

$$SS_{bc} = \frac{(195)^2}{10} + \frac{(180)^2}{10} - \frac{(375)^2}{20} = 11.25$$

   b. The between-rows sum of squares is the sum of the squared deviations caused by the difference between the row means and the grand mean. Find it by applying the formula

$$SS_{br} = \frac{(\Sigma X_{r_1})^2}{n_{r_1}} + \frac{(\Sigma X_{r_2})^2}{n_{r_2}} + \cdots - \frac{(\Sigma X)^2}{N} \tag{7.17}$$

   For the data presented in Table 7.7, this value is

$$SS_{br} = \frac{(200)^2}{10} + \frac{(175)^2}{10} - \frac{(375)^2}{20} = 31.25$$

   c. The sum of squares interaction is the part of the deviation between the group means and the overall mean that is due neither to row differences nor to column differences. Expressed as a formula, it becomes

$$SS_{int} = SS_b - (SS_{bc} + SS_{br}) \tag{7.18}$$

   Expressed in words, the interaction sum of squares is equal to the between-groups sum of squares minus the sum of the between-columns sum of squares and the between-rows sum of squares.

   For the data presented in Table 7.7, this interaction sum of squares is

$$SS_{int} = 103.75 - (11.25 + 31.25) = 61.25$$

3. Determine the number of degrees of freedom associated with each source of variation. They are found as follows:

$$df \text{ for between-columns sum of squares} = C - 1$$
$$df \text{ for between-rows sum of squares} = R - 1$$
$$df \text{ for interaction} = (C - 1)(R - 1)$$
$$df \text{ for between-groups sum of squares} = G - 1$$
$$df \text{ for within-groups sum of squares} = N - G$$
$$df \text{ for total sum of squares} = N - 1$$

where

$C$ = number of columns
$R$ = number of rows
$G$ = number of groups
$N$ = number of subjects in all groups

4. Find the mean square values by dividing each sum of squares by its associated number of degrees of freedom.
5. Compute the $F$ ratios for the main and the interaction effects by dividing the between-groups mean squares by the within-groups mean square for each of the three components.

The results of the calculations based on the data presented in Table 7.7 are summarized in Table 7.8. Three $F$ ratios are listed in this table. To find the significance of each of these values, consult Table A.4 (the table of $F$ values) as before. To enter this table, use the number of degrees of freedom associated with each $F$ ratio ($df$ for the numerator) and the number of degrees of freedom associated with the within-groups mean square ($df$ for the denominator). For example, the between-columns $F$ ratio is 3.913. Consulting the table, you see that with 1 and 16 $df$, an $F$ ratio of 4.49 or more is needed for significance at the .05 level. Because the $F$ ratio is smaller than the value shown in the table, you can conclude that the high-stress versus low-stress difference is not statistically significant.

Because the between-rows degrees of freedom are the same as the degrees of freedom between columns, the same levels of 4.49 (.05 level) and 8.53 (.01 level) apply. Because the obtained value of 10.869 exceeds both of these values, it is significant at the .01 level.

**Table 7.8** Summary of a 2 × 2 Multifactor Analysis of Variance

| Source of Variance | *SS* | *df* | *MS* | *F* | Level of Significance |
|---|---|---|---|---|---|
| Between columns (stress) | 11.25 | 1 | 11.25 | 3.913 | — |
| Between rows (achievement need) | 31.25 | 1 | 31.25 | 10.869 | .01 |
| Columns by rows (interaction) | 61.25 | 1 | 61.25 | 21.304 | .01 |
| Between groups | 103.75 | 3 | 34.583 | | |
| Within groups | 46.00 | 16 | 2.875 | | |
| Total | 149.75 | 19 | | | |

In the example, the degrees of freedom for interaction are the same (1 and 16). Therefore, the same values of $F$ are required. The $F$ value of 21.304 is greater than 8.35; therefore, the interaction is significant at the .01 level. You can reject the null hypothesis—that the difference between the high-stress minus low-stress group means among the subjects with low need for achievement is the same as the difference between the high-stress minus low-stress group means among the subjects with high need for achievement.

### Interpreting the *F* Ratios

The first $F$ ratio (between columns) in Table 7.8 is not significant and shows that the stress conditions do not differ significantly from one another in their effect on the performance of the subjects in the experiment. This analysis, a comparison of the combined performance of groups 1 and 2 ($\overline{X} = 19.5$) with the combined performance of groups 3 and 4 ($\overline{X} = 18.0$), was not statistically significant, so we do not have sufficient evidence to make any conclusions about the stress conditions.

The second $F$ ratio (between rows), which is significant at the .01 level, is based on comparing the performance of the subjects in groups 1 and 3 ($\overline{X} = 20.0$) with the performance of those in groups 2 and 4 ($\overline{X} = 17.5$). From the significance of this $F$ ratio you can infer that the difference between the performance of those subjects with high achievement need and those with low achievement need is beyond chance expectation. Because you have a significant $F$ ratio for the difference, you conclude that under conditions similar to those of your experiment, a higher level of task performance can be expected from people with high achievement need than people with low achievement need.

The third $F$ ratio shows the interaction effect between the two variables: stress level and achievement need. The significance of the $F$ ratio in this case means that the effect of stress level on performance in a problem-solving task depends on the degree of achievement need.

Among the students with a high need for achievement, the difference between high stress and low stress was only 2 points ($21 - 19$). Among the students with a low need for achievement, you find a difference between high stress and low stress of 5 points ($20 - 15$). Because your $F$ test indicated that the interaction was significant, you can conclude that high stress makes more difference when combined with low achievement need than when combined with high achievement need.

Again, we can calculate estimates of the effect size for the three conditions by using $\eta^2$:

$$\eta^2 \text{ for stress} = 11.25/149.75 = .08$$
$$\eta^2 \text{ for need achievement} = 31.25/149.75 = .21$$
$$\eta^2 \text{ for interaction} = 61.25/149.75 = .41$$

The effect size for stress was small; the effects for need achievement and interaction were large.

Multifactor analysis has been of great value in educational research because many questions educators need to investigate are inherently complex in nature. These techniques enable educators to analyze the combined effects of two or more independent variables in relation to a dependent variable. For example, a simple comparison of the dependent variable means of two groups of pupils

taught by different methods may yield insignificant results. However, if you incorporate aptitude into the experiment as a measured independent variable, you may find that one method works better with the more capable pupils, whereas the other works better with the less capable pupils. (We present more information on interpreting factorial analysis of variance in Chapter 11.) We realize you will not need to calculate "by hand" an *F* ratio or other statistics for your research projects. A computer will calculate them for you in a matter of seconds. Actually working through the calculations, however, will help you to understand exactly what a statistic represents and how to interpret it.

Multifactor analysis of variance is not limited to two independent variables as in our example. Any number of independent variables may be incorporated in this technique. Several intermediate statistics books, including Glass and Hopkins (1996), explain the computation and interpretation of these procedures.

## THE CHI-SQUARE TESTS OF SIGNIFICANCE

When dealing with nominal data, the most widely used **tests of significance** are the **chi-square tests.** They compare **observed frequencies** and **expected frequencies.** Observed frequencies, as the name implies, are the actual frequencies obtained by observation. Expected frequencies are theoretical frequencies that would be observed when the null hypothesis is true.

### The One-Variable Chi Square (Goodness of Fit)

Consider the hypothesis that the proportion of male and female students in advanced placement (AP) classes is different from that of male and female students in a high school as a whole. If you know that 40 percent of the total high school enrollment is male and that 300 students are enrolled in AP courses, the expected frequencies of male and female students enrolled in AP classes will be 120 males and 180 females. Now suppose that the observed frequencies are found to be 140 males and 160 females.

| | *f* Observed | *f* Expected |
|---|---|---|
| Male | 140 | 120 |
| Female | 160 | 180 |
| Total | 300 | 300 |

There is a difference between observed and expected frequencies. Is it likely that this difference is due to chance alone? The null hypothesis is that the observed frequencies will not differ from what we expect from chance.

To determine whether the difference between the expected and observed frequencies is statistically significant, apply the chi-square formula:

$$\chi^2 = \sum\left[\frac{(f_o - f_e)^2}{f_e}\right] \tag{7.19}$$

where

$\chi^2$ = value of chi square
$f_0$ = observed frequency
$f_e$ = expected frequency

Applying this formula to the data, you obtain

$$\chi^2 = \frac{(140-120)^2}{120} + \frac{(160-180)^2}{180} = 5.55$$

To determine whether this chi-square value is significant, consult the table of $\chi^2$ values in the Appendix (Table A.5). The first column in this table shows the number of degrees of freedom involved in any given chi-square problem. The remaining columns present the values needed for different levels of significance. The number of degrees of freedom, as discussed previously, is based on the number of observations that are free to vary once certain restrictions are placed on the data. When you have a fixed number of observations divided into only two categories, as soon as the number falling into one category has been determined, the other is fixed. Thus, when you find that the number of male students is 140, the number of female students in the total of 300 must be 160. In this example, there is only 1 degree of freedom. In the one-variable chi square, the number of degrees of freedom equals $K-1$, where $K$ is the number of categories used for classification. By consulting the table of $\chi^2$ values, you find that your observed value of 5.55 is statistically significant at the .05 (and .02) level. You would reject the null hypothesis that there is no difference in the proportion of male and female students taking AP classes and the proportion of male and female students in the entire high school.

Interpreting this result, you can now state that the proportion of males who take AP classes is significantly greater than the proportion in the entire school at the .05 level of confidence. The significance level of .05 means that there are less than 5 chances in 100 of observing such a difference between the proportions of male and female students through chance alone. Thus, the data lend support to your research hypothesis that the proportion of male and female students who take AP classes is significantly different from the proportion of males and females in the entire high school.

The use of the chi-square test is not limited to situations in which there are only two categories of classification; this test can also be used to test a null hypothesis stating that there is no significant difference between the proportion of the subjects falling into any number of different categories. Suppose, for example, you asked a random sample of 120 college freshmen whether they prefer to live in a dormitory or in town or have no preference, with the results shown in Table 7.9.

If there were no difference between the three categories of response, you would have 40 responses in each category. These are your expected frequencies, as shown in Table 7.9.

A comparison of the two sets of frequencies presented in Table 7.9 shows that there are differences between your expected and observed data. To determine

**Table 7.9** Observed and Expected Frequencies of Responses of 120 Undergraduate Students on Their Preference for Living Accommodations

| | Dormitory | Town | No Preference | Total |
|---|---|---|---|---|
| Observed | 40 | 50 | 30 | 120 |
| Expected | 40 | 40 | 40 | 120 |

whether they are significant, apply the chi-square test. The value of $\chi^2$ for these data, using Formula 7.19, would be

$$\chi^2 = \frac{(40-40)^2}{40} + \frac{(50-40)^2}{40} + \frac{(30-40)^2}{40} = 5.00$$

The degrees of freedom, again, equal the number of categories minus 1 ($K-1$) or, in this case, $3-1=2$. Referring to the table of $\chi^2$ values, you can see that with 2 degrees of freedom, a $\chi^2$ value of 5.991 or greater is required for significance at the .05 level. However, your obtained $\chi^2$ value is smaller than this value and therefore is not statistically significant. This means that the observed differences between categories could easily have happened by chance. Consequently, the null hypothesis—that there is no significant difference between the frequencies of the three categories—cannot be rejected. In other words, if the proportions of preferences for the three categories in the entire freshman population were equal, you would expect to observe sample differences as great as those in your sample more often than 5 times in 100 through chance. We conclude that there is insufficient evidence that the college freshmen have a preference among the living accommodations.

## THINK ABOUT IT 7.3

You want to know if high school students prefer Coca-Cola (Coke) or Pepsi among soft drinks. You randomly select 100 students and ask them to indicate their preference for Coca-Cola or Pepsi. The null hypothesis is that there is no preference between the two; thus you would expect 50 students to choose each brand. In your survey, you find that 60 students say they prefer Coke and 40 prefer Pepsi. Calculate a chi-square for these data:

| | Coke | Pepsi |
|---|---|---|
| Observed | 60 | 40 |
| Expected | 50 | 50 |

$$\text{chi-square} = \frac{(60-50)2}{50} + \frac{(40-50)2}{50} = 100/50 + 100/50 = 2+2 = 4$$

We have two categories, thus $df = 1$. Consulting A.5, we find that the critical value of chi-square at the .05 level is 3.84. Because the calculated chi-square of 4 is larger than the critical value, you would reject the null hypothesis and conclude that high school students prefer Coca-Cola over Pepsi.

### The Two-Variable Chi Square (Chi-Square Test of Independence)

So far, we have only considered examples in which observations were classified along a single dimension. Sometimes, however, researchers wish to use more than one dimension for classification. The two-variable chi-square design uses two independent variables, each with two or more levels, and a dependent variable in the form of a frequency count. The purpose of the test is to determine whether or not the two variables in the design are independent of one another. Suppose, for example, you add another dimension to the previous problem and ask both graduate and undergraduate students to state their preferences regarding the

**Table 7.10** Observed Frequencies of Responses of 200 Undergraduate and Graduate Students on Their Preference for a University Health Plan

| Subjects | Plan A | Plan B | No Preference | Total |
|---|---|---|---|---|
| Undergraduate students | 40 | 50 | 30 | 120 |
| Graduate students | 20 | 40 | 20 | 80 |
| Total | 60 | 90 | 50 | 200 |

two proposed university student health plans. Assume the frequencies shown in Table 7.10 were the result.

In this case, the null hypothesis is that the preference for health plans is the same for graduates as it is for undergraduates—that is, the variables "student status" and "preference for health plan" are unrelated, or independent. The null hypothesis in this **chi-square test of independence** is always that the variables are independent in the population. Your observations show that 30 percent (60/200) of all students prefer Health Plan A, 45 percent (90/200) prefer Plan B, and 25 percent state no preference. If the null hypothesis is true, you would expect to find these same proportions among both graduates and undergraduates.

You can compute expected cell frequencies by multiplying the row frequency associated with a cell by the column frequency associated with that cell and then dividing this product by the grand total ($E = f_r f_c / N$). For example, the expected frequency of response for undergraduate students who favor Plan A is $120 \times 60 \div 200 = 36$; for those undergraduate students who prefer Plan B, it is $120 \times 90 \div 200 = 54$; and for graduate students who prefer Plan A, it is $80 \times 60 \div 200 = 24$. Using this approach, you can find the expected frequencies for each cell as shown in Table 7.11.

Note that all the row and column totals in Table 7.11 are exactly the same as those shown in Table 7.10. Now you ask if the observed frequencies differ enough from the expected frequencies to enable you to reject the likelihood that these differences could have occurred merely by chance. Applying the formula, you obtain

$$\begin{aligned}\chi^2 &= \frac{(40-36)^2}{36} + \frac{(50-54)^2}{54} + \frac{(30-30)^2}{30} + \frac{(20-24)^2}{24} \\ &\quad + \frac{(40-36)^2}{36} + \frac{(20-20)^2}{20} \\ &= 1.8518\end{aligned}$$

The number of degrees of freedom for a two-way table is found by applying the formula

$$df = (C-1)(R-1) \tag{7.20}$$

**Table 7.11** Expected Frequencies of Responses of 200 Undergraduate and Graduate Students Regarding Their Preference for a University Health Plan

| Subjects | Plan A | Plan B | No Preference | Total |
|---|---|---|---|---|
| Undergraduate students | 36 | 54 | 30 | 120 |
| Graduate students | 24 | 36 | 20 | 80 |
| Total | 60 | 90 | 50 | 200 |

where

$df$ = number of degrees of freedom
$C$ = number of columns
$R$ = number of rows

Applying this formula to the problem under consideration, you obtain

$$df = (3-1)(2-1) = 2$$

Referring to Table A.4, you can see that with 2 degrees of freedom, a $\chi^2$ value of 5.991 is needed for significance at the .05 level. But your obtained $\chi^2$ value of 1.8518 is smaller than this tabled value and is therefore not significant. This means that the differences between expected and observed frequencies are not beyond what you would expect by chance. In other words, you do not have reliable evidence that there is a relationship between the variables "student status" and "preference for a health care plan" in the population from which your sample was drawn. The null hypothesis of independence cannot be rejected.

### Assumptions of Chi Square

Chi square is so easy to use that researchers may forget that there are assumptions that must be met if valid interpretations are to be made.

1. Observations must be independent—that is, the subjects in each sample must be randomly and independently selected.
2. The categories must be mutually exclusive: Each observation can appear in one and only one of the categories in the table.
3. The observations are measured as frequencies.

## SUMMARY

Investigators hope to form generalizations about populations by studying groups of individuals selected from populations. These generalizations will be sound only if the selected groups (the samples) used in these studies are representative of the larger groups (the populations) from which they are chosen.

Statisticians distinguish between two major types of sampling procedures: probability sampling and nonprobability sampling. Probability sampling is characterized by random selection of population elements. In nonprobability sampling, the researcher's judgment replaces random selection. Simple random sampling, stratified sampling, cluster sampling, and systematic sampling are forms of probability sampling. In simple random sampling, all members of a population have an equal chance of being included within the sample. In stratified sampling, researchers select independent samples from different subgroups, or strata, of a population. In cluster sampling, researchers randomly select naturally occurring groups, or clusters, from a population; then they use as the sample all individuals within the selected clusters. Finally, in systematic sampling, researchers take as the sample every $K$th case from a list of the population. Forms of nonprobability sampling are convenience sampling, purposive sampling, and quota sampling. In accidental sampling, researchers use the available cases as the sample. In purposive sampling, cases judged as typical of the population of interest constitute the sample. Finally, in quota sampling, quotas are assigned to various strata of a population; then cases judged to be typical of each stratum are selected.

Inferential statistics provide tools by means of which researchers can estimate how confident they can be in inferring that phenomena observed in samples will also appear in the populations from which the samples were drawn. In other words, they tell you how much trust to put in what your observations seem to be telling you.

A basic strategy in inferential statistics is to compute the difference among observations that would be likely to arise by chance alone. The result of this computation is called the error term. Then the observed differences among observations are compared with the error term. If the observed differences are similar to the differences that could arise by chance, the researcher cannot reject the likelihood that the observed differences were merely a function of chance. If the observed differences are greater than the error term, the researcher consults the tabled values of the statistic to determine whether the ratio of observation to error is great enough to reject the chance explanation at a predetermined level of significance. Inferential statistics are used when individuals are randomly assigned to treatments as well as when individuals are randomly selected from a population.

Effect size and sample size determine statistical significance. Power calculations determine the sample size needed to give a specified probability of rejecting the null hypothesis at a specified level of significance for a given effect size.

The indexes most commonly used in inferential statistics are the *t* test, analysis of variance (ANOVA), and the chi-square test. The *t* test is used to determine whether the difference between two sample means is statistically significant. There are three types of *t* tests: (1) the *t* test for independent groups, which is used to compare two sample means (two groups) when the samples have been drawn independently from a population; (2) the *t* test for dependent groups, which is employed with two samples in which the subjects are matched (e.g., twins) or with two repeated measures such as pre- and posttests obtained from the same subjects; and (3) the *t* test for Pearson product moment correlation.

Analysis of variance is used to compare the means of two or more samples and to test the null hypothesis that no significant differences exist between the means of these samples. Multifactor analysis of variance enables researchers to test the effect of more than one independent variable and the interaction effect of such variables. There are other inferential statistics available to test hypotheses when the data are in the form of ranks.

The chi-square statistic is an index employed to find the significance of differences between proportions of subjects, objects, events, and so forth that fall into different categories (nominal data) by comparing observed frequencies and frequencies expected under a true null hypothesis.

## KEY CONCEPTS

accessible population
analysis of variance (ANOVA)
biased sample
chi-square tests
chi-square test of independence
cluster sampling
convenience sampling
critical values for Pearson *r*
degrees of freedom (*df*)
directional test
effect size
error term
$\eta^2$
expected frequency
factorial design
*F* ratio
*F* test
interaction effect
level of significance
main effect
multifactor analysis of variance
nondirectional test
nonprobability sampling
null hypothesis
observed frequency
one-tailed test
parameter
Pearson *r*
population
power calculation
power formulas
probability sampling
proportional stratified sampling
purposive (judgment) sampling
quota sampling
random assignment
random sample
random sampling

sample
sampling errors
simple random sampling
standard error of the mean
statistic
stratified sampling
systematic sampling
table of random numbers
target population
tests of significance
*t* curve
*t* test for dependent samples
*t* test for independent samples
two-tailed test
Type I error
Type II error
*z* score

## EXERCISES

1. Does the accuracy of a sample in representing the characteristics of the population from which it was drawn always increase with the size of the sample? Explain.
2. You have been asked to determine whether teachers in the Springfield School District favor the "year-round school" concept. Because the district is rather large, you are asked to contact only 500 teachers. Determine the number you would choose from each of the following levels to draw a proportional stratified random sample:

| Level | Total Number |
|---|---|
| Elementary | 3500 |
| Middle school | 2100 |
| High school | 1400 |
| Total | 7000 |

3. You are asked to conduct an opinion survey on a college campus with a population of 15,000 students. How would you proceed to draw a representative sample of these students for your survey?
4. A national magazine has 1 million subscribers. The editorial staff wants to know which aspects of the magazine are liked and which are not. The staff decides that a personal interview is the best method to obtain the information. For practical and economic reasons, only 500 people in five cities will be surveyed. In this situation, identify the following:
   **a.** The population
   **b.** The sample
5. Which of the following are probability samples? Which are nonprobability samples?
   **a.** Random sample
   **b.** Convenience sample
   **c.** Cluster sample
   **d.** Stratified sample
   **e.** Purposive sample
   **f.** Quota sample
   **g.** Systematic sample
6. What is the difference between random sampling and random assignment?
7. Do the laws of probability apply in both?
8. Investigators wish to study the question, "Do blondes have more fun?"
   **a.** What is the null hypothesis in this question?
   **b.** What would be a Type I error in this case?
   **c.** What would be a Type II error in this case?
   **d.** If one investigator uses a .05 level of significance in investigating this question and another investigator uses a .001 level of significance, which would be more likely to make a Type I error?
   **e.** If one investigator uses a .05 level of significance in investigating this question and another investigator uses a .001 level of significance, which would be more likely to make a Type II error?
9. Inferential statistics enable researchers to do which of the following?
   **a.** Reach infallible conclusions
   **b.** Reach reasonable conclusions with incomplete information
   **c.** Add an aura of legitimacy to what is really sheer guesswork
10. What two conditions are necessary for a Type I error to occur?
11. Which of the following statements describes the role of the null hypothesis in research?
    **a.** It enables researchers to determine the probability of an event occurring through chance alone when there is no real relationship between variables.
    **b.** It enables researchers to prove there is a real relationship between variables.
    **c.** It enables researchers to prove there is no real relationship between variables.

**12.** A Type II error occurs when a researcher does which of the following?
- **a.** Rejects a false null hypothesis
- **b.** Rejects a true null hypothesis
- **c.** Has already made a Type I error
- **d.** Retains a false null hypothesis
- **e.** Retains a true null hypothesis

**13.** The phrase "level of significance" refers to which of the following situations?
- **a.** The probability of an event being due to chance alone, which is calculated after the data from an experiment are analyzed
- **b.** The probability of a Type I error that an investigator is willing to accept
- **c.** The actual probability of a Type II error
- **d.** The probability of reporting there is no relationship when there really is one

**14.** How do you determine the level of significance to use in an experiment?

**15.** A pharmaceutical company has employed researchers to investigate the frequency of heart attacks among those who have taken their drug regularly and those who have not. Would the manufacturer urge the researchers to be especially careful to avoid making a Type I error or a Type II error?

**16.** What is concluded when the results of a study are not statistically significant?

**17.** Compare stratified sampling with quota sampling.

**18.** What are the three elements to consider in determining sample size?

**19.** How large a sample would be needed for an effect size of .10 or greater to have an 80 percent chance of being statistically significant at the two-tailed .05 level of significance?

**20.** Find the chi-square value for the following set of data. Then indicate if the obtained value of chi square is significant at the .05 level of significance.

| | | |
|---|---|---|
| 2 | 8 | 10 |
| 6 | 4 | 10 |
| 8 | 12 | |

**21.** You have the responses of two groups of students (education and noneducation majors) to an item in an attitude scale. Find the chi-square value for this problem and indicate whether your obtained chi-square value is statistically significant.

| | **Strongly Agree** | **Agree** | **Undecided** | **Disagree** | **Strongly Disagree** |
|---|---|---|---|---|---|
| Education students | 7 | 8 | 10 | 9 | 6 |
| Noneducation students | 8 | 10 | 7 | 8 | 7 |

**22.** For testing each of the following hypotheses, indicate which of the following statistical procedures would be appropriate:

Test for independent means
Test for nonindependent means
Factorial analysis of variance
Chi-square one-variable or goodness-of-fit test
Chi-square test of independence

- **a.** The proportion of students in the class who receive A's will be the same as the proportion of students who receive B's.
- **b.** The mean score of the 1 p.m. section on the final examination will be significantly higher than the mean score of the 7 p.m. section.
- **c.** Attending the evening class instead of an afternoon class is an advantage, grade-wise, for female students and a disadvantage for male students.
- **d.** If the same statistics quiz that was given to students on the first day of class were administered on the last day of class, students would show a significant gain in achievement.
- **e.** There is a relationship between a student's passing or failing the class and whether that student expresses satisfaction or dissatisfaction with the course.
- **f.** Any differences among the mean scores of the fall semester classes and the summer session classes of the same instructor can easily be accounted for by chance.
- **g.** There is a significant interaction effect between teaching experience and research experience and success in a research course.

**23.** Are there statistical tests that can be used when the data are in ranks?

**24.** A directional test differs from a nondirectional test in that in a directional test the researcher is interested in changes that take place in which of the following ways?
- **a.** Only in a positive direction
- **b.** Only in a negative direction
- **c.** In both positive and negative directions
- **d.** Either in a positive or in a negative direction, but not both

**25.** Two randomly selected groups have been used in an experiment in which group I received treatment and group II received no treatment. The researcher's hypothesis is that the mean performance of group I will be higher than the mean performance of group II. Apply the *t* test to the following information and state if the researcher's hypothesis could be supported with a two-tailed .01 $\alpha$.

| | $\overline{X}$ | $N$ | $\Sigma X^2$ |
|---|---|---|---|
| Group I | 45.32 | 30 | 382.02 |
| Group II | 41.78 | 30 | 264.32 |

**26.** A researcher wants to test the hypothesis that the correlation between variable A and variable B is significantly greater than zero. He has obtained an $r = .21$ between the two variables using 22 subjects. Use Table A.8 to determine if the hypothesis can be rejected at the .05 level (two-tailed).

**27.** The following data are for a $2 \times 3$ experimental design. Apply the multifactor analysis of variance to test the significance of the main effects and the interaction effect.

| | Columns | | |
|---|---|---|---|
| **Rows** | **A** | **B** | **C** |
| A | 25, 23, 20, 17, 15 | 22, 20, 18, 16, 14 | 20, 18, 16, 14, 12 |
| B | 16, 14, 12, 10, 8 | 18, 16, 14, 12, 10 | 19, 18, 16, 14, 13 |

Provide a table in which the sums of squares (*SS*), degrees of freedom (*df*), mean squares (*MS*), and *F* values are shown. Then answer the following questions:
- **a.** Which *F* values are significant and at what level?
- **b.** How many null hypotheses are being tested in this problem?
- **c.** How many of these hypotheses can the researcher reject?

**28.** Explain the logic of the *t* test.

**29.** What test of significance would you use to analyze data from the following studies?
- **a.** You want to determine the effectiveness of three different methods of teaching spelling on third-graders' achievement in spelling. Students are randomly assigned to one of the three methods.
- **b.** Many high school students work part-time during the school term. You want to know if the part-time work has an adverse effect on these students' academic performance. You draw a sample of students who work and ask them the number of hours they worked per week during the academic year; you use grade point average as a measure of their academic performance.
- **c.** You want to compare the popularity of three different brands of jeans among high school girls. You draw a random sample of high school girls and ask them their preference.
- **d.** You want to compare the effectiveness of a film on global warming on the attitudes of college students toward environmental issues. You randomly select two groups of college students currently enrolled in introductory psychology. You show your film to one group and an interesting but unrelated film to the other group. Following the film, you compare the two groups on a measure of attitudes toward environmental issues.

**30.** The following is a partial summary table for an ANOVA based on a study with three groups and $N = 30$. Complete the summary table and interpret. Also, calculate $\eta^2$ as a measure of effect size.

| **Source of variance** | *SS* | *df* | *MS* | *F* | **Level of significance** |
|---|---|---|---|---|---|
| Between groups | 87.4 | | | | |
| Within groups | 151.4 | | | | |
| Total | | | | | |

## ANSWERS

**1.** A larger randomly drawn sample is more likely to be representative of the population than is a smaller random sample. A large sample obtained with a method that permits systematic bias will not be any more representative than a small biased sample.

**2.** To obtain a proportional stratified sample, divide the 500 teachers in proportion to their representation in the population as follows:

Elementary $3500/7000 \times 500 = 250$
Middle school $2100/7000 \times 500 = 150$
High school $1400/7000 \times 500 = 100$
Total sample 500

**3.** Obtain a list of all students and assign a number to each one. Then select a random sample of a given number by using a table of random numbers. Starting at a random point in the table, go up or down the column and include those students whose numbers are drawn. You may also use a web-based random number generator.

**4. a.** All subscribers to the magazine
**b.** Five hundred individuals who are interviewed

**5.** Probability samples: a, c, d, g
Nonprobability samples: b, e, f

**6.** In random sampling, a chance procedure is used to select a sample from a list of all those in a population. In random assignment, a chance procedure is used to decide which of the available subjects will get which treatment.

**7.** Yes

**8. a.** There is no relationship between hair color and how much fun a person has.
**b.** The investigators make a Type I error if they declare that blondes have more fun than nonblondes or that blondes have less fun than nonblondes, when in fact the two groups have an equal amount of fun.
**c.** The investigators make a Type II error if they fail to conclude that blondes have more fun or less fun, when in fact they do.
**d.** The investigator with the .05 level of significance
**e.** The investigator with the .001 level of significance

**9.** b

**10.** The null hypothesis must be true, and the investigator must reject it.

**11.** a

**12.** d

**13.** b

**14.** By weighing the consequences of Type I and Type II errors

**15.** A Type I error

**16.** The results could easily be a function of chance: The evidence is insufficient to justify a conclusion.

**17.** In stratified sampling, representativeness in each stratum is achieved through random selection, whereas in quota sampling representativeness in various strata is attained by way of judgment.

**18.** (1) The effect size considered to be the boundary between a meaningful and a nonmeaningful difference, (2) level of significance ($\alpha$), and (3) desired probability of rejecting a false null hypothesis ($\beta$)

**19.** $N = (1/.10)^2\ (1.96 + .84)^2$
$= 100(7.84)$
$= 784$
(Note that with a small effect size a large number is required.)

**20.** 3.33 $df = 1$—nonsignificant

**21.** 0.954 $df = 2$—nonsignificant

**22. a.** Chi-square one-variable fit test
**b.** *t* Test for independent means, or one-way ANOVA
**c.** Factorial analysis of variance
**d.** *t* Test for dependent means
**e.** Chi-square test of independence
**f.** *t* Test for independent means
**g.** Factorial analysis of variance

**23.** Yes

**24.** d

**25.** $t = 4.11$, $df = 58$, significant at .001 level

**26.** With 20 degrees of freedom, an $r$ of .4227 or greater is needed to reject the null hypothesis. Therefore, the null hypothesis is retained.

**27.**

| Source of Variance | *SS* | *df* | *MS* | *F* |
|---|---|---|---|---|
| Between columns | 0 | 2 | 0 | 0 |
| Between rows | 120 | 1 | 120 | 11.34 |
| *Interaction* | *80* | *2* | *40* | *3.78* |
| Between groups | 200 | 5 | 40 | 3.78 |
| Within groups | 254 | 24 | 10.58 | — |

**a.** Between rows at .01 and interaction at .05
**b.** Three
**c.** Two

**28.** Sample statistics are used to determine the difference between two means that is expected due to pure chance when the null hypothesis is true. The observed difference is compared to this expected difference to determine the likelihood that the observed difference is due only to chance.

**29. a.** one-way ANOVA
**b.** Pearson coefficient of correlation
**c.** Chi-square (goodness-of-fit)
**d.** *t* Test for independent means

**30.**

| Source of Variance | *SS* | *df* | *MS* | *F* | Level of Significance |
|---|---|---|---|---|---|
| Between groups | 87.4 | 2 | 43.70 | 7.79 | 0.01 |
| Within groups | 151.4 | 27 | 5.61 | | |
| | 238.8 | 29 | | | |

$$\eta^2 = SS \text{ between}/SS \text{ total}$$
$$= 87.4/238.8$$
$$= .366$$

**Conclusion:** There is a significant difference between the groups on the dependent variable of the study. The effect size would be considered large.

## REFERENCES

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Fisher, R. A. (1925). *Statistical methods for research workers*. London: Oliver & Boyd. [Revised 1970]

Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology*. Englewood Cliffs, NJ: Prentice Hall.

Olejnik, S. F. (1984). Planning educational research: Determining the necessary sample size. *Journal of Experimental Education*, *53*(1), 40–48.

Siegel, S., & Castellan, N. J., Jr. (1988). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

CHAPTER 8

# Tools of Research

Operational definitions transform constructs into observable measures.

## INSTRUCTIONAL OBJECTIVES

After studying this chapter, the student will be able to:

1. Explain the role of measurement in research.
2. Access sources such as *Mental Measurements Yearbook* and *Tests in Print* to obtain information necessary for evaluating standardized tests and other measuring instruments.
3. State the difference between a test and a scale.
4. Distinguish between norm-referenced and criterion-referenced tests.
5. Distinguish between measures of aptitude and achievement.
6. Distinguish between ceiling effect and floor effect and discuss why these may be of concern.
7. Describe the steps to follow in preparing a Likert scale for measuring attitudes.
8. Define performance assessment and discuss its advantages and disadvantages.
9. Describe the characteristics of a bipolar adjective scale.
10. State the kinds of errors that are common to rating scales.
11. State advantages and disadvantages of self-report personality measures.
12. List at least five guidelines that a researcher should follow when using direct observation as a data-gathering technique.
13. Define a situational test, and tell when it might be used in research.
14. State the essential characteristic of a projective technique and name at least two well-known projective techniques.

One aim of quantitative research is to obtain greater understanding of relationships among variables in populations. For example, you might ask, What is the relationship between intelligence and creativity among 6-year-olds? You cannot directly observe either intelligence or creativity. Nor can you directly observe all 6-year-olds. But this does not mean that you must remain in ignorance about this and similar questions. There are observable behaviors that are accepted as being valid indicators of constructs such as intelligence and creativity. Using indicators to approximate constructs is the measurement aspect of research.

Some measurement is very straightforward, using a single indicator to represent a variable. For example, you could measure a person's educational background by asking about the highest grade he or she had completed. Similarly, such variables as grade level, nationality, marital status, or number of children could be measured by a single indicator simply because these variables refer to phenomena that are very clear and for which a single indicator provides an acceptable measure. Other variables, however, are more complex and much more difficult to measure. In these cases, using a single indicator is not appropriate.

Selecting appropriate and useful measuring instruments is critical to the success of any research study. One must select or develop scales and instruments that can measure complex constructs such as intelligence, achievement, personality, motivation, attitudes, aptitudes, interests, and self-concept. There are two basic ways to obtain these measures for your study: Use one that has already been developed or construct your own.

To select a measuring instrument, the researcher should look at the research that has been published on his or her question to determine what other researchers have used to measure the construct of interest. These reports will generally indicate whether the instrument worked well or whether other procedures might be better. Other useful sources for identifying published instruments for one's research purposes are the *Seventeenth Mental Measurements Yearbook* (Geisinger, Spies, Carlson, & Plake, 2007) and a companion volume, *Tests in Print VII* (Murphy, Plake, & Spies, 2006). Each edition of *Tests in Print* provides an index of all known commercially available tests in print at the time, with information on publisher and date of publication. A subject index helps one to locate tests in a specific category. The Buros Center for Testing website (www.unl.edu/buros) allows you to examine a large amount of information on tests and testing. Once you locate an available test, you then consult the *Mental Measurements Yearbook* for more information and a critical review of the test. The "Test Reviews Online," a service of the Buros Center for Testing, provides reviews exactly as they appear in the *Mental Measurements Yearbook* series. Another good source of information about both published and unpublished tests is the Educational Testing Service (ETS) Test Collection. The ETS Test Collection is a library of more than 20,000 commercial and research tests and other measuring devices designed to provide up-to-date test information to educational researchers. It is available on the web (www.ets.org/testcoll). ETS also has the collection *Tests in Microfiche*, which provides not only an index of unpublished tests but also copies of the tests on microfiche.

If researchers cannot find a previously developed instrument, then they must develop their own. The procedure involves identifying and using behavior that can be considered an indicator of the construct. To locate these indicators, researchers should turn first to the theory behind the construct. A good theory generally suggests how the construct will manifest itself and the changes that can be observed; that is, it suggests ways to measure the construct(s). For example, the general ($g$ factor) theory of intelligence influenced the choice of tasks in the construction of early intelligence tests. Shavelson, Huber, and Stanton's (1976) multidimensional theory of self-concept served as the blueprint for a number of self-concept measures that have had a major influence on both theory and classroom practice. For instance, the Shavelson model was the basis for Marsh's (1988)

widely used SDQ (Self-Description Questionnaire), which measures self-concept in preadolescents, adolescents, and late adolescents/young adults. Following construction of an instrument, additional research is used to support or revise both the instrument and the theory upon which it is based. Researchers can also use their own experiences and expertise to decide on the appropriate indicators of the construct. In this chapter, we briefly discuss some of the main types of measuring instruments that are used in educational research: achievement and aptitude tests, personality tests, attitude scales, and observational techniques.

## TESTS

Tests are valuable measuring instruments for educational research. A **test** is a set of stimuli presented to an individual in order to elicit responses on the basis of which a numerical score can be assigned. This score, based on a representative sample of the individual's behavior, is an indicator of the extent to which the subject has the characteristic being measured.

The utility of these scores as indicators of the construct of interest is in large part a function of the objectivity, validity, and reliability of the tests. Objectivity is the extent of agreement among scorers. Some tests, such as multiple-choice and true–false tests, are described as objective because the scoring is done by comparing students' answers with the scoring key, and scorers need make no decisions. Essay tests are less objective because scores are influenced by the judgment and opinions of the scorers. In general, validity is the extent to which a test measures what it claims to measure. Reliability is the extent to which the test measures accurately and consistently. We discuss validity and reliability in Chapter 9.

### ACHIEVEMENT TESTS

**Achievement tests** are widely used in educational research, as well as in school systems. They are used to measure what individuals have learned. Achievement tests measure mastery and proficiency in different areas of knowledge by presenting subjects with a standard set of questions involving completion of cognitive tasks. Achievement tests are generally classified as either standardized or teacher/researcher made.

#### Standardized Tests

**Standardized tests** are published tests that have resulted from careful and skillful preparation by experts and cover broad academic objectives common to the majority of school systems. These are tests for which comparative norms have been derived, their validity and reliability established, and directions for administering and scoring prescribed. The directions are contained in the manuals provided by the test publishers. To establish the norms for these tests, their originators administer them to a relevant and representative sample. The norm group may be chosen to represent the nation as a whole or the state, city, district, or local school. The *mean* for a particular grade level in the sample becomes the norm for that grade level. It is important to distinguish between a norm and a standard. A *norm* is not necessarily a goal or a criterion of what should be. It is a

measure of what *is*. Test norms are based on the actual performance of a specified group, not on standards of performance. The skills measured are not necessarily what "ought" to be taught at any grade level, but the use of norms does give educators a basis for comparing their groups with an estimate of the mean for all children at that grade level. Currently, as part of the accountability movement, standardized tests are being widely used to measure students' achievement. The No Child Left Behind Act of 2001 mandated that states have instruments that ensure accurate measurement of a body of skills and knowledge judged to be important and that the instruments be administered and scored under standardized conditions. The measurement aims to determine the number of students at a particular grade level who know a particular set of facts or are proficient in a particular set of skills. For example, Indiana has the ISTEP (Indiana Student Test of Educational Progress), Illinois has the ISAT (Illinois Standard Achievement Test), and California has the CST (California Standards Test).

Standardized achievement tests are available for single school subjects, such as mathematics and chemistry, and also in the form of comprehensive batteries measuring several areas of achievement. An example of the latter is the California Achievement Test (CAT/5), which contains tests in the areas of reading, language, and mathematics and is appropriate for grades kindergarten to 12. Other widely used batteries include the Iowa Tests of Basic Skills (ITBS), the Metropolitan Achievement Tests (MAT-8), the SRA Achievement Series, and the Stanford Achievement Test Series (SAT-9). Some well-known single-subject achievement tests are the Gates–MacGinitie Reading Test, the Nelson–Denny Reading Test, and the Modern Math Understanding Test (MMUT). If one is interested in measuring achievement in more than one subject area, it is less expensive and time-consuming to use a battery. The main advantage of the test battery is that each subtest is normed on the same sample, which makes comparisons across subtests, both within and between individuals, easier and more accurate.

In selecting an achievement test, researchers must be careful to choose one that is reliable and is appropriate (valid) for measuring the aspect of achievement in which they are interested. There should be a direct link between the test content and the curriculum to which students have been exposed. The test must also be valid and reliable for the type of subjects included in the study. Sometimes a researcher is not able to select the test but must use what the school system has already selected. The *Mental Measurements Yearbooks* present a comprehensive listing, along with reviews of the different achievement tests available.

If an available test measures the desired behavior and if the reliability, validity, and the norms are adequate for the purpose, then there are advantages in using a standardized instrument. In addition to the time and effort saved, investigators realize an advantage from the continuity of testing procedures—the results of their studies can be compared and interpreted with respect to those of other studies using the same instrument.

### Researcher-Made Tests

When using standardized tests of achievement is not deemed suitable for the specific objectives of a research study, research workers may construct their own tests. It is much better to construct your own test than to use an inappropriate

standardized one just because it is available. The advantage of a **researcher-made test** is that it can be tailored to be content specific; that is, it will match more closely the content that was covered in the classroom or in the research study. For example, suppose a teacher wants to compare the effects of two teaching methods on students' achievement in mathematics. Although there are excellent standardized tests in mathematics, they are generally designed to measure broad objectives and may not focus sufficiently on the particular skills the researcher wishes to measure. It would be wise in this case to construct the measuring instrument, paying particular attention to evidence of its validity and reliability. The researcher should administer a draft of the test to a small group who will not participate in the study but who are similar to those who will participate. An analysis of the results enables the researcher to check the test's validity and reliability and to detect any ambiguities or other problems before employing the test. For suggestions on achievement test construction, refer to specialized texts in measurement, such as those by Popham (2005), Thorndike (2005), Kubiszyn and Borich (2006), and Haladyna (2004).

### Norm-Referenced and Criterion-Referenced Tests

On the basis of the type of interpretation made, standardized and **teacher-made tests** may be further classified as **norm-referenced** or **criterion-referenced.** Norm-referenced tests permit researchers to compare individuals' performance on the test to the performance of other individuals. An individual's performance is interpreted in terms of his or her relative position in a specified reference group known as the *normative group*. Typically, standardized tests are norm referenced, reporting performance in terms of percentiles, standard scores, and similar measures.

In contrast, criterion-referenced tests enable researchers to describe what a specific individual can do, without reference to the performance of others. Performance is reported in terms of the level of mastery of some well-defined content or skill domain. Typically, the level of mastery is indicated by the percentage of items answered correctly. For example, a criterion-referenced test might be used to ascertain what percentage of Indiana fourth-graders know the capitals of the 50 states. Predetermined cutoff scores may be used to interpret the individual's performance as pass–fail. The state tests used in the mandated accountability testing programs are criterion referenced. A well-known standardized instrument, the National Assessment of Educational Progress (NAEP), is criterion referenced. It is administered to a national sample of all U.S. schools to measure student knowledge in a wide variety of subject areas.

Before designing a measuring instrument, you must know which type of interpretation is to be made. In norm-referenced tests, items are selected that will yield a wide range of scores. A researcher must be concerned with the range of difficulty of the items and the power of the items to discriminate among individuals. In criterion-referenced tests, items are selected solely on the basis of how well they measure a specific set of instructional objectives. They may be easy or difficult, depending on what is being measured. The major concern is to have a representative sample of items measuring the stated objectives so that individual performance can be described directly in terms of the specific knowledge and skills that these people are able to achieve.

### Test Performance Range

The range of performance that an achievement test permits is important. Researchers want a test designed so that the subjects can perform fully to their ability level without being restricted by the test. Two types of testing effects may occur. A **ceiling effect** occurs when many of the scores on a measure are at or near the maximum possible score. Tests with a ceiling effect are too easy for many of the examinees, and we do not know what their scores might have been if there had been a higher ceiling. For example, if we gave a 60-item test and most of the scores fell between 55 and 60, we would have a ceiling effect. A graph of the frequency distribution of scores would be negatively skewed (see Chapter 6).

Likewise, test performance may be restricted at the lower end of the range, resulting in a **floor effect.** A floor effect occurs when a test is too difficult and many scores are near the minimum possible score. For example, a statistics test administered as a pretest before students had a statistics class would likely show a floor effect. A graph of the frequency distribution of scores would be positively skewed. A test with a floor effect would not detect true differences in examinees' achievement either. Standardized tests typically cover a wide range of student performance, so it is not likely that many students would get all or almost all questions correct (ceiling effect) or almost all questions wrong (floor effect). A researcher should, however, consult the test manual for information about ceiling and floor effects so that he or she can select an instrument that permits a wide range of performance. Test developers gather extensive data on subjects' performance during the test standardization process. Researchers who construct their own tests can try them out with various groups and examine the results for evidence of ceiling and floor effects. If it appears that performance range is restricted, then the researcher needs to revise the test.

### Performance Assessments

Another way to classify achievement tests is whether they are verbal or **performance tests.** The most common achievement tests are paper-and-pencil tests measuring cognitive objectives. This familiar format, usually administered to groups, requires individuals to compose answers or choose responses on a printed sheet. In some cases, however, a researcher may want to measure performance—what an individual can *do* rather than what he or she *knows*. Performance assessment, usually administered individually, is a popular alternative to traditional paper-and-pencil tests among educators. A performance test is a technique in which a researcher directly observes and assesses an individual's performance of a certain task and/or judges the finished product of that performance. The test taker is asked to carry out a *process* such as playing a musical instrument or tuning a car engine or to produce a *product* such as a written essay. The performance or product is judged against established criteria. An everyday example of a performance test is the behind-the-wheel examination taken when applying for a driver's license. A paper-and-pencil test covering knowledge of signs and rules for driving is not sufficient to measure driving skill. In investigating a new method of teaching science, for example, you would want to know the effect of the method not only on students' cognitive behavior but also on their learning of various laboratory procedures and techniques or their ability to complete experiments. In this case, the researcher's test would require the students to perform a real task or

use their knowledge and skills to solve a science problem. Performance assessment is important in areas such as art, music, home economics, public speaking, industrial training, and the sciences, which typically involve individuals' ability to do something or produce something. Portfolios that contain a collection of student work such as poetry, essays, sketches, musical compositions, audiotapes of speeches, and even mathematics worksheets are popular in performance assessments. They provide an opportunity for teachers and researchers to gain a more holistic view of changes in students' performance over time.

*Constructing a Performance Test* To create a performance test, follow these three basic steps:

1. Begin with a clear statement of the objectives and what individuals will be asked to do and the conditions under which the task will be performed. A set of test specifications listing the critical dimensions to be assessed will lead to a more comprehensive coverage of the domain. State whether there will be time limits, whether reference books will be available, and so on.
2. Provide a problem or an exercise that gives students an opportunity to perform—either a simulation or an actual task. All individuals must be asked to perform the same task.
3. Develop an instrument (checklist, rating scale, or something similar) that lists the relevant criteria to use in evaluating the performance and/or the product. Make sure that the same criteria are used for each individual's performance or product.

Performance tests are useful for measuring abilities and skills that cannot be measured by paper-and-pencil tests. However, they are time intensive and thus more expensive to administer and score.

## APTITUDE TESTS

**Aptitude tests** differ from achievement tests in that aptitude tests attempt to measure general ability or potential for learning a body of knowledge and skills, whereas achievement tests attempt to measure the actual extent of acquired knowledge and skills in specific areas. Aptitude tests measure a subject's ability to perceive relationships, solve problems, and apply knowledge in a variety of contexts. Some critics question the distinction made between aptitude and achievement tests. They point out that an aptitude test measures achievement to some extent, and an achievement test has an aptitude element. Aptitude tests were formerly referred to as **intelligence tests**, but the latter term has declined in use because of controversy over the definition of intelligence and because people tend to associate intelligence with inherited ability. Aptitude tests should *not* be considered as measures of innate (or "pure") intelligence. As noted previously, performance on such tests partly depends on the background and schooling of the subject.

Educators have found aptitude tests useful and generally valid for the purpose of predicting school success. Many of the tests are referred to as **scholastic aptitude tests,** a term pointing out specifically that the main function of these tests is to predict school performance. Well-known aptitude tests are the ACT (American College Testing Assessment) and the SAT (Scholastic Assessment Test)

for high school students and the GRE (Graduate Record Exam) and MAT (Miller Analogies Test) for college seniors.

Researchers often use aptitude tests. Aptitude or intelligence is frequently a variable that needs to be controlled in educational experiments. To control this variable, the researcher may use the scores from a scholastic aptitude test. Of the many tests available, some have been designed for use with individuals and others for use with groups.

### Individual Aptitude Tests

The most widely used individually administered instruments for measuring aptitude are the Stanford–Binet Intelligence Scale (4th ed.) and the three Wechsler tests. The Stanford–Binet currently in use is the outcome of several revisions of the device first developed in France in 1905 by Alfred Binet and Theodore Simon for identifying children who were not likely to benefit from normal classroom instruction. It was made available for use in the United States in 1916. This test originally reported an individual's mental age. Later, the concept of *intelligence quotient* (IQ) was introduced. This quotient was derived by dividing mental age (MA) by chronological age (CA) and multiplying the result by 100. The present revision of the Stanford–Binet no longer employs the MA/CA ratio for determining IQ. The IQ is found by comparing an individual's performance (score) with norms obtained from his or her age group through the use of standard scores (see Chapter 6). The latest revision of the test has 15 subtests organized into four areas: Verbal Reasoning, Quantitative Reasoning, Abstract/Visual Reasoning, and Short-Term Memory. The scores on the 15 subtests are standard scores with a mean of 50 and a standard deviation of 8. The four area scores and the total IQ score all have a mean of 100 and standard deviation of 16. The Stanford–Binet is appropriate for ages 2 years through adult.

The tests David Wechsler developed to measure aptitude now come in several forms: the Wechsler Intelligence Scale for Children—Third Edition (WISC–III, 1991), the Wechsler Adult Intelligence Scale-III (WAIS–III, 1997), and the Wechsler Preschool and Primary Scale of Intelligence-Revised (WPPSI–R, 1989), which was introduced for the 4 to 6½-year age group. The Wechsler tests yield verbal IQ scores, performance IQ scores, and full-scale IQ scores derived by averaging the verbal subtest scores, the performance subtest scores, and all subtest scores, respectively. The Wechsler scales are more popular than the Stanford–Binet primarily because they require less time to administer.

### Group Tests of Aptitude

A Stanford–Binet or Wechsler test must be given by a trained psychometrician to an individual subject, a procedure expensive in both time and money. Thus, they are impractical as aptitude measures for large groups of individuals. In this situation, group tests are used. The first group test of mental ability was developed during World War I for measuring the ability of men in military service. One form of this test, the Army Alpha, was released for civilian use after the war and became the model for a number of group tests. Today, many group tests of mental aptitude are available. Among the most widely used are the Cognitive Abilities Tests (CogAT), Test of Cognitive Skills (TCS/2), and the Otis–Lennon School Ability Tests (OLSAT-7). The CogAT and the OLSAT-7 are appropriate for grades kindergarten to 12, whereas the TCS/2 is used for grades 2 to 12.

## TESTING AND TECHNOLOGY

New technologies are presenting opportunities for alternatives to paper-and-pencil tests. For example, the PRAXIS I test designed to assess basic skills prior to entry into teacher education is given electronically with immediate scoring and feedback on performance provided to the examinee. A computer is also used to administer the GRE and a number of other well-known tests. Many of you may have encountered computer-based testing when you took the knowledge portion of your test to obtain a driver's license.

## MEASURES OF PERSONALITY

Educational researchers often use measures of personality. There are several different types of personality measures, each reflecting a different theoretical point of view. Some reflect trait and type theories, whereas others have their origins in psychoanalytic and motivational theories. Researchers must know precisely what they wish to measure and then select the instrument, paying particular attention to the evidence of its validity. Two approaches are used to measure personality: objective personality assessment and projective personality assessment.

### OBJECTIVE PERSONALITY ASSESSMENT

**Self-report inventories** present subjects with an extensive collection of statements describing behavior patterns and ask them to indicate whether or not each statement is characteristic of their behavior by checking *yes*, *no*, or *uncertain*. Other formats use multiple choice and true–false items. The score is computed by counting the number of responses that agree with a trait the examiner is attempting to measure. For example, someone with paranoid tendencies would be expected to answer *yes* to the statement "People are always talking behind my back" and *no* to the statement "I expect the police to be fair and reasonable." Of course, similar responses to only two items would not indicate paranoid tendencies. However, such responses to a large proportion of items could be considered an indicator of paranoia.

Some of the self-report inventories measure only one trait, such as the California F-Scale, which measures authoritarianism. Others, such as Cattell's Sixteen Personality Factor Questionnaire, measure a number of traits. Other multiple-trait inventories used in research are the Minnesota Multiphasic Personality Inventory (MMPI-2), the Guilford–Zimmerman Temperament Survey, the Mooney Problem Check List, the Edwards Personal Preference Schedule (EPPS), the Myers–Briggs Type Indicator, and the Strong Interest Inventory. A popular inventory, the Adjective Checklist, asks individuals to check from a list of adjectives those that are applicable to themselves. It is appropriate for individuals in grade 9 through adults and only takes 15 minutes to complete. It yields scores on self-confidence, self-control, needs, and other aspects of personality adjustment.

**Inventories** have been used in educational research to obtain trait descriptions of certain defined groups, such as underachievers and dropouts. They are useful for finding out about students' self-concepts, their concerns or problems, and their study skills and habits. Inventories have also been used in research concerned with interrelationships between personality traits and such variables as aptitude, achievement, and attitudes.

Inventories have the advantages of economy, simplicity, and objectivity. They can be administered to groups and do not require trained psychometricians. Most of the disadvantages are related to the problem of validity. The validity of self-report inventories depends in part on the respondents' being able to read and understand the items, their understanding of themselves, and especially their willingness to give frank and honest answers. As a result, the information obtained from inventories may be superficial or biased. This possibility must be taken into account when using results obtained from such instruments. Some inventories have built in validity scales to detect faking, attempts to give socially desirable responses, or reading comprehension problems.

### PROJECTIVE PERSONALITY ASSESSMENT

**Projective techniques** are measures in which an individual is asked to respond to an ambiguous or unstructured stimulus. They are called *projective* because a person is expected to project into the stimulus his or her own needs, wants, fears, beliefs, anxieties, and experiences. On the basis of the subject's interpretation of the stimuli and his or her responses, the examiner attempts to construct a comprehensive picture of the individual's personality structure. Projective methods are used mainly by clinical psychologists for studying and diagnosing people with emotional problems. They are not frequently used in educational research because of the necessity of specialized training for administration and scoring and the expense involved in individual administration. Furthermore, many researchers question their validity primarily because of the complex scoring. The two best known projective techniques are the Rorschach Inkblot Technique and the Thematic Apperception Test (TAT). The Rorschach consists of 10 cards or plates each with either a black/white or a colored inkblot. Individuals are asked what they "see." Their responses are scored according to whether they used the whole or only a part of the inkblot or if form or color was used in structuring the response, whether movement is suggested, and other aspects. In the TAT, the respondent is shown a series of pictures varying in the extent of structure and ambiguity and asked to make up a story about each one. The stories are scored for recurrent themes, expression of needs, perceived problems, and so on. The TAT is designed for individuals age 10 years through adult. There is also a form available for younger children (Children's Apperception Test) and one for senior citizens (Senior Apperception Test).

## SCALES

Scales are used to measure attitudes, values, opinions, and other characteristics that are not easily measured by tests or other measuring instruments. A **scale** is a set of categories or numeric values assigned to individuals, objects, or behaviors for the purpose of measuring variables. The process of assigning scores to those objects in order to obtain a measure of a construct is called *scaling*. Scales differ from tests in that the results of these instruments, unlike those of tests, do not indicate success or failure, strength or weakness. They measure the degree to which an individual exhibits the characteristic of interest. For example, a researcher may use a scale to measure the attitude of college students toward religion or any other topic. A number of scaling techniques have been developed throughout the years.

## ATTITUDE SCALES

**Attitude scales** use multiple responses—usually responses to statements—and combine the responses into a single scale score. Rating scales, which we discuss later in this chapter, use judgments—made by the individual under study or by an observer—to assign scores to individuals or other objects to measure the underlying constructs.

Attitudes of individuals or groups are of interest to educational researchers. An attitude may be defined as a positive or negative affect toward a particular group, institution, concept, or social object. The measurement of attitudes presumes the ability to place individuals along a continuum of favorableness–unfavorableness toward the object.

If researchers cannot locate an existing attitude scale on their topic of interest, they must develop their own scales for measuring attitudes. We discuss two types of attitude scales: summated or Likert (pronounced *Lik'ert*) scales and bipolar adjective scales.

### Likert Scales: Method of Summated Ratings

The Likert scale (1932), named for Rensis Likert who developed it, is one of the most widely used techniques to measure attitudes. A **Likert scale** (a **summated rating scale**) assesses attitudes toward a topic by presenting a set of statements about the topic and asking respondents to indicate for each whether they strongly agree, agree, are undecided, disagree, or strongly disagree. The various agree–disagree responses are assigned a numeric value, and the total scale score is found by summing the numeric responses given to each item. This total score assesses the individual's attitude toward the topic.

A Likert scale is constructed by assembling a large number of statements about an object, approximately half of which express a clearly favorable attitude and half of which are clearly unfavorable. Neutral items are not used in a Likert scale. It is important that these statements constitute a representative sample of all the possible opinions or attitudes about the object. It may be helpful to think of all the subtopics relating to the attitude object and then write items on each subtopic. To generate this diverse collection of items, the researcher may find it helpful to ask people who are commonly accepted as having knowledge about and definite attitudes toward the particular object to write a number of positive and negative statements. Editorial writings about the object are also good sources of potential statements for an attitude scale. Figure 8.1 shows items from a Likert scale designed to measure attitudes toward capital punishment.

For pilot testing, the statements, along with five response categories arranged on an agreement–disagreement continuum, are presented to a group of subjects. This group should be drawn from a population that is similar to the one in which the scale will be used. The statements should be arranged in random order so as to avoid any response set on the part of the subjects.

The subjects are directed to select the response category that best represents their reaction to each statement: *strongly agree* (SA), *agree* (A), *undecided* (U), *disagree* (D), or *strongly disagree* (SD). There has been some question regarding whether the undecided option should be included in a Likert scale. Most experts in the field recommend that the researcher include a neutral or undecided choice

1. Capital punishment serves as a deterrent to premeditated crime.
   SA A U D SD
*2. Capital punishment is morally wrong.
   SA A U D SD
3. The use of capital punishment is the best way for society to deal with hardened criminals.
   SA A U D SD
*4. I would sign a petition in favor of legislation to abolish the death penalty.
   SA A U D SD
*5. Capital punishment should not be used because there is always the possibility that an innocent person could be executed.
   SA A U D SD
6. Capital punishment reduces the use of tax monies for the care of prison inmates.
   SA A U D SD
*7. Only God has the right to take a human life.
   SA A U D SD
8. If more executions were carried out, there would be a sharp decline in violent crime.
   SA A U D SD
*9. Capital punishment should only be considered after all rehabilitation efforts have failed.
   SA A U D SD
10. I believe murder deserves a stronger penalty than life imprisonment.
   SA A U D SD
*11. Capital punishment should be abolished because it is in conflict with basic human rights.
   SA A U D SD
*12. I would be willing to participate in an all-night vigil to protest the execution of a criminal in my state.
   SA A U D SD

*These are negative items, agreement with which is considered to reflect a negative or unfavorable attitude toward capital punishment.

**Figure 8.1** Example of a Likert Scale

*Source*: These items were taken from an attitude scale constructed by a graduate student in an educational research class.

because some respondents actually feel that way and do not want to be forced into agreeing or disagreeing.

*Scoring Likert Scales* To score the scale, the response categories must be weighted. For favorable or positively stated items, *strongly agree* is scored 5, *agree* is scored 4, *undecided* is scored 3, *disagree* is scored 2, and *strongly disagree* is scored 1. For unfavorable or negatively stated items, the weighting is reversed because disagreement with an unfavorable statement is psychologically equivalent to agreement with a favorable statement. Thus, for unfavorable statements, *strongly agree* would receive a weight or score of 1 and *strongly disagree* a weight of 5. (The weight values do not appear on the attitude scale presented to respondents, nor do the asterisks seen in Figure 8.1.)

The sum of the weights of all the items checked by the subject is the individual's total score. The highest possible scale score is $5 \times N$ (the number of items); the lowest possible score is $1 \times N$.

Let us consider an example of scoring a Likert scale by looking at just the first six statements of the scale shown in Figure 8.1. An individual would complete this scale by circling the appropriate letter(s) for each statement.

The following are the responses circled by a hypothetical respondent and the score for each item:

| Response | Score |
|---|---|
| 1. D | 2 |
| 2. SA | 1 |
| 3. D | 2 |
| 4. A | 2 |
| 5. A | 2 |
| 6. U | 3 |

The individual's total score on the six items is 12 (out of a possible 30). Divide the total score by the number of items to arrive at a mean attitude score: $2 + 1 + 2 + 2 + 2 + 3)/6 = 2.0$. Because the mean score is less than 3, we conclude that this individual has a moderately negative attitude toward capital punishment.

*Item Analysis* After administering the attitude scale to a preliminary group of respondents, the researcher does an **item analysis** to identify the best functioning items. The item analysis typically yields three statistics for each item: (1) an item discrimination index, (2) the percentage of respondents marking each choice to each item, and (3) the item mean and standard deviation.

The item discrimination index shows the extent to which each item discriminates among the respondents in the same way as the total score discriminates. The item discrimination index is calculated by correlating item scores with total scale scores, a procedure usually done by computer. If high scorers on an individual item have high total scores and if low scorers on this item have low total scores, then the item is discriminating in the same way as the total score. To be useful, an item should correlate at least .25 with the total score. Items that have very low correlation or negative correlation with the total score should be eliminated because they are not measuring the same thing as the total scale and hence are not contributing to the measurement of the attitude. The researcher will want to examine those items that are found to be nondiscriminating. The items may be ambiguous or double barreled (containing two beliefs or opinions in one statement), or they may be factual statements not really expressing feelings about the object. Revising these items may make them usable. The item analysis also shows the percentage of respondents choosing each of the five options and the mean and standard deviation for each item. Items on which respondents are spread out among the options are preferred. Thus, if most respondents choose only one or two of the options, the item should be rewritten or eliminated. After selecting the most useful items as indicated by the item analysis, the researcher should then try out the revised scale with a different group of subjects and again check the items for discrimination and variability.

*Validity* Validity concerns the extent to which the scale really measures the attitude construct of interest. It is often difficult to locate criteria to be used

in obtaining evidence for the validity of attitude scales. Some researchers have used observations of actual behavior as the criterion for the attitude being measured. This procedure is not often used because it is often difficult to determine what behavior would be the best criterion for the attitude and also because it is expensive.

One of the easiest ways to gather validity evidence is to determine the extent to which the scale is capable of discriminating between two groups whose members are known to have different attitudes (see Chapter 9). To validate a scale that measures attitudes toward organized religion, a researcher would determine if the scale discriminated between active church members and people who do not attend church or have no church affiliation. A scale measuring attitudes toward abortion should discriminate between members of pro-life groups and members of pro-choice groups. By "discriminate," we mean that the two groups would be expected to have significantly different mean scores on the scale. Another method of assessing validity is to correlate scores on the attitude scale with those obtained on another attitude scale measuring the same construct and whose validity is well established.

*Reliability* The reliability of the new scale must also be determined. Reliability is concerned with the extent to which the measure would yield consistent results each time it is used. The first step in ensuring reliability is to make sure that the scale is long enough—that it includes enough items to provide a representative sampling of the whole domain of opinions about the attitudinal object. Other things being equal, the size of the reliability coefficient is directly related to the length of the scale. Research shows, however, that if the items are well constructed, scales having as few as 20 to 22 items will have satisfactory reliability (often above .80). The number of items needed depends partly on how specific the attitudinal object is; the more abstract the object, the more items are needed.

You would also want to calculate an index of reliability. The best index to use for an attitude scale is coefficient alpha (see Chapter 9), which provides a measure of the extent to which all the items are positively intercorrelated and working together to measure one trait or characteristic (the attitude). Many statistical computer programs routinely calculate coefficient alpha as a measure of reliability. For further discussion on the construction of Likert and other attitude scales, the reader is referred to Mueller (1986).

### Bipolar Adjective Scales

The **bipolar adjective scale** presents a respondent with a list of adjectives that have bipolar or opposite meanings. Respondents are asked to place a check mark at one of the seven points in the scale between the two opposite adjectives to indicate the degree to which the adjective represents their attitude toward an object, group, or concept. Figure 8.2 shows a bipolar adjective scale designed to measure attitude toward school. Notice that the respondent checked the extreme right position for item a and the extreme left position for item d. The adjective pairs making up a scale are listed in both directions; on some pairs the rightmost position is the most positive response, and on other pairs the leftmost position is the most positive. This is done to minimize a response set or a tendency to favor certain positions in a list of options. An individual might have a tendency to choose

School

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| a. bad | : | : | : | : | : | : ✓ | good |
| b. fast | : ✓ | : | : | : | : | : | slow |
| c. dull | : | : | : | : | : ✓ | : | sharp |
| d. pleasant | ✓ : | : | : | : | : | : | unpleasant |
| e. light | : | : ✓ | : | : | : | : | heavy |
| f. passive | : | : | : | : | : | : ✓ | active |
| g. worthless | : | : | : | : | : ✓ | : | valuable |
| h. strong | : | : | : ✓ | : | : | : | weak |
| i. still | : | : | : | : ✓ | : | : | moving |

**Figure 8.2** Bipolar Adjective Scale Showing Responses of One Subject Toward the Concept "School"

the extreme right end and would check that position for each item. However, if the direction of the scale is changed in a random way so that the right end is not always the more favorable response, the individual must read each item and respond in terms of its content rather than in terms of a positional preference. The responses are scored by converting the positions checked into ratings (1 to 7). Seven represents the most positive and 1 the least positive response on each scale. The weights on each item would then be summed and averaged. In Figure 8.2, item weights are $7+6+6+7+3+7+6+4+5=51/9=5.67$. The score of 5.67 indicates a very positive attitude toward school.

The bipolar adjective scale is a very flexible approach to measuring attitudes. A researcher can use it to investigate attitudes toward any concept, person, or activity in any setting. It is much easier and less time-consuming to construct than a Likert scale. Instead of having to come up with approximately 20 statements, you need only select four to eight adjective pairs. It requires very little reading time by participants. The main difficulty is the selection of the adjectives to use. If one has a problem with this task, there are references such as Osgood, Suci, and Tannenbaum (1967) that provide lists of bipolar adjectives. It is probably better, however, to think of adjective pairs that are especially relevant to one's own project.

## RATING SCALES

**Rating scales** present a number of statements about a behavior, an activity, or a phenomenon with an accompanying scale of categories. Observers or respondents are asked to indicate their assessment or judgment about the behavior or activity on the rating scale. For example, a teacher might be asked to rate the leadership ability of a student. The teacher would indicate his or her assessment of the student's characteristic leadership behavior by checking a point on a continuum or choosing a response category. It is assumed that raters are familiar with the behavior they are asked to assess. A numeric value may be attached to the points or categories so that an overall score could be obtained.

One of the most widely used rating scales is the **graphic scale,** in which the respondent indicates the rating by placing a check at the appropriate point on a

| | Low | Medium | High |
|---|---|---|---|
| Personal appearance | | | |
| Social acceptability | | | |
| Speaking skills | | | |

**Figure 8.3** Example of a Graphic Scale

horizontal line that runs from one extreme of the behavior in question to the other. Figure 8.3 is an example of a graphic scale. The rater can check any point on the continuous line. Graphic scales usually assign numeric values to the descriptive points. Such scales are referred to as *numeric rating scales*. The speaking skills item in Figure 8.3 could look like this in a numeric scale:

| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| one of the poorest speakers | | | an average speaker | | | one of the very best speakers |

### Category Scales

The **category scale** consists of a number of categories that are arranged in an ordered series. Five to seven categories are most frequently used. The rater picks the one that best characterizes the behavior of the person being rated. Suppose a student's abilities are being rated and one of the characteristics being rated is creativity. The following might be one category item:

How creative is this person? (check one)

exceptionally creative _____
very creative _____
not creative _____
not at all creative _____

To provide greater meaning, brief descriptive phrases are sometimes used to comprise the categories in this type of scale. Clearly defined categories contribute to the accuracy of the ratings. For example,

How creative is this person? (check one)

always has creative ideas _____
has many creative ideas _____
sometimes has creative ideas _____
rarely has creative ideas _____

### Comparative Rating Scales

In using the graphic and category scales, raters make their judgments without directly comparing the person being rated to other individuals or groups. In **comparative rating scales,** in contrast, raters are instructed to make their judgment with direct reference to the positions of others with whom the individual might be compared. The positions on the rating scale are defined in terms of a given population with known characteristics. A comparative rating scale is shown in Figure 8.4. Such a scale might be used in selecting applicants for admission to graduate school. Raters are asked to judge the applicant's ability to do graduate work compared with that of all the students the rater has known. If the rating is to be valid, the judge must understand the range and distribution of abilities in the total group of graduate students.

*Errors in Rating* Because ratings depend on the perceptions of human observers, who are susceptible to various influences, rating scales are subject to

| Area of Competency (to be rated) | Unusually low | Poorer than most students | About average among students | Better than most | Really superior | Not able to judge |
|---|---|---|---|---|---|---|
| 1. Does this person show evidence of clear-cut and worthy professional goals? | | | | | | |
| 2. Does this person attack problems in a constructive manner? | | | | | | |
| 3. Does he or she take well-meant criticism and use it constructively? | | | | | | |

**Figure 8.4** Example of a Comparative Rating Scale

considerable error. Among the most frequent systematic errors in rating people is the **halo effect,** which occurs when raters allow a generalized impression of the subject to influence the rating given on very specific aspects of behavior. This general impression carries over from one item in the scale to the next. For example, a teacher might rate a student who does good academic work as also being superior in intelligence, popularity, honesty, perseverance, and all other aspects of personality. Or, if you have a generally unfavorable impression of a person, you are likely to rate the person low on all aspects.

Another type of error is the **generosity error,** which refers to the tendency for raters to give subjects the benefit of any doubt. When raters are not sure, they tend to rate people favorably. In contrast, the **error of severity** is a tendency to rate all individuals too low on all characteristics. Another source of error is the **error of central tendency,** which refers to the tendency to avoid either extreme and to rate all individuals in the middle of the scale. For example, the ratings that teachers of English give their students have been found to cluster around the mean, whereas mathematics teachers' ratings of students show greater variation.

One way of reducing such errors is to train the raters thoroughly before they are asked to make ratings. They should be informed about the possibility of making these "personal bias" types of errors and how to avoid them. It is absolutely essential that raters have adequate time to observe the individual and his or her behavior before making a rating. Another way to minimize error is to make certain that the behavior to be rated and the points on the rating scale are clearly defined. The points on the scale should be described in terms of overt behaviors that can be observed, rather than in terms of behaviors that require inference on the part of the rater.

The accuracy or reliability of ratings is usually increased by having two (or more) trained raters make independent ratings of an individual. These independent ratings are pooled, or averaged, to obtain a final score. A researcher may also correlate the ratings of the two separate raters in order to obtain a coefficient of interrater reliability (see Chapter 9). The size of the coefficient indicates the extent to which the raters agree. An **interrater reliability** coefficient of .70 or higher is considered acceptable for rating scales.

## DIRECT OBSERVATION

In many cases, systematic or **direct observation** of behavior is the most desirable measurement method. Observation is used in both quantitative and qualitative research. When observations are made in an attempt to obtain a comprehensive picture of a situation, and the product of those observations is notes or narratives, the research is qualitative. In Chapter 15, we discuss the use of observation in qualitative research. The current chapter focuses on observation in quantitative research where the product of using the various observational instruments is numbers. The purpose of direct observation is to determine the extent to which a particular behavior(s) is present. The observer functions like a camera or recording device to provide a record of the occurrence of the behavior in question. The researcher identifies the behavior of interest and devises a systematic procedure for identifying, categorizing, and recording the behavior in either a natural or a contrived situation. The behaviors observed in quantitative studies may be categorized as high inference and low inference. High-inference behaviors such as teacher warmth or creativity require more judgment on the part of the observer. Low-inference behaviors require less judgment by the observer. Examples of low-inference behaviors include classroom behaviors such as teachers' asking questions, praising students, or accepting students' ideas. In educational research, one of the most common uses of direct observation is in studying classroom behavior. For example, if you were interested in investigating the extent to which elementary teachers use positive reinforcement in the classroom, you could probably obtain more accurate data by actually observing classrooms rather than asking teachers about their use of reinforcement. Or if you wanted to study students' disruptive behavior in the classroom and how teachers deal with it, direct observation would provide more accurate data than reports from students or teachers.

There are five important preliminary steps to take in preparing for quantitative direct observation:

1. *Select the aspect of behavior to be observed*. Because it is not possible to collect data on everything that happens, the investigator must decide beforehand which behaviors to record and which not to record.
2. *Clearly define the behaviors falling within a chosen category*. Know what behaviors would be indicators of the attribute. In studying aggressive behavior in the classroom, would challenging the teacher or speaking out of turn be classified as aggressive, or would it be restricted to behaviors such as pushing, hitting, throwing objects, and name-calling? If observing multiple categories of behavior, make sure the categories are mutually exclusive.
3. *Develop a system for quantifying observations*. The investigator must decide on a standard method for counting the observed behaviors. For instance, establish beforehand whether an action and the reaction to it are to count as a single incident of the behavior observed or as two incidents. A suggested approach is to divide the observation period into brief time segments and to record for each period—for example, 10 seconds—whether the subject showed the behavior or not.

4. *Develop specific procedures for recording the behavior*. Record the observations immediately after they are made because observers' memory is not sufficiently reliable for accurate research. The best solution is a coding system that allows the immediate recording of what is observed, using a single letter or digit. A coding system is advantageous in terms of the observers' time and attention.
5. *Train the people who will carry out the observations*. Training and opportunity for practice are necessary so that the investigator can rely on the observers to follow an established procedure in observing and in interpreting and reporting observations. Having the observers view a videotape and discuss the results is a good training technique.

## DEVICES FOR RECORDING OBSERVATIONS

Researchers use checklists, rating scales, and coding sheets to record the data collected in direct observation.

### Checklists

The simplest device used is a **checklist,** which presents a list of the behaviors that are to be observed. The observer then checks whether each behavior is present or absent. A checklist differs from a scale in that the responses do not represent points on a continuum but, rather, nominal categories. For example, a researcher studying disruptive behavior would prepare a list of disruptive behaviors that might occur in a classroom. An observer would then check items such as "Passes notes to other students" or "Makes disturbing noises" each time the behavior occurs. The behaviors in a checklist should be operationally defined and readily observable.

### Rating Scales

Rating scales, discussed previously, are often used by observers to indicate their evaluation of an observed behavior or activity. Typically, rating scales consist of three to five points or categories. For example, an observer studying teachers' preparation for presentation of new material in a classroom might use a scale with the following points: 5 (*extremely well prepared*), 4 (*well prepared*), 3 (*prepared*), 2 (*not well prepared*), or 1 (*totally unprepared*). A 3-point scale might include 3 (*very well prepared*), 2 (*prepared*), or 1 (*not well prepared*). Scales with more than five rating categories are not recommended because it is too difficult to accurately discriminate among the categories.

### Coding Systems

**Coding systems** are used in observational studies to facilitate the categorizing and counting of specific, predetermined behaviors as they occur. The researcher does not just indicate whether a behavior occurred as with a checklist but, rather, uses agreed-on codes to record what actually occurred. Whereas rating scales can be completed after an observation period, coding is completed at the time the observer views the behavior.

Two kinds of coding systems are typically used by researchers: sign coding and time coding. *Sign coding* uses a set of behavior categories; each time one of the behaviors occurs, the observer codes the happening in the appropriate category. If a coding sheet used in classroom observational research listed "summarizing" as a teacher behavior, the observer would code a happening every time a teacher summarized material.

In a study using sign coding, Skinner, Buysse, and Bailey (2004) investigated how total duration and type of social play of preschool children with disabilities varied as a function of the chronological and developmental age of their social partners. They hypothesized that developmental age of each partner would better predict the duration of social play than chronological age. The 55 focal children were preschool children with mild to moderate developmental delays who were enrolled in some type of inclusive developmental day program. Each focal child was paired with 4 different same-sex partners in a standardized dyadic play situation. The observations took place outside the classroom in a specially designed and well-equipped play area. The observation consisted of two 15-minute sessions with each of the 4 play partners, or a total of 120 minutes per focal child over a period of 2 days. A video camera recorded the play behavior and trained coders used Parten's (1932) seven categories of play to code the extent to which children were engaged socially. The Battelle Developmental Inventory (Newborg, Stock, Wnek, Guidubaldi, & Svinicki, 1988) was used to assess the overall developmental status of both focal children and their social partners. A mixed-model regression analysis was employed, with the independent variables being the chronological and developmental ages of both the focal children and the partners; the dependent variable was the total duration of the category called associative play. No impact was observed for the focal children's chronological age once they accounted for developmental age. Also, they found that the influence of partner's developmental age on social play was different depending on the developmental age of the focal child. The researchers concluded that advantages accrued to preschoolers with disabilities from mixed-aged play groupings depend on the child's developmental age and those of available social partners.

In the second type of coding, called *time coding*, the observer identifies and records all predetermined behavior categories that occur during a given time period. The time period might be 10 seconds, 5 minutes, or some other period of time. Miller, Gouley, and Seifer (2004) used time coding in a study designed to document observed emotional and behavioral dysregulation in the classroom and to investigate the relationships between observed dysregulation and teachers' ratings of children's classroom adjustment and their social engagement with peers. Dysregulation was defined as emotional and behavioral displays disruptive to the preschool classroom setting. The participants were 60 low-income children attending Head Start classes. Each child was observed in a naturalistic context for two sessions of 10 minutes each, or a total of 20 minutes. The researchers used handheld computers with *The Observer* (Noldus Information Technology, 1995) software, which permitted coding of behavior along several dimensions. Analysis showed that although the majority of children did not display much dysregulated emotion or behavior in the classroom, almost one-fourth of children did display high levels of dysregulation in the observation period.

High levels of classroom dysregulation were related to teacher ratings of poor classroom adjustment and observed peer conflict behaviors, as well as negative emotional displays.

Coding has the advantage of recording observations at the time the behavior occurs, and it may yield more objective data than do rating scales. The disadvantage is that a long training period may be required for observers to learn to code behavior reliably. A number of standardized coding systems and observation forms are available. Beginning researchers should check references such as the ETS Test Collection Database for a suitable one before attempting to construct their own.

## ADVANTAGES AND DISADVANTAGES OF DIRECT OBSERVATION

The most obvious advantage of systematic observation is that it provides a record of the actual behavior that occurs. We do not have to ask subjects what they would do or what they think; we have a record of their actions. Probably the most important advantage of systematic observation is its appropriateness for use with young children. It is used extensively in research on infants and on preschool children who have difficulty communicating through language and may be uncomfortable with strangers. Another advantage is that systematic observation can be used in natural settings. It is often used in educational research to study classroom or playground behavior.

The main disadvantage of systematic observation is the expense. Observations are more costly because of the time required of trained observers. Subjects may be observed for a number of sessions, requiring extended hours.

## VALIDITY AND RELIABILITY OF DIRECT OBSERVATION

As with other types of measures, the validity and reliability of direct observation must be assessed. The best way to enhance validity is to carefully define the behavior to be observed and to train the people who will be making the observations. Observers must be aware of two sources of bias that affect validity: observer bias and observer effect. **Observer bias** occurs when the observer's own perceptions, beliefs, and biases influence the way he or she observes and interprets the situation. Having more than one person make independent observations helps to detect the presence of bias. **Observer effect** occurs when people being observed behave differently just because they are being observed. One-way vision screens may be used in some situations to deal with this problem. In many cases, however, after an initial reaction the subjects being observed come to pay little attention to the observer, especially one who operates unobtrusively. Some studies have used interactive television to observe classrooms unobtrusively. Videotaping for later review and coding may also be useful. Researchers who have used videotapes, for example, have found that although the children initially behaved differently with the equipment in the room, after a few days they paid no attention and its presence became routine. Handheld technologies, such as a PalmPilot, can be used to record data during observations rather than the traditional pencil-and-paper recording techniques. Professional

software such as *The Observer XT* 8.0 (Noldus Information Technology, 2008) is available for use in the collection, analysis, and presentation of observational data. Information on *The Observer XT 8.0* is available at www.noldus.com/site/doc200806003.

The accuracy or reliability of direct observation is usually investigated by having at least two observers independently observe the behavior and then determining the extent to which the observers' records agree. Reliability is enhanced by providing extensive training for the observers so that they are competent in knowing what to observe and how to record the observations. Further discussion of methods for assessing the reliability of direct observation is presented in Chapter 9.

### CONTRIVED OBSERVATIONS

In **contrived observations,** the researcher arranges for the observation of subjects in simulations of real-life situations. The circumstances have been arranged so that the desired behaviors are elicited.

One form of contrived observation is the **situational test.** A classic example of a situational test—although not labeled as such at the time—was used in a series of studies by Hartshorne and May (1928) for the Character Education Inquiry (CEI). These tests were designed for use in studying the development of such behavior characteristics as honesty, self-control, truthfulness, and cooperativeness. Hartshorne and May observed children in routine school activities but also staged some situations to focus on specific behavior. For example, they gave vocabulary and reading tests to the children, collected the tests, and without the children's knowledge made duplicate copies of their answers. Later, the children were given answer keys and were asked to score their original papers. The difference between the scores the children reported and the actual scores obtained from scoring the duplicate papers provided a measure of cheating. Another test asked the children to make a mark in each of 10 small, irregularly placed circles while keeping their eyes shut. Previous control tests under conditions that prevented peeking indicated that a score of more than 13 correctly placed marks in a total of three trials was highly improbable. Thus, a score of more than 13 was recorded as evidence that the child had peeked.

Hartshorne and May (1928) found practically no relationship between cheating in different situations, such as on a test and in athletics. They concluded that children's responses were situationally specific—that is, whether students cheated depended on the specific activity, the teacher involved, and other situations rather than on some general character trait.

## DATA COLLECTION IN QUALITATIVE RESEARCH

Qualitative researchers also have a number of data-gathering tools available for their investigations. The most widely used tools in qualitative research are interviews, document analysis, and observation. We discuss these methods in Chapter 15.

## SUMMARY

One of the most important tasks of researchers in the behavioral sciences is the selection and/or development of dependable measuring instruments. A research study can be no better than the instruments used to collect the data. A variety of tests, scales, and inventories are available for gathering data in educational research, especially for quantitative studies. Researchers need to be aware of the strengths and limitations of these data-gathering instruments so that they can choose the one(s) most appropriate for their particular investigation. If an appropriate standardized instrument is available, the researcher would be wise to choose it because of the advantage in terms of validity, reliability, and time saved.

A test is a set of stimuli presented to an individual to elicit responses on the basis of which a numerical score can be assigned. Achievement tests measure knowledge and proficiency in a given area and are widely used in educational research. Standardized achievement tests permit the researcher to compare performance on the test to the performance of a normative reference group. Tests may be classified as paper-and-pencil or as performance tests, which measure what someone can *do* rather than what he or she *knows*. Aptitude tests are used to assess an individual's verbal and nonverbal capacities. Personality inventories are designed to measure the subject's personal characteristics and typical performance.

Attitude scales are tools for measuring individuals' beliefs, feelings, and reactions to certain objects. The major types of attitude scales are Likert-type scales and the bipolar adjective scale.

Rating scales permit observers to assign scores to the assessments made of observed behavior or activity. Among the types of rating scales are the graphic scale, the category scale, and comparative rating scales.

Rating scales, checklists, and coding systems are most commonly used to record the data in quantitative direct observation research. In coding systems, behavior can be categorized according to individual occurrences (sign coding) or number of occurrences during a specified time period (time coding).

## KEY CONCEPTS

achievement test
aptitude test
attitude scale
bipolar adjective scale
category scale
ceiling effect
checklist
coding system
comparative rating scales
contrived observation
criterion-referenced test
direct observation
error of central tendency
error of severity
floor effect
generosity error
graphic scale
halo effect
intelligence test
interrater reliability
inventories
item analysis
Likert scale
norm-referenced test
observer bias
observer effect
performance test
projective technique
rating scale
researcher-made test
scale scholastic aptitude test
self-report inventories
situational test
standardized test
summated rating scale
teacher-made test
test

## EXERCISES

1. What is the meaning of the term *standardized* when applied to measuring instruments?
2. What is the difference between comparative rating scales and graphic and category scales?
3. List some of the common sources of bias in rating scales.
4. What type of instrument would a researcher choose in order to obtain data about each of the following?

**a.** How college professors feel about the use of technology in their teaching
**b.** The potential of the seniors at a small college to succeed in graduate school
**c.** To determine if high school chemistry students can analyze an unknown chemical compound
**d.** How well the students at Brown Elementary School compare to the national average in reading skills
**e.** The advising-style preferences of a group of college freshmen
**f.** How well students perform in a public speaking contest
**g.** To determine the winner in a history essay contest
**h.** The general verbal and nonverbal abilities of a student with attention deficit disorder
**i.** The extent to which elementary teachers use negative reinforcement in the classroom, and the effect of that reinforcement on students' behavior
**j.** The problems faced by minority students during the first year at a large research university
**k.** How parents in a school system feel about moving the sixth grade from the elementary school to the middle school

5. How would you measure parents' attitudes toward a new dress code proposed for a middle school?
6. What are some procedures for increasing the accuracy of direct observation techniques?
7. Construct a five-item Likert scale for measuring peoples' attitudes toward stem cell research.
8. Intelligence tests can most accurately be described as
   **a.** Measures of innate mental capacity
   **b.** Academic achievement measures
   **c.** Reading tests
   **d.** Scholastic aptitude tests
9. List and briefly describe the instruments available for recording data in observational research.
10. What type of instrument would be most appropriate to measure each of the following?
    **a.** To determine if high school chemistry students can use laboratory scales to weigh specified amounts of a given chemical compound
    **b.** How students in the various elementary schools in Brown County compare in math skills
    **c.** How parents feel about an extended school day for elementary schools in the district
    **d.** The general verbal and nonverbal abilities of a child with dyslexia
    **e.** To study bullying in an elementary classroom
    **f.** To get a major professor's evaluation of the potential of a student for advanced work in chemistry
    **g.** To get a quick measure of students' attitudes toward the extracurricular programs available at the school

## ANSWERS

1. *Standardized* refers to instruments for which comparative norms have been derived, their reliability and validity have been established, and directions for administration and scoring have been prescribed.
2. In judging an individual on a comparative rating scale, the rater must have knowledge of the group with which the individual is being compared. In judging an individual on graphic and category scales, raters do not make a direct comparison of the subject with other people.
3. Raters may be less than objective in judging individuals when influenced by such tendencies as the halo effect, the generosity error, the error of severity, or the error of central tendency.
4. **a.** Attitude scale
   **b.** Aptitude test (group)
   **c.** Performance test
   **d.** Standardized reading achievement test
   **e.** Inventory
   **f.** Rating scale (performance test)
   **g.** Performance test
   **h.** Aptitude or intelligence test (individual)
   **i.** Direct observation
   **j.** Inventory
   **k.** Attitude scale

5. Construct a Likert scale containing approximately 20 statements expressing positive and negative feelings about the proposed dress code or construct a bipolar adjective scale.
6. The behaviors to be observed must be specified; behaviors falling within a category must be defined; a system for quantification must be developed; and the observers must be trained to carry out the observations according to this established procedure.
7. Answers will vary.
8. d
9. Checklists indicate the presence or absence of certain behaviors. Rating scales and coding schemes both yield quantitative measures. In ratings, the person indicates his or her judgment of the behavior on a continuum. Ratings are sometimes completed in retrospect. Coding schemes are used to categorize observed behavior as it occurs.
10. **a.** Performance test
    **b.** Standardized achievement test
    **c.** Attitude scale
    **d.** Individual intelligence test, such as the Wechsler
    **e.** Observation
    **f.** Comparative rating scale
    **g.** Bipolar adjective scale

## REFERENCES


Geisinger, K., Spies, R., Carlson, J., & Plake, B. (Eds.). (2007). *The Seventeenth Mental Measurements Yearbook*. Lincoln: University of Nebraska, Buros Institute of Mental Measurements.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Mahwah, NJ: Erlbaum.

Hartshorne, H., & May, M. A. (1928). *Studies in the nature of character: Studies in deceit* (Vol. 1). New York: Macmillan. [Reprinted in 1975 by Ayer, New York]

Kubiszyn, T., & Borich, G. (2006). *Educational testing and measurement*. New York: Wiley.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, no. 140.

Marsh, H. W. (1988). *Self-description questionnaire: A theoretical and empirical basis for the measurement of multiple dimensions of preadolescent self-concept: A test manual and a research monograph*. San Antonio, TX: The Psychological Corporation.

Miller, A., Gouley, K., & Seifer, R. (2004). Emotions and behaviors in the Head Start classroom: Associations among observed dysregulation, social competence, and preschool adjustment. *Early Education and Development*, *15*(2), 147–165.

Mueller, D. J. (1986). *Measuring social attitudes: A handbook for researchers and practitioners*. New York: Teachers College Press.

Murphy, L., Plake, B., & Spies, R. (Eds.). (2006). *Tests in print VII: An index to tests, test reviews, and the literature on specific tests*. Lincoln: University of Nebraska, Buros Institute of Mental Measurements.

Newborg, J., Stock, J., Wnek, L., Guidubaldi, J., & Svinicki, J. (1988). *The Battelle Developmental Inventory (BDI)*. Chicago: Riverside.

Noldus Information Technology. (1995). *The observer: System for collection and analysis of observational data* (Version 3.0). Sterling, VA: Author.

Noldus Information Technology. (2008). *The observer XT 8.0*. Sterling, VA: Author.

Osgood, C. E., Suci, G. J., & Tannenbaum, P.H. (1967). *The measurement of meaning*. Urbana: University of Illinois Press.

Parten, M. B. (1932). Social participation among preschool children. *Journal of Abnormal Social Psychology*, *27*, 243–269.

Popham, W. J. (2005). *Classroom assessment: What teachers need to know*. Boston: Allyn & Bacon.

Shavelson, R. J., Huber, J. J., & Stanton, G. C. (1976). Self-concept: Validation of construct interpretations. *Review of Educational Research*, *46*, 407–441.

Skinner, M., Buysse, V., & Bailey, D. (2004). Effects of age and developmental status of partners on play of preschoolers with disabilities. *Journal of Early Intervention*, *26*(3), 194–203.

Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education*. Upper Saddle River, NJ: Pearson Education.

CHAPTER 9

# Validity and Reliability

Information derived from measuring instruments ranges from excellent to useless to downright misleading. There are systematic ways to assess the usefulness of the scores derived from measuring instruments.

## INSTRUCTIONAL OBJECTIVES

After studying this chapter, the student will be able to:

1. Distinguish between validity and reliability.
2. List the major types of evidence used to support the valid interpretation of test scores.
3. Define construct underrepresentation and construct-irrelevant variance and explain their relevance to the validity of test scores.
4. Distinguish between convergent and discriminant evidence of validity.
5. Distinguish between random and systematic errors of measurement and their relationship to validity and reliability of test scores.
6. State the different sources of random error in educational and psychological measures.
7. Describe the different procedures (test–retest, equivalent forms, split-half, Kuder–Richardson, and others) for estimating the reliability of a measure.
8. Compute reliability coefficients for given data.
9. Define interobserver reliability and explain how it is calculated.
10. Apply the Spearman–Brown formula to determine the effect of lengthening a test on test reliability.
11. Explain the factors affecting the size of a reliability coefficient.
12. Compute the standard error of measurement and interpret score bands as indications of reliability.
13. Compute indexes to show the reliability of a criterion-referenced test.

Quantitative research always depends on measurement. Chapter 8 introduced you to some of the measuring instruments used in research. Two very important concepts that researchers must understand when they use measuring instruments are *validity* and *reliability*. Validity is defined as the extent to which scores on a test enable one to make meaningful and appropriate interpretations. Reliability indicates how consistently a test measures whatever it does measure. Researchers must be concerned about the validity and reliability of the scores derived from instruments used in a study and must include this information in the research report. If a researcher's data are not obtained with instruments that allow valid and reliable interpretations, one can have little faith in the results obtained or in the conclusions based on the results.

## VALIDITY

Validity is the most important consideration in developing and evaluating measuring instruments. Historically, **validity** was defined as the extent to which an instrument measured what it claimed to measure. The focus of recent views of validity is not on the instrument itself but on the interpretation and meaning of the scores derived from the instrument. The most recent *Standards for Educational and Psychological Testing* (1999),* prepared by the American Educational Research Association, the National Council on Measurement in Education, and the American Psychological Association, defines validity as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). Measuring instruments yield scores; however, the important issue is the interpretation we make of the scores, which may or may not be valid. For example, a fourth-grade math test that might allow a teacher to make valid interpretations about the math achievement of her fourth-grade students would not yield valid interpretations about the fourth-graders' abilities to solve algebra problems. If one tried to use the math achievement test for this purpose, it would be the interpretations about the students' ability to solve algebra problems that would be invalid, not the test. Thus, we no longer speak of the validity of the instrument but, rather, the validity of the interpretations or inferences that are drawn from the instrument's scores. Validity does not travel with the instrument. A test may be valid for use with one population or setting but not with another.

Assessing the validity of score-based interpretations is important to the researcher because most instruments used in educational and psychological investigations are designed for measuring hypothetical constructs. Recall that constructs such as intelligence, creativity, anxiety, critical thinking, motivation, self-esteem, and attitudes represent abstract variables derived from theory or observation. Researchers have no direct means of measuring these constructs such as exist in the physical sciences for the measurement of characteristics such as length, volume, and weight. To measure these hypothetical constructs, you must move from the theoretical domain surrounding the construct to an empirical level that operationalizes the construct. That is, we use an operational definition to measure the construct. We do this by selecting specific sets of observable tasks believed to serve as indicators of the particular theoretical construct. Then we assume that performance (scores) on the tasks reflects the particular construct of interest as distinguished from other constructs. Essentially, validity deals with how well the operational definition fits with the conceptual definition.

Tests may be imprecise measures of the constructs they are designed to assess because they leave out something that theory states should be included, include something that should be left out, or both. Messick (1995) identified two problems that threaten the interpretation (validity) of test scores: construct underrepresentation and construct-irrelevant variance. The term **construct underrepresentation** refers to assessment that is too narrow and fails to include important dimensions of the construct. The test may not adequately sample some kinds of content or some types of responses or psychological processes and thus fails to adequately represent the theoretical domain of the construct. Individuals'

*The 1999 edition of the *Standards* is currently being revised (see www.apa.org/science/standards.html).

scores on a math test may be misleading because the test did not measure some of the relevant skills that, if represented, would have allowed the individuals to display their competence. Or a scale designed to measure general self-concept might measure only social self-concept and not academic and physical components of self-concept.

The term **construct-irrelevant variance** refers to the extent to which test scores are affected by variables that are extraneous to the construct. Low scores should not occur because the test contains something irrelevant that interferes with people's demonstration of their competence. Construct-irrelevant variance could lower scores on a science achievement test for individuals with limited reading skills or limited English skills. Reading comprehension is thus a source of construct-irrelevant variance in a science achievement test and would affect the validity of any interpretations made about the individuals' science achievement.

## VALIDATION

The process of gathering evidence to support (or fail to support) a particular interpretation of test scores is referred to as validation. We need evidence to establish that the inferences, which are made on the basis of the test results, are appropriate. Numerous studies may be required to build a body of evidence about the validity of these score-based interpretations. The *Standards for Educational and Psychological Testing* lists three categories of evidence used to establish the validity of score-based interpretations: evidence based on content, evidence based on relations to a criterion, and construct-related evidence of validity. Using these categories does not imply that there are distinct types of validity but, rather, that different types of evidence may be gathered to support the intended use of a test. The categories overlap and all are essential to a unitary concept of validity.

### 1. Evidence Based on Test Content

**Evidence based on test content** involves the test's content and its relationship to the construct it is intended to measure. The *Standards* defines content-related evidence as "The degree to which the sample of items, tasks, or questions on a test are representative of some defined universe or domain of content." That is, the researcher must seek evidence that the test to be used represents a balanced and adequate sampling of all the relevant knowledge, skills, and dimensions making up the content domain. Evidence based on test content is especially important in evaluating achievement tests. In this age of educational accountability, content validity is receiving renewed attention. Crocker (2003) wrote, "When scores are used for educational accountability, the 'load-bearing wall' of that validity argument is surely content representativeness" (p. 7). Validation of an achievement test, for instance, would consider the appropriateness of the test's content to the total content area to be measured as well as how adequately the test samples the total domain. One would not attempt to measure chemistry students' knowledge of oxidation, for example, with only two questions.

A researcher who wished to develop a test in fourth-grade mathematics for a particular school system would need to determine what kinds of content (skills and objectives) are covered in fourth-grade classes throughout the system. After examining textbooks, syllabi, objectives, and talking with teachers, the researcher

would prepare an outline of the topics, computational and conceptual skills, and performances that make up fourth-grade mathematics (content domain) in that system, along with an indication of the emphasis given to each. Using the outline as a guide, the researcher would write a collection of test items that cover each topic and each objective in proportion to the emphasis given to each in the total content domain. The result should be a representative sample of the total domain of knowledge and skills included in that school system's fourth-grade math classes.

If a math test were designed to be used nationally, the researcher would need to examine widely used textbooks, states' curriculum guides, syllabi, and so on throughout the country to determine what content (concepts and skills) is included in fourth-grade math. The test content would be sampled to provide a representative and balanced coverage of this national curriculum. Subject matter experts and curriculum specialists would be asked to judge the adequacy of the test's content for measuring fourth-grade math achievement. Developers of nationally used achievement tests, such as the Stanford Achievement Test, are expected to provide extensive evidence of content validity. If a publisher says a test measures reading comprehension, for example, then the publisher should provide evidence that higher scores on the test are attributable to higher levels of reading comprehension rather than, for example, better memory.

To ensure content validity in a classroom test, a teacher should prepare a "blueprint" showing the content domain covered and the relative emphasis given to each aspect of the domain. If the pie chart in Figure 9.1 represents a teacher's assessment of the relative importance of topics within a unit on the American Civil War, a 50-item exam should include 10 items on topic A; 5 each on B, C, D, and G; 12 on E; and 8 on F.

There is no numerical index to indicate content validity. Evidence based on content is mainly the result of a logical examination or analysis by content experts that shows whether the instrument adequately represents the content

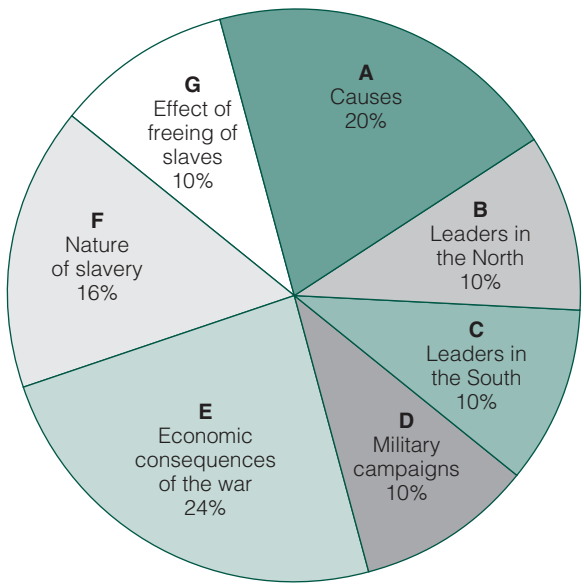



**Figure 9.1** Unit on the American Civil War

and objectives making up the domain. An achievement test may have content validity when used for the purposes defined by the test maker but not yield valid interpretations for a user who defines the content domain in a different way. Only the user of a test can ultimately judge its validity for his or her purpose. Brennan (2001) stated, "For test users, the single best thing to do in advancing proper score use and interpretation is to take the test, or at least, study its content" (p.12).

Although **content-related validity evidence** is especially important for achievement tests, it is also a concern for other types of measuring instruments, such as personality and aptitude measures. An instrument for measuring attitudes toward capital punishment, for example, would be examined to ensure that it contains, in sufficient number, a balanced set of positive and negative statements about capital punishment. An academic aptitude test should measure skills and abilities judged to be important to success in academic tasks. If you were developing a test to select among applicants for a particular job, you would need to specify all the major aspects of the job and then write test items that measure each aspect.

**Face validity** is a term sometimes used in connection with a test's content. Face validity refers to the extent to which examinees believe the instrument is measuring what it is supposed to measure. The question is, "on the face of it," does the test appear to be valid? Although it is not a technical form of validity, face validity can be important to ensure acceptance of the test and cooperation on the part of the examinees. Students taking a test to qualify for an advanced chemistry class would not expect it to contain items dealing with world history or geography.

### 2. Evidence Based on Relations to a Criterion

**Criterion-related validity evidence** refers to the extent to which test scores are systematically related to one or more outcome criteria. The emphasis is on the criterion because one will use the test scores to infer performance on the criterion. Historically, two types of criterion-related validity evidence have been distinguished: concurrent and predictive. The distinction is made on the basis of the time the criterion data are collected.

*Concurrent Validity* **Concurrent validity evidence** is the relationship between scores on a measure and criterion scores obtained at the same time. Assume a researcher has developed a foreign language aptitude test and needs evidence that the test really measures foreign language aptitude. The researcher could select a well-known and previously validated foreign language aptitude test, administer it and the new test to a group of students, and determine the correlation between the two sets of scores. A substantial correlation between the new test and the widely accepted test is evidence that the new test is also measuring foreign language aptitude. Other criteria available at the time might be current grades in a foreign language class or scores on a teacher-made test. Or assume a researcher at Educational Testing Service has developed a new scholastic aptitude test that might replace the more expensive Scholastic Assessment Test (SAT). In order to obtain evidence about the meaningfulness of the scores from this new test, the researcher would administer both the new test and the SAT

(the criterion) to a representative sample of high school students. A substantial correlation between the two sets of scores would indicate that inferences made on the basis of the new test's scores would have validity for measuring scholastic aptitude. A low correlation would indicate that the validity of the inferences based on the new test's scores would be suspect. One would not consider the test a worthwhile replacement for the SAT.

*Predictive Validity* **Predictive validity evidence** is the relationship between scores on a measure and criterion scores available at a future time. In gathering predictive validity evidence of a foreign language aptitude test, one would look at the relationship between scores on the test and the grades students eventually earned in a future foreign language course (criterion). If a relationship is demonstrated, the scores on the aptitude test could be used later to predict performance in foreign language courses. In the case of a new scholastic aptitude test, predictive validity evidence would involve administering the test to a sample of high school juniors or seniors and then putting the scores away until the students complete their first semester or two of college. When the students' college grade point averages (GPAs) become available, one would correlate the test scores and GPAs. If the correlation were high, one has evidence for the usefulness of the aptitude test for predicting college achievement. Large numbers of high school students take the SAT or the ACT test each year because evidence has revealed a correlation between SAT and ACT scores and freshman college GPA. Likewise, the GRE is used in the selection process for admission to graduate school because there is evidence that scores on the GRE are correlated with achievement in graduate school and thus have validity for predicting future achievement. Figure 9.2 illustrates concurrent- and predictive-related evidence used in the validation of an aptitude test.

*Choosing the Criterion* The choice of the criterion and its measurement are crucial in criterion-related evidence. What does one look for when choosing a criterion?

1. The worth of the entire procedure depends first and foremost on the *relevance* of the criterion. The criterion must well represent the attribute being measured or else it would be meaningless to use it. For example, GPA is considered a relevant measure of success in college and is generally chosen

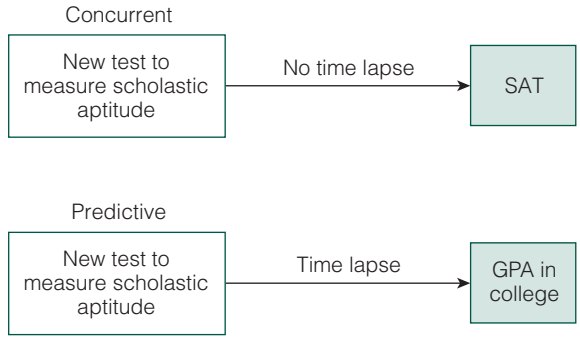



**Figure 9.2** Criterion-Related Evidence of Validity

as the criterion for validation studies of scholastic aptitude tests. A relevant criterion for a test designed to select salespeople might be the dollar value of sales made in a specified time. Supervisor ratings might be used as a criterion in the validation of a test designed to predict success in data-entry positions at a corporation. It is sometimes difficult to find a relevant criterion measure, as in the validation of measures designed to predict teacher effectiveness. With neither an agreed-on description of teacher effectiveness nor an effective method of measuring that variable, it is extremely difficult to validate such instruments.

2. The criterion must also first be *reliable*, which means that it is a consistent measure of the attribute over time or from situation to situation. If the criterion is not consistent, you would not expect it to relate consistently to any tests.
3. The criterion should also be *free from bias*, which means that the scoring of the criterion measure itself should not be influenced by any factors other than actual performance on the criterion. For example, if ratings are used as the criterion, it is essential that the raters be trained and be very careful not to let any factors other than actual performance influence their ratings. The criterion may also be biased through contamination, which occurs when scores on the criterion are influenced by the scorer's knowledge of the individuals' predictor test scores. For example, assume that the criterion used to validate an art aptitude test is grades in art class. If the teachers who grade the students' work are aware of the students' scores on the aptitude test, this awareness may influence the teachers' evaluation of the students and hence the grades. This type of contamination of the criterion can be prevented by not permitting the person who grades or rates the criterion to see the original scores on the test.

### Validity Coefficient

The coefficient of correlation between test scores and criterion is called a **validity coefficient** ($r_{xy}$). Like any correlation coefficient, the size of a validity coefficient is influenced by the strength of the relationship between test and criterion and the range of individual differences in the group. As usual, the nearer the coefficient is to 1.00 (+ or −), the stronger the evidence is that the test is useful for the stated purpose.

Validity coefficients indicate whether the test will be useful as a predictor or as a substitute measure. If it has been shown that a test has a high correlation with a future criterion, then that test can later be used to predict that criterion. Accumulating predictive evidence requires time and patience. In some cases, researchers must wait for several years to determine whether performance on a measure is useful for predicting success on a criterion.

Concurrent criterion-related validity evidence is important in tests used for classification, certification, or diagnosis. For example, one would seek concurrent validity evidence for a new psychiatric screening device by examining its correlation with a well-established instrument already available. If there is a substantial correlation between the new test and the established instrument, one would assume they are measuring the same construct, and the new test could

be used as a substitute for the older instrument. Concurrent validity evidence is necessary when new tests are designed to replace older tests that may be more expensive or more difficult and time-consuming to administer.

Students often ask, "How high does a validity coefficient need to be?" As a general rule, the higher the validity coefficient, the better the evidence. But whether high or low, useful or not useful, depends on the purpose of the test and the context in which it is to be used. A correlation coefficient of .40 could be very helpful in cases for which no predictive instrument has previously been available. In other cases, a correlation of .65 might be considered low and unsatisfactory if other predictors are available that have a higher relationship with the criterion. In general, an instrument has "good" validity as a selection device if evidence shows it has a higher correlation with the criterion than do competing instruments.

### 3. Construct-Related Evidence of Validity

**Construct-related evidence of validity** focuses on test scores as a measure of a psychological construct. To what extent do the test scores reflect the theory behind the psychological construct being measured? Recall that psychological constructs such as intelligence, motivation, anxiety, or critical thinking are hypothetical qualities or characteristics that have been "constructed" to account for observed behavior. They cannot be seen or touched or much less measured directly. How does one know that a measure of a particular construct really reflects this hypothetical characteristic? The test developer of such a measure would have to provide evidence that the scores really reflect the construct in question. The process begins with a definition of the construct based on the theory and previous research. The test developer then specifies the aspects of the construct that are to be measured in the test and develops items that require test takers to demonstrate the behaviors that define the construct. One collects any logical and empirical evidence that supports the assertion that a test measures the construct as defined and not something else. Construct-related evidence is more comprehensive than content- and criterion-related evidence and subsumes the other types. In general, any information that sheds light on the construct being measured is relevant.

Let us consider some of the strategies used to gather construct-related evidence.

1. *Related measures studies:* The aim is to show that the test in question measures the construct it was designed to measure and not some other theoretically unrelated construct. The *Standards* (1999) distinguishes two types of evidence based on relations to other variables: **convergent** and **discriminant.** "Relationships between test scores and other measures intended to assess *similar* constructs provide convergent evidence, whereas relationships between test scores and measures purportedly of *different* constructs provide discriminant evidence" (*Standards*, p. 14). In the case of convergent evidence, the researcher tries to show that the intended construct is being measured; in the case of **divergent evidence,** he or she shows that a wrong construct is not being measured. A mathematical reasoning test would be expected to correlate with grades in mathematics or with other math reasoning tests (convergent evidence). The math test and these

other measures correlate because they all converge on the same construct. Conversely, the scores on the math reasoning test would be expected to have little or no relationship (discriminant evidence) with measures of other skills, such as reading. If a substantial correlation is found between the math test and the reading test, then the math test is being affected by reading ability, and instead of measuring mathematical reasoning, it is really measuring reading ability. Such evidence would lead one to conclude that the math test is not measuring the intended construct (math reasoning) and thus would not yield valid interpretations about math reasoning. Of course, a mathematical reasoning test will inevitably involve some reading skill, so one would not expect a zero correlation with a reading test. However, if two mathematical reasoning tests are alike in all other aspects, the one with a correlation of .15 with a reading test would be preferred over the one with v correlation of .35.

In a classic article, Campbell and Fiske (1959) discussed a **multitrait–multimethod matrix** (MTMM) of corrrelation coefficients as a straightforward way to simultaneously evaluate convergent and discriminant validity of a construct. Their approach was based on the belief that measures of the same construct should correlate with each other even if they use different methods (convergent validity), and that measures of different constructs should not correlate with each other even if they employ the same method (discriminant validity). To illustrate, let's assume a researcher has a theory about a personality characteristic called teacher warmth and has developed an attitude scale as a measure of this construct. In order to establish its construct validity, he or she would need to show not only that it converges with other measures of teacher warmth but also that it could be distinguished from other teacher traits such as sociability. The researcher could administer the attitude scale (method A) to assess teacher warmth and also assess teacher warmth through face-to-face interviews (method B) with the same group of participants. Sociability would similarly be measured in two ways: scores on an existing attitude scale designed to measure sociability (method C) and through face-to-face interviews (method D) with the same participants.

The next step is to calculate the intercorrelations of the participants' scores on all four measures and present the intercorrrelations in what is called a multitrait–multimethod matrix. Table 9.1 shows hypothetical correlations between teacher warmth measured by A, attitude scale, and B,

**Table 9.1** Multitrait–Multimethod Matrix of Correlations between Two Teacher Traits across Two Methods of Measurement

| | | Warmth | | Sociability | |
|---|---|---|---|---|---|
| | | Scale | Interview | Scale | Interview |
| | | A | B | C | D |
| Warmth | Scale | A | .75 | .30 | .10 |
| | Interview | B | | .25 | .20 |
| Sociability | Scale | C | | | .70 |
| | Interview | D | | | |

interview, and sociability measured by C, attitude scale, and D, interview. Let us look at the correlations that are relevant to the construct validity of the attitude scale for measuring teacher warmth.

The high correlation of .75 between teacher warmth measured by method A (attitude scale) and by method B (interview) is evidence of convergent validity. The low correlations of .30 between teacher warmth measured by method A and sociability measured by method C and .10 between method A and method D are evidence of divergent validity of the teacher warmth measure. These data provide evidence for the construct validity of the teacher warmth attitude scale. Of course, one would want to conduct further analyses involving more traits and more measures to determine if the pattern of correlations fits the theory behind the constructs. The rule is as follows: If there is a good fit between the theory and the data, then keep both the theory and the measures. If not, you need to revise the theory or the measures or both. The previous example, using only two traits and two methods, is the simplest possible form of a multitrait–multimethod analysis.

2. *Known-groups technique:* Another procedure for gathering construct-related evidence is the **known-groups technique,** in which the researcher compares the performance of two groups already known to differ on the construct being measured. One hypothesizes that the group known to have a high level of the construct will score higher on the measure than the group known to have a low level of the construct. If the expected difference in performance is found, one concludes that the test is measuring that construct. You would expect that scores on a musical aptitude test, for instance, would differ for students currently enrolled in a school of music versus an unselected group of college students. If an inventory measures psychological adjustment, the scores of a group previously identified as adjusted and a group previously identified as maladjusted should be markedly different on the inventory.
3. *Intervention studies:* Another strategy for gathering construct-related evidence is to apply an experimental manipulation and determine if the scores change in the hypothesized way. You would expect the scores on a scale designed to measure anxiety to increase if individuals are put into an anxiety-provoking situation. The scores of a control group not exposed to the experimental manipulation should not be affected. If anxiety were manipulated in a controlled experiment and the resulting scores change in the predicted way, you have evidence that the scale is measuring anxiety.
4. *Internal structure studies:* Analyzing the internal structure of a test is another source of evidence that the test is measuring the construct it is supposed to be measuring. This procedure involves showing that all the items making up the test or scale are measuring the same thing—that is, that the test has internal consistency. We would expect that individuals who answer some questions in a certain way would also answer similar questions in the same way. In a scale measuring attitudes toward stem cell research, for instance, one would determine if individuals who support stem cell research were consistent in their agreeing with positive statements and disagreeing with the negative statements in the scale. A procedure called **factor analysis**

provides a way to study the constructs that underlie performance of a test. Factor analysis calculates the correlations among all the items and then identifies factors by finding groups of items that are correlated highly with one another but have low correlations with other groups. More than one factor may be needed to account for the correlations among the items. You then decide if the observed intercorrelations conform to the theory behind the construct being measured. If the theory suggests a single one-dimension construct, then we look for high intercorrelations among all the items. If the theory suggests more than one dimension, we should have subscales to measure each separate dimension. In that case, the subscales should have high internal consistency, but they should not correlate highly with other subscales. A measure of feminism, for example, would probably have several subscales covering family, work, pay, politics, authority relations, and the like. The extent to which the observed item intercorrelations agree with the theoretical framework provides evidence concerning the construct being measured. Further discussion of factor analysis is presented in Chapter 13.

5. *Studies of response processes:* Another way to obtain evidence about how well a test is measuring the construct of interest is to look at the **evidence based on response processes** of individuals actually taking the test. Questioning test takers about the mental processes and skills that they use when responding to the items of a test can provide information about what construct is being measured. If one were gathering validity evidence about a new verbal reasoning test, for instance, one might ask individuals to "think aloud" as they work through the test. This procedure may reveal that the test is measuring verbal reasoning, or it may reveal that other factors such as vocabulary or reading comprehension are being measured. Examining response processes may indicate certain construct-irrelevant factors that differentially influence the performance of different subgroups. Thus, it provides evidence about whether the test scores have the same meaning or can be interpreted in the same way across different subgroups. Table 9.2 summarizes the three major types of evidence for validity.

## VALIDITY GENERALIZATION

A concern in validity studies of educational and employment tests is the extent to which evidence of validity based on test–criterion relationships can be generalized to new settings without further investigations of validity in the new setting. Research shows that test–criterion correlations may vary greatly from time to time and place to place because of the type of criterion measure used, the way the predictor is measured, the type of test takers, and the time period involved.

Validity generalization studies have used meta-analysis, which provides statistical summaries of past validation studies in similar situations. If the meta-analytic database is large and the studies adequately represent the type of situation to which a researcher wishes to generalize, we find support for validity generalization. In other circumstances in which the findings of the meta-analytic studies are less consistent and in which there are more differences between the new and old settings, it is more risky to generalize. Local validation studies providing situation-specific evidence would be more valuable.

**Table 9.2** Types of Evidence for Validity of a Test

| Type | Question | Method |
|---|---|---|
| Content related | Is the test a representative sample of the domain being measured? | Make a logical analysis of the content to determine how well it covers the domain. |
| Criterion related (concurrent) | Does a new test correlate with a currently available test (criterion) so that the new test could be a substitute? | Correlate scores from new test with scores of a criterion available at the time. |
| Criterion related (predictive) | Does a new test correlate with a future criterion so that the test can be used to predict later performance on the criterion? | Correlate test scores with a measure (criterion) available at a future time. |
| Construct related | Does the test really measure the intended construct? | Gather various kinds of evidence: convergent and divergent evidence, known-groups technique, intervention study, internal structure, and response processes. |

## VALIDITY OF CRITERION-REFERENCED TESTS

Recall that criterion-referenced tests are designed to measure a rather narrow body of knowledge or skills. Thus, the main concern in assessing the validity of criterion-referenced tests is *content validity*. The basic approach to determining content validity is to have teachers or subject matter experts examine the test and judge whether it is an adequate sample of the content and objectives to be measured.

Another approach that has been used is to administer the test and divide the examinees into two groups: masters versus nonmasters. Then, one determines the proportion of examinees in each group who answered each item correctly. Valid items are those for which the success rate in the master group is substantially higher than the success rate in the nonmaster group. To be very strict, the success rate on each item should be 100% for masters, whereas nonmasters have a very low or a chance rate of success. In the ideal test, there should be no misclassifications (Thorndike, 2005, p. 192).

## APPLICATION OF THE VALIDITY CONCEPT

Validity is always specific to the particular purpose for which the instrument is being used. "It is incorrect to use the unqualified phrase 'the validity of the test.' No test is valid for all purposes or in all situations" (*Standards*, 1999, p. 17). Validity should be viewed as a characteristic of the interpretation and use of test scores and not of the test itself. A test that has validity in one situation and for one purpose may not be valid in a different situation or for a different purpose. A teacher-made achievement test in high school chemistry might be useful for measuring end-of-year achievement in chemistry but not useful for predicting achievement in college chemistry. A German proficiency test might be appropriate for placing undergraduates in German classes at a university but not be a valid exit exam for German majors. Thus, validation is always a responsibility of the test user as well as of the test developer.

We have viewed "test validation" as a process of gathering different types of evidence (content, criterion-related, and construct) in support of score-based interpretations and inferences. The goal of the process is to derive the best possible case for the inferences we want to make.

### THINK ABOUT IT 9.1

Identify the type of validity evidence (content, concurrent criterion, predictive criterion, or construct related) being gathered in each of the following examples:

a. A test administered to applicants for law school correlates .65 with first semester grades in law school.
b. A group of math professors examine the math placement test administered to freshmen at the university. They conclude that the test is an excellent sample of the math skills students need to succeed in college-level courses.
c. A high school teacher administers a standardized chemistry test and correlates the scores with the scores that students earned the next day on a teacher-made chemistry test.
d. As predicted, the scores for a group of Young Republicans on a scale measuring political conservatism were markedly higher than those for a group of Young Democrats.
e. Scores on a new scale to detect depression are correlated with scores on a well-established scale measuring optimism. The correlation was negligible.

**Answers**

a. Predictive (criterion related)
b. Content related
c. Concurrent (criterion related)
d. Construct related
e. Construct related (theory would predict that measures of depression would not be correlated with measures of optimism; this is divergent construct-related validity evidence)

## RELIABILITY

As we mentioned at the beginning of this chapter, the **reliability** of a measuring instrument is the degree of consistency with which it measures whatever it is measuring. This quality is essential in any kind of measurement. A post office will soon take action to repair a scale if it is found that the scale sometimes underestimates and sometimes overestimates the weight of packages. A bathroom scale would be reliable if it gives you nearly the same weight on five consecutive days. However, if you got vastly different readings on each of the five days, you would consider the scale unreliable as a measure of your weight and would probably replace it. Psychologists and educators are concerned about the consistency of their measuring devices when they attempt to measure such complex constructs as scholastic aptitude, achievement, motivation, anxiety, and the like. They would not consider a scholastic aptitude test worthwhile if it yielded markedly different results when administered to the same students on two occasions within the same time frame. People who use such measuring instruments must identify and use techniques that will help them determine to what extent their measuring instruments are consistent and reliable.

On a theoretical level, reliability is concerned with the effect of error on the consistency of scores. In this world measurement always involves some error. There are two kinds of errors: **random errors of measurement** and **systematic errors of measurement.** Random error is error that is a result of pure chance. Random errors of measurement may inflate or depress any subject's score in an unpredictable manner. Systematic errors, on the other hand, inflate or depress scores of identifiable groups in a predictable way. Systematic errors are the root of validity problems; random errors are the root of reliability problems.

## SOURCES OF RANDOM ERROR

Chance or random error that leads to inconsistency in scores can come from three sources:

1. *The individual being measured may be a source of error*. Fluctuations in individuals' motivation, interest, level of fatigue, physical health, anxiety, and other mental and emotional factors affect test results. As these factors change randomly from one measurement to the next, they result in a change or inconsistency in one's scores. Individuals may make more lucky guesses at one time than another. A student's breaking a pencil point on a timed test would increase the error component in the test results.
2. *The administration of the measuring instrument may introduce error*. An inexperienced person may depart from standardized procedures in administering or scoring a test. Testing conditions such as light, heat, ventilation, time of day, and the presence of distractions may affect performance. Instructions for taking the test may be ambiguous. The scoring procedure may be a source of error. Objectivity and precise scoring procedures enhance consistency, whereas subjectivity and vague scoring instructions depress it.
3. *The instrument may be a source of error*. Brevity of a test is a major source of unreliability. A small sample of behavior results in an unstable score. For example, if a test is very short, those subjects who happen to know the few answers required will get higher scores than they deserve, whereas those who do not know those few answers will get lower scores than they deserve. For example, if a test is given to assess how well students know the capitals of the 50 states but only five questions are asked, it is possible that a student who knows only 10 capitals could get all five questions correct, whereas a student who knows 40 could get none correct. Luck is more of a factor in a short test than in a long test.

If a test is too easy and everyone knows most of the answers, students' relative scores again depend on only a few questions and luck is a major factor. If questions are ambiguous, "lucky" examinees respond in the way the examiner intended, whereas "unlucky" subjects respond in another equally correct manner, but their answers are scored as incorrect.

One element in a physical fitness test for elementary students is the baseball throw. Subjects are instructed to throw a baseball as far as they can, and the distance of the throw is measured. Although the object of the test is to get a score that is typical of a subject's performance, certainly if you had a single subject

throw a baseball on several occasions, you would find that the child does not throw it the same distance every time.

Assume you had each student make a throw on two consecutive days. If you then compared the two scores (distance thrown) for each student, you would find that they were almost never exactly the same. Most of the differences would be small, but some would be moderately large and a few would be quite large. Because the results are inconsistent from one day's throw to the next, one throw is not completely reliable as a measure of a student's throwing ability. Three types of chance, or random, influences lead to inconsistency between scores on the two days:

1. The student may change from one time to another. On one day he or she may feel better than on the other. On one day the student may be more motivated or less fatigued. Maybe the student loses balance when starting to throw the ball, or maybe his or her fingers slip while gripping the ball. Perhaps the student's father, hearing about the task, decides to coach the child in throwing a baseball before the next day.
2. The task may change from one measurement to the next. For example, the ball used one day may be firm, whereas on the second day it may be wet and soggy. One day perhaps the examiner permits the students to take a running start up to the throwing line, whereas on the second day a different examiner permits only a couple of steps. There may be gusts of wind at certain times that help some students more than others.
3. The limited sample of behavior results in a less reliable score. The average of a student's baseball throw scores on two days would yield a better estimate of his or her true baseball throwing skill than one day's score. The average of three days' scores would be a still better estimate and so on.

Reliability is concerned with the effect of such random errors of measurement on the consistency of scores. But some errors involved in measurement are predictable or systematic. Using the example of the baseball throw, imagine a situation in which the instructions for the throw are given in English but not all the subjects understand English. The scores of the non-English-speaking subjects could be systematically depressed because the subjects do not comprehend what they are expected to do. Such systematic errors of measurement are a validity problem. The validity of score-based inferences is lowered whenever scores are systematically changed by the influence of anything other than what you are trying to measure (irrelevant variance). In this instance, you are measuring not only baseball-throwing skill but also, in part, English comprehension.

To decide whether you are dealing with reliability or validity, you determine whether you are considering random errors or systematic errors. If a class is being given the baseball throw test and two balls are being employed, one firm and one soggy, and it is purely a matter of chance who gets which ball, the variation caused by the ball used is a reliability problem. The variation caused by the ball represents random error that affects the consistency of the measurements. If the girls are tested using a dry, firm ball and the boys get a wet, soggy ball, scores are a function of gender as well as of skill, resulting in systematic errors that give rise to a validity problem.

## RELATIONSHIP BETWEEN RELIABILITY AND VALIDITY

Reliability is concerned with how consistently you are measuring whatever you are measuring. It is not concerned with the meaning and interpretation of the scores, which is the validity question. We express the relationship between these two concepts as follows: A measuring instrument can be reliable without being valid, but it cannot be valid unless it is first reliable. For example, someone could decide to measure intelligence by determining the circumference of the head. The measures might be very consistent from time to time (reliable), but this method would not yield valid inferences about intelligence because circumference of the head does not correlate with any other criteria of intelligence, nor is it predicted by any theory of intelligence. So a test can be very reliable but consistently yield scores that are meaningless.

To be able to make valid inferences from a test's scores, the test must first be consistent in measuring whatever is being measured. Reliability is a necessary but not a sufficient condition for valid interpretations of test scores.

## EQUATIONS FOR RELIABILITY

It is generally accepted that all measurements of human qualities contain random error. Although scientists cannot remove all this error, they do have ways to assess the aggregate magnitude of measurement errors. Reliability procedures are concerned with determining the degree of inconsistency in scores caused by random error.

When you administer a test to a student, you get a score, which is called the **observed score.** If you had tested this student on some other occasion with the same instrument, you probably would not have obtained exactly the same observed score because of the influence of random errors of measurement. Assuming that test scores have an error component implies that there is a hypothetical error-free score for an individual that would be obtained if the measurement were perfectly accurate. This error-free value is called the individual's **true score** on the test. The true score is conceptualized as "the hypothetical average score resulting from many repetitions of the test or alternate forms of the instrument" (*Standards*, 1999, p. 25).

We conclude, therefore, that every test score consists of two components: the *true score* plus some *error of measurement*. As noted previously, this error component may be caused by any one, or a combination, of a number of factors associated with variations within the examinee from time to time or with the test and its administration.

The reliability of a test is expressed mathematically as the best estimate of what proportion of the total variance of scores on the test is true variance. As we explained in Chapter 6, variance is an index of the spread of a set of scores. If you administer a test to a group of students, some of the spread (variance) of the students' scores is due to true differences among the group and some of the spread (variance) is due to errors of measurement.

The idea of error component and true component in a single test score may be represented mathematically by Formula 9.1:

$$X = T + E \tag{9.1}$$

where

$X$ = observed score
$T$ = true score component
$E$ = error-of-measurement component

The true score component may be defined as the score an individual would obtain under conditions in which a perfect measuring device is used. The error-of-measurement component can be either positive or negative. If it is positive, the individual's true score will be overestimated by the observed score; if it is negative, the person's true score will be underestimated. Because researchers assume that an error of measurement is just as likely to be positive as it is to be negative, they can conclude that the sum of the errors and the mean of the errors would both be 0 if the same measuring instrument or an equivalent form of the instrument were administered an infinite number of times to a subject. Under these conditions, the true component would be defined as the individual's mean score on an infinite number of measurements. The true score is a theoretical concept because an infinite number of administrations of a test to the same subject is not feasible.

In the usual research situation, the investigator has one measure on each of a group of subjects, a single set of scores, to consider. Each observed score has a true score component and an error score component. It has been shown mathematically that the variance of the observed scores of a large group of subjects ($\sigma_x^2$) is equal to the variance of their true scores ($\sigma_t^2$) plus the variance of their errors of measurement ($\sigma_e^2$) or

$$\sigma_x^2 = \sigma_t^2 + \sigma_e^2 \tag{9.2}$$

Reliability may be defined theoretically as the ratio of the true score variance to the observed score variance in a set of scores, as expressed by the following formula:

$$r_{xx} = \frac{\sigma_t^2}{\sigma_x^2} \tag{9.3}$$

where

$r_{xx}$ = reliability of the test
$\sigma_t^2$ = variance of the true scores
$\sigma_x^2$ = variance of the observed scores

Reliability is the proportion of the variance in the observed scores that is free of error. This notion can be expressed in the following formula, derived from Formulas 9.2 and 9.3:

$$r_{xx} = 1 - \frac{\sigma_e^2}{\sigma_x^2} \tag{9.4}$$

The **coefficient of reliability** $r_{xx}$ can range from 1, when there is no error in the measurement, to 0, when the measurement is all error. When there is no error in the measurement, $\sigma_e^2$ in the reliability formula is 0 and $r_{xx} = 1$.

$$r_{xx} = 1 - \frac{\sigma_e^2}{\sigma_x^2} \quad r_{xx} = 1 - \frac{0}{\sigma_x^2} = 1 - 0 = 1$$

If measurement is all error, $\sigma_e^2 = \sigma_x^2$ and $r_{xx} = 0$.

$$r_{xx} = 1 - \frac{\sigma_e^2}{\sigma_x^2} = 1 - 1 = 0$$

The extent of error is indicated by the degree of departure of the reliability coefficient from 1. A coefficient of .80 on a test, for example, indicates the best estimate is that 80 percent of the observed variance in the scores is true variance and 20 percent is error. Thus, the greater the error, the more the reliability coefficient is depressed below 1 and the lower the reliability. Conversely, if the reliability coefficient is near 1.00, the instrument has relatively little error and high reliability.

## APPROACHES TO RELIABILITY

A test is reliable to the extent that the scores made by an individual remain nearly the same in repeated measurements. That is, individuals will have the same, or nearly the same, rank on the repeated administrations. There are two ways to express the consistency of a set of measurements.

1. The first method indicates the amount of variation to be expected within a set of repeated measurements of a *single* individual. If it were possible to weigh an individual on 200 scales, you would get a frequency distribution of scores to represent his or her weight. This frequency distribution would have an average value, which you could consider the "true" weight. It would also have a standard deviation, indicating the spread. This standard deviation is called the **standard error of measurement** because it is the standard deviation of the "errors" of measuring the weight for one person. With psychological or educational data, researchers do not often make repeated measurements on an individual. Time would not permit such repetition; in addition, the practice and fatigue effects associated with repeated measurement would have an influence on the scores. Thus, instead of measuring one person many times, researchers measure a large group on two occasions. Using the pair of measurements for each individual, they can estimate what the spread of scores would have been for the average person had the measurement been made again and again.
2. The consistency of a set of scores is also indicated by the extent to which each individual maintains the same relative position in the group. With a reliable test, the person who scores highest on a test today should also be one of the highest scorers the next time the same test is given. Each person in the group would stay in approximately the same relative position. The more individuals shift in relative position, the lower the test's reliability. You can compute a coefficient of correlation between two administrations of the same test to determine the extent to which the individuals maintain the same relative position. This coefficient is called a **reliability coefficient** ($r_{xx}$). A reliability coefficient of 1.00 indicates that each individual's relative position on the two administrations remained exactly the same and the test would be perfectly reliable.

Thus, the consistency of a measure is indicated by (1) its standard error of measurement or (2) its reliability coefficient. We discuss standard error of measurement later in the chapter. Let us now consider the various reliability coefficients.

## RELIABILITY COEFFICIENTS

There are three broad categories of reliability coefficients used with norm-referenced tests: (1) coefficients derived from correlating individuals' scores on the same test administered on different occasions (test–retest coefficients), (2) coefficients derived from correlating individuals' scores on different sets of equivalent items (equivalent-forms coefficients), and (3) coefficients based on the relationship among scores derived from individual items or subsets of items within a test (internal-consistency coefficients). The internal-consistency coefficient requires only a single administration of a test, whereas the other coefficients require two administrations.

### Test–Retest Reliability

An obvious way to estimate the reliability of a test is to administer it to the same group of individuals on two occasions and correlate the two sets of scores. The correlation coefficient obtained by this procedure is called a **test–retest reliability coefficient.** For example, a physical fitness test may be given to a class during one week and the same test given again the following week. If the test has good reliability, each individual's relative position on the second administration of the test will be near his or her relative position on the first administration of the test.

The test–retest reliability coefficient, because it indicates consistency of subjects' scores over time, is sometimes referred to as a **coefficient of stability.** A high coefficient tells you that you can generalize from the score a person receives on one occasion to a score that person would receive if the test had been given at a different time. A test–retest coefficient assumes that the characteristic being measured by the test is stable over time, so any change in scores from one time to another is caused by random error. The error may be caused by the condition of the subjects themselves or by testing conditions. The test–retest coefficient also assumes there is no practice effect or memory effect. For example, students may learn something just from taking a test and thus will react differently on the second taking of the test. These practice effects from the first testing will not likely be the same across all students, thus lowering the reliability estimate. If the interval of time is short, there may also be a memory effect; students may mark a question the same way they did previously just because they remember marking it that way the first time. This memory effect tends to inflate the reliability estimate, but it can be controlled somewhat by increasing the time between the first test and the retest. However, if the time between testings is too long, differential learning may be a problem—that is, students will learn different amounts during the interval, which would affect the reliability coefficient. Thus, the period of time between the two administrations is an issue that must be considered.

Because of these problems, the test–retest procedure is not usually appropriate for tests in the cognitive domain. Use of this procedure in schools is largely restricted to measures of physical fitness and athletic prowess.

### Equivalent-Forms Reliability

Researchers use the **equivalent-forms technique** of estimating reliability, which is also referred to as the **alternate-forms technique** or **parallel-forms technique,** when it is probable that subjects will recall their responses to the test items. Here, rather than correlating the scores from two administrations of the same

test to the same group, the researcher correlates the results of alternate (equivalent) forms of the test administered to the same individuals. If the two forms are administered at essentially the same time (in immediate succession), the resulting reliability coefficient is called the **coefficient of equivalence.** This measure reflects variations in performance from one specific set of items to another. It indicates whether you can generalize a student's score to what the student would receive if another form of the same test had been given. The question is, To what extent does the student's performance depend on the particular set of items used in the test? If subjects are tested with one form on one occasion and with an equivalent form on a second occasion and their scores on the two forms are correlated, the resulting coefficient is called the **coefficient of stability and equivalence.** This coefficient reflects two aspects of test reliability: variations in performance from one time to another and variations from one form of the test to another. A high coefficient of stability and equivalence indicates that the two forms are measuring the same skill and measuring consistently over time. This is the most demanding and the most rigorous measure available for determining the reliability of a test.

Designing alternate forms of a test that are truly equivalent is a challenge with this technique of estimating reliability. If a successful design is not achieved, then the variation in scores from one form to another could not be considered error variance. Alternate forms of a test are independently constructed tests that must meet the same specifications—that is, they must have the same number of items, instructions, time limits, format, content, range, and level of difficulty—but the actual questions are not the same. Ideally, you should have pairs of equivalent items and assign one of each pair to each form. In a world geography test, for example, form A might ask, "On what continent is the Nile River?" whereas form B asks, "On what continent is the Amazon River?" Form A might ask, "What is the capital of Italy?" and form B, "What is the capital of France?" The distribution of the test scores must also be equivalent.

The alternate-forms technique is recommended when you want to avoid the problem of recall or practice effect and in cases in which you have available a large number of test items from which to select equivalent samples. Researchers generally consider that the equivalent-forms procedure provides the best estimate of the reliability of academic and psychological measures.

### Internal-Consistency Measures of Reliability

Other reliability procedures are designed to determine whether all the items in a test are measuring the same thing. These are called the **internal-consistency procedures** and require only a single administration of one form of a test.

*Split-Half Reliability* The simplest of the internal-consistency procedures, known as the *split-half*, artificially splits the test into two halves and correlates the individuals' scores on the two halves. Researchers administer the test to a group and later divide the items into two halves, obtain the scores for each individual on the two halves, and calculate a coefficient of correlation. This **split-half reliability coefficient** is like a coefficient of equivalence because it reflects fluctuations from one sample of items to another. If each subject has a very similar position on the two halves, the test has high reliability. If there is little consistency in positions,

the reliability is low. The method requires only one form of a test, there is no time lag involved, and the same physical and mental influences will be operating on the subjects as they take the two halves. A problem with this method is in splitting the test to obtain two comparable halves. If, through item analysis, you establish the difficulty level of each item, you can place each item into one of the two halves on the basis of equivalent difficulty and similarity of content. The most common procedure, however, is to correlate the scores on the odd-numbered items of the test with the scores on the even-numbered items. However, the correlation coefficient computed between the two halves systematically underestimates the reliability of the entire test because the correlation between the 50 odd-numbered and 50 even-numbered items on a 100-item test is a reliability estimate for a 50-item test, not a 100-item test. To transform the split-half correlation into an appropriate reliability estimate for the entire test, the **Spearman–Brown prophecy formula** is employed:

$$r_{xx} = \frac{2r_{\frac{1}{2}\frac{1}{2}}}{1 + r_{\frac{1}{2}\frac{1}{2}}} \tag{9.5}$$

where

$r_{xx}$ = estimated reliability of the entire test
$r_{\frac{1}{2}\frac{1}{2}}$ = Pearson $r$ correlation between the two halves

For example, if we find a correlation coefficient of .65 between two halves of a test, the estimated reliability of the entire test, using the Spearman–Brown formula, would be

$$r_{xx} = \frac{(2)(.65)}{1 + .65} = .79$$

The Spearman–Brown procedure is based on the assumption that the two halves are parallel. Because this assumption is seldom exactly correct, in practice, the split-half technique with the Spearman–Brown correction tends to overestimate the reliability that would be obtained with test–retest or equivalent-forms procedures. Bear this in mind when evaluating the reliabilities of competing tests.

Split-half reliability is an appropriate technique to use when time-to-time fluctuation in estimating reliability is to be avoided and when the test is relatively long. For short tests the other techniques, such as test–retest or equivalent-forms, are more appropriate. The split-half procedure is not appropriate to use with speed tests because it yields spuriously high coefficients of equivalence in such tests. A speed test is one that purposefully includes easy items so that the scores mainly depend on the speed with which subjects can respond. Errors are minor, and most of the items are correct up to the point where time is called. If a student responds to 50 items, his or her split-half score is likely to be 25–25; if another student marks 60 items, his or her split-half score is likely to be 30–30, and so on. Because individuals' scores on odd- and even-numbered items are very nearly identical, within-individual variation is minimized and the correlation between the halves would be nearly perfect. Thus, other procedures are recommended for use with speed tests.*

*There are computer programs for calculating all the reliability formulas in this chapter. We included the formulas and worked examples so you can see *how* the procedures work.

*Homogeneity Measures* Other internal-consistency measures of reliability do not require splitting the test into halves and scoring each half separately. These procedures assess the interitem consistency, or **homogeneity,** of the items. They reflect two sources of error: (1) the content sampling as in split-half and (2) the heterogeneity of the behavior domain sampled. The more heterogeneous the domain, the lower the interitem consistency; conversely, the more homogeneous the domain, the higher the interitem consistency.

*Kuder–Richardson Procedures* Kuder and Richardson (1937) developed procedures that have been widely used to determine homogeneity or internal consistency. Probably the best known index of homogeneity is the **Kuder–Richardson formula** 20 (K–R 20), which is based on the proportion of correct and incorrect responses to each of the items on a test and the variance of the total scores:

$$r_{xx} = \frac{K}{K-1}\left(\frac{s_x^2 - \Sigma pq}{s_x^2}\right) \qquad \text{K–R 20} \quad (9.6)$$

where

$r_{xx}$ = reliability of the whole test
$K$ = number of items on the test
$s_x^2$ = variance of scores on the total test (squared standard deviation)
$p$ = proportion of correct responses on a single item
$q$ = proportion of incorrect responses on the same item

The product $pq$ is computed for each item, and the products are summed over all items to give $\Sigma pq$. K–R 20 is applicable to tests whose items are scored dichotomously (0 or 1); thus, it is useful with test items that are scored as true/false or right/wrong. Many machine-scoring procedures for tests routinely provide a K–R 20 coefficient along with a split-half coefficient.

Another formula, **Kuder–Richardson 21,** is computationally simpler but requires the assumption that all items in the test are of equal difficulty. This assumption is often unrealistic:

$$r_{xx} = \frac{Ks_x^2 - \overline{X}(K - \overline{X})}{s_x^2(K-1)} \qquad \text{K–R 21} \quad (9.7)$$

where

$r_{xx}$ = reliability of the whole test
$K$ = number of items in the test
$s_x^2$ = variance of the scores
$\overline{X}$ = mean of the scores

This method is by far the least time-consuming of all the reliability estimation procedures. It involves only one administration of a test and employs only easily available information. As such, it can be recommended to teachers for classroom use if the test is not machine scored and the K–R 20 cannot be calculated by computer.

For example, suppose a teacher has administered a 50-item test to a class and has computed the mean as 40 and the standard deviation as 6. Applying Formula 9.7, the reliability could be estimated as follows:

$$r_{xx} = \frac{(50)6^2 - 40(50-40)}{6^2(50-1)} = \frac{1800-400}{1764} = .79$$

Because the Kuder–Richardson procedures stress the equivalence of all the items in a test, they are especially appropriate when the intention of the test is to measure a single trait. For a test with homogeneous content (e.g., math test covering fractions), the reliability estimate will be similar to that provided by the split-half. For a test designed to measure several traits, the Kuder–Richardson reliability estimate is usually lower than reliability estimates based on a correlational procedure.

Analysts have shown through deductive reasoning that the Kuder–Richardson reliability for any test is mathematically equivalent to the mean of the split-half reliability estimates computed for every possible way of splitting the test in half. This fact helps explain the relationship between the two procedures. If a test is of uniform difficulty and is measuring a single trait, any one way of splitting that test in half is as likely as any other to yield similar half scores. Therefore, the Spearman–Brown and Kuder–Richardson methods will yield similar estimates. If a test has items of varying difficulty and is measuring various traits, the Kuder–Richardson estimate is expected to be lower than the split-half estimate. For example, suppose a secretarial skills test samples typing, shorthand, spelling, and English grammar skills. In applying the split-half method, the test maker would assign equal numbers of items from each subtest to each half of the test. If the test is doing a good job of measuring this combination of skills, the split-half reliability will be high. The Kuder–Richardson method, which assesses the extent to which all the items are equivalent to one another, would yield a considerably lower reliability estimate.

*Coefficient Alpha* Another widely used measure of homogeneity is **coefficient alpha,** also called **Cronbach alpha** after Lee Cronbach, who developed it in 1951. Coefficient alpha has wider applications than the K–R 20 formula. When items are scored dichotomously, it yields the same result as the K–R 20, but it can also be used when items are not scored dichotomously. The formula for alpha is as follows:

$$\alpha = \left(\frac{K}{K-1}\right)\left(\frac{s_x^2 - \Sigma s_i^2}{s_x^2}\right) \tag{9.8}$$

where

$K$ = number of items on the test
$\Sigma s_i^2$ = sum of variances of the item scores
$s_x^2$ = variance of the test scores (all $K$ items)

The formula for alpha is similar to the K–R 20 except that the $\Sigma pq$ is replaced by $\Sigma s_i^2$, the sum of the variances of item scores. To calculate, you determine the variance of all the scores for *each* item and then add these variances across all items to get $s_x^2$.

Researchers use Cronbach alpha when measures have items that are not scored simply as right or wrong, such as attitude scales or essay tests. The item score may take on a range of values; for example, on a Likert attitude scale the individual may receive a score from 1 to 5 depending on which option was

**Table 9.3** Summary of Reliability Coefficients

| | | Number of Test Forms Required | |
|---|---|---|---|
| | | One | Two |
| Number of Administrations Required | One | Split-half<br>K–R 20<br>Coefficient alpha | Equivalent forms<br>(no time lapse) |
| | Two | Test–retest | Equivalent-forms<br>(time lapse) |

chosen. Similarly, on essay tests a different number of points may be assigned to each answer. Many computer programs for reliability, such as the one included in SPSS, provide a coefficient alpha as the index of reliability.

If the test items are heterogeneous—that is, they measure more than one trait or attribute—the reliability index as computed by either coefficient alpha or K–R 20 will be lowered. Furthermore, these formulas are not appropriate for timed tests because item variances will be accurate only if each item has been attempted by every person.

Table 9.3 presents a summary of the different types of reliability coefficients arranged according to the number of forms and number of administrations required.

## INTERPRETATION OF RELIABILITY COEFFICIENTS

The interpretation of a reliability coefficient should be based on a number of considerations. Certain factors affect reliability coefficients, and unless these factors are taken into account, any interpretation of reliability will be superficial.

1. *The reliability of a test is in part a function of the length of the test*. Other things being equal, the longer the test, the greater its reliability. A test usually consists of a number of sample items that are, theoretically, drawn from a universe of test items. You know from what you have studied about sampling that the greater the sample size, the more representative it is expected to be of the population from which it is drawn. This is also true of tests. If it were possible to use the entire universe of items, the score of a person who takes the test would be his or her true score. A theoretical universe of items consists of an infinite number of questions and is obviously not a practical possibility. You therefore construct a test that is a sample from such a theoretical universe. The greater the number of items included in the test, the more representative it should be of the true scores of the people who take it. Because reliability is the extent to which a test represents the true scores of individuals, the longer the test, the greater its reliability, provided that all the items in the test belong in the universe of items.
2. *Reliability is in part a function of group heterogeneity*. The reliability coefficient increases as the spread, or heterogeneity, of the subjects who take the test increases. Conversely, the more homogeneous the group is with respect to the trait being measured, the lower will be the reliability coefficient. One

explanation of reliability is that it is the extent to which researchers can place individuals, relative to others in their groups, according to certain traits. Such placement is easier when you are dealing with individuals who are more heterogeneous than homogeneous on the trait being measured. It does not take a sensitive device to determine the placement of children in a distribution according to their weights when the age range of these children is from 5 to 15 years. In fact, this placement is possible with some degree of accuracy even without using any measuring device. It does take a sensitive device, however, to carry out the same placement if all those who are to be compared and placed in the distribution are 5 years old. Thus, the heterogeneity of the group with whom a measuring instrument is used is a factor that affects the reliability of that instrument. The more heterogeneous the group used in the reliability study, the higher the reliability coefficient. Keep this fact in mind when selecting a standardized test. The publisher may report a high reliability coefficient based on a sample with a wide range of ability. However, when the test is used with a group having a much narrower range of ability, the reliability will be lower.

3. *The reliability of a test is in part a function of the ability of the individuals who take that test*. A test may be reliable at one level of ability but unreliable at another level. The questions in a test may be difficult and beyond the ability level of those who take it—or the questions may be easy for the majority of the subjects. This difficulty level affects the reliability of the test. When a test is difficult, the subjects are guessing on most of the questions and a low reliability coefficient will result. When it is easy, all subjects have correct responses on most of the items, and only a few difficult items are discriminating among subjects. Again, we would expect a low reliability. There is no simple rule by which you can determine how difficult, or how easy, a test should be. That depends on the type of test, the purpose, and the population with which it will be used.
4. *Reliability is in part a function of the specific technique used for its estimation*. Different procedures for estimating the reliability of tests result in different reliability coefficients. The alternate forms with time lapse technique gives a lower estimation of reliability than either test–retest or split-half procedures because in this technique form-to-form as well as time-to-time fluctuation is present. The split-half method, in contrast, results in higher reliability coefficients than do its alternatives because of the speed element in most tests. Thus, in evaluating the reliability of a test, you would give preference to a test whose reliability coefficient has been estimated by the alternate-forms technique, rather than by other techniques, when the reported reliabilities are similar. Standardized test manuals report reliability coefficients based on test–retest and alternate-forms techniques, but teachers generally do not use these procedures for estimating reliability. Repeated testing and alternate forms are not feasible in most classroom situations. Instead, teachers use the split-half, the Kuder–Richardson, or one of the other measures of internal consistency as a measure of reliability.
5. *Reliability is in part a function of the nature of the variable being measured*. Some variables of interest to researchers yield consistent measures more

often than do other variables. For instance, because academic achievement is relatively easy to measure, most established tests of academic achievement have quite high reliability (coefficients of .90 or higher). Aptitude tests that are designed to predict future behavior—a more difficult task—have somewhat lower reliability (.80 or lower). Reliable measures of personality variables are most difficult to obtain; thus, these measures typically have only moderate reliability (.60 to .70).

6. *Reliability is influenced by the objectivity of the scoring*. Inconsistent scoring introduces error that reduces the reliability of a test. The potential unreliability of the scoring of essay tests, for example, means that essay tests are generally considered to be not as reliable as multiple-choice and other types of selected-response tests.

Table 9.4 summarizes the factors affecting reliability.

What is the minimum reliability acceptable for an instrument? Perhaps the best response to this question is that a good reliability is one that is as good as or better than the reliability of competing measures. A spelling achievement test with a reliability of .80 is unsatisfactory if competing tests have reliability coefficients of .90 or better. A coefficient of .80 for a test of creativity would be judged excellent if other tests of the same construct have reliabilities of .60 or less.

The degree of reliability you need in a measure depends to a great extent on the use you will make of the results. The need for accurate measurement increases as the consequences of decisions and interpretation become more important. If the measurement results are to be used for making a decision about a group or for research purposes, or if an erroneous initial decision can be easily corrected, scores with modest reliability (coefficients in the range of .50 to .60) may be acceptable. However, if the results are to be used as a basis for making decisions about individuals, especially important or irreversible decisions (e.g., rejection or admission of candidates to a professional school or the placement of children in special education classes), only instruments with the highest reliability are acceptable. Measurement experts state that in such situations a reliability of .90 is the minimum that should be tolerated, and a reliability of .95 should be the desired standard.

**Table 9.4** Factors Affecting Reliability of a Test

| Factor | Potential Effect |
|---|---|
| 1. Length of the test | The longer the test, the greater the reliability. |
| 2. Heterogeneity of group | The more heterogeneous the group, the greater the reliability. |
| 3. Ability level of group | A test that is too easy or too difficult for a group results in lower reliability. |
| 4. Techniques used to estimate reliability | Test–retest and split-half give higher estimates. Equivalent forms give lower estimates. |
| 5. Nature of the variable | Tests of variables that are easier to measure yield higher reliability estimates. |
| 6. Objectivity of scoring | The more objective the scoring, the greater the reliability. |

**PICTURE THIS**


Joe Rocco

## THINK ABOUT IT 9.2

Indicate the type of reliability coefficient illustrated in each of the following exercises:

a. A teacher prepares two forms of a math achievement test, administers the two forms to a group of students on consecutive days, and correlates the students' scores from the two administrations.
b. A college professor administers a 40-item multiple-choice test in educational psychology. The scoring office provides the professor a reliability index found by dividing the test into two forms and calculating the correlation between the students' scores on the two.
c. A teacher questions the results of a verbal aptitude test administered to her English class. She decides to have the students take the same test on the following day. She then correlates the two sets of scores and finds a coefficient of .90.
d. A commercial test developed two forms of a standardized reading test and administered the two forms of the test to a representative sample of elementary school students in the fall and again in the spring.
e. A teacher wanted a reliability estimate of an essay test in history administered at the end of the semester. She used a computer program that calculated the variance of all the scores for each item and then plugged the total variances across all items into a formula.

**Answers**

a. Alternate forms (coefficient of equivalence)
b. Split-half reliability coefficient

c. Test–retest (coefficient of stability)
d. Alternate forms (coefficient of stability and equivalence)
e. Coefficient alpha

## STANDARD ERROR OF MEASUREMENT

The reliability coefficient looks at the consistency of test scores for a group, but it does not tell us anything about the amount of error in individual test scores. Suppose you had an aptitude test score of 105 for an individual child. If we retested, we would probably not obtain that same score. How much variability could we expect in the child's score on retesting? Recall that measurement theory states that any obtained score is made up of the true score plus an error score: $X = T + E$. Because of error, the obtained score is sometimes higher than the true score and sometimes lower than the true score. Returning to the example of the aptitude test, you would expect with repeated administration to obtain a number of different scores for the same individual. In fact, you would have a frequency distribution of aptitude scores. The mean of this distribution of scores would be the best approximation of the child's true score, and the standard deviation would be an indicator of the errors of measurement. Because this standard deviation is the standard deviation of the errors of measurement, it is called the standard error of measurement. Test theory tells us that the distribution of error scores approximates a normal distribution, and we can use the normal distribution to represent it. Measurement errors are normally distributed with a mean of zero. There may be many small errors, but there will be few very large ones. The standard deviation of this distribution of errors (standard error of measurement, $s_M$) would give an estimate of how frequently errors of a given size might be expected to occur when the test is administered many times.

In practice, you usually do not have repeated measures for an individual but you can get an estimate of the standard error of measurement from one group administration of a test. The formula for standard error of measurement is

$$s_M = s_x\sqrt{1 - r_{xx}}$$

where

$s_M$ = standard error of measurement
$s_x$ = standard deviation of test scores
$r_{xx}$ = reliability coefficient

Thus, using the standard deviation of the obtained scores and the reliability of the test, we can estimate the amount of error in individual scores. If the aptitude test has a reliability coefficient of .96 and a standard deviation of 15, then

$$s_M = 15\sqrt{1 - .96} = 15\sqrt{.04} = 3$$

What does the standard error of measurement tell us? It tells us something about how accurate an individual's score is on a test. We can use what we know about a normal distribution to make statements about the percentage of scores that fall between different points in a distribution. Given a student's obtained score, you

use the $s_M$ to determine the range of score values that will, with a given probability, include the individual's true score. This range of scores is referred to as a **confidence band.** Assuming that the errors of measurement are normally distributed about a given score and equally distributed throughout the score range, you could be 68 percent confident that a person's true score (the score if there were no errors of measurement) lies within one $s_M$ on either side of the observed score. For example, if a subject has an observed score of 105 on an aptitude test where the standard error of measurement is 3, you could infer at the 68 percent confidence level that the subject's true score lies somewhere between 102 and 108. Or you can state at the 95 percent confidence level that the true score will fall within 1.96 (or rounded to 2) $s_M$ of the obtained score (between 99 and 111). You can also use the standard error of measurement to determine how much variability could be expected on retesting the individual. If the subject could be retested on the same aptitude test a number of times, you could expect that in approximately two-thirds of the retests the scores would fall within a range of 6 points of the observed score, and in 95 percent of retests the scores would fall within a range of 12 points. Figure 9.3 shows (a) the distribution of error scores

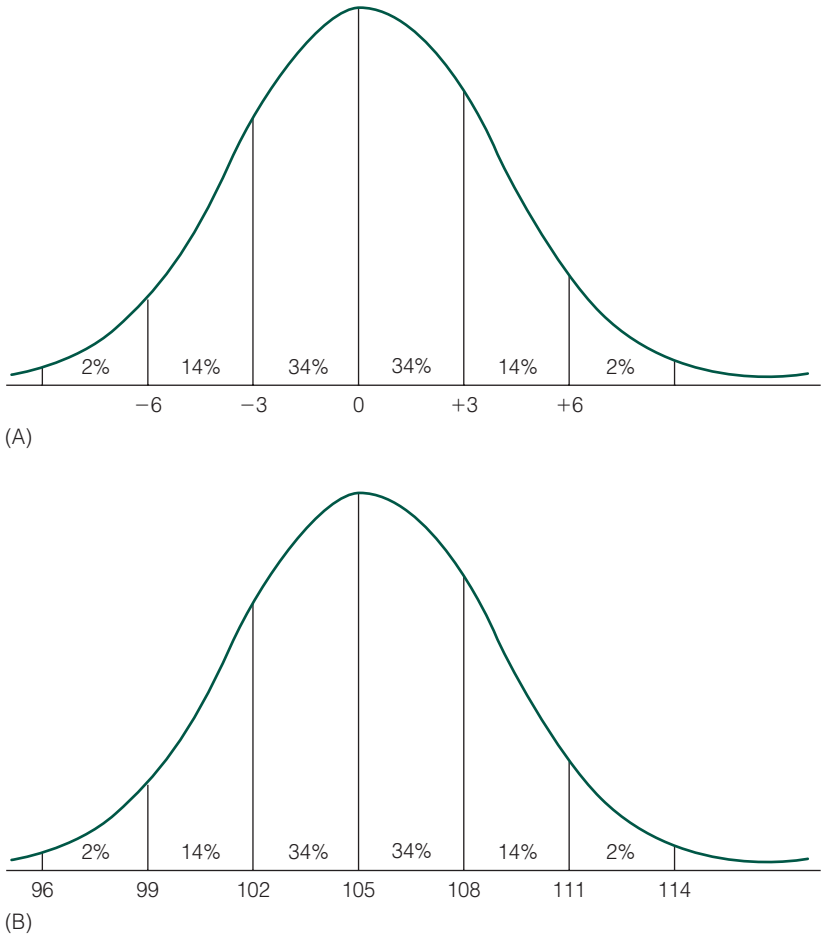



**Figure 9.3** (A) The Distribution of Error Scores When $s_M = 3.00$ and (B) the Distribution around an Obtained Score of 105 with $s_M = 3.00$

(standard error of measurement of the test) and (b) the distribution of errors around an obtained score of 105 with $s_M = 3$.

The standard error of measurement ($s_M$) and the reliability coefficient ($r_{xx}$) are alternative ways of expressing how much confidence we can place in an observed score. The reliability coefficients provide an indicator of the consistency of a group of scores or items making up a test. The standard error of measurement provides an estimate of the consistency of an individual's performance on a test. How accurate or precise an estimate of the true score any observed score will provide is indicated by the size of these two indexes of reliability. As the reliability coefficient increases, the standard error of measurement decreases; as reliability decreases, the standard error of measurement increases. Look for a low standard error of measurement or a high reliability coefficient as an indicator of the stability of test scores. No one method of estimating reliability is optimal in all situations. The standard error of measurement is recommended for use when interpreting individual scores, and the reliability coefficient is recommended for use when comparing the consistency of different tests. You always want scores that are sufficiently consistent to justify anticipated uses and interpretations.

It is meaningless, however, to make a general statement that a test is "reliable." You must report the methods used to estimate the reliability index, the nature of the group from which data were derived, and the conditions under which data were obtained. Potential users of a test then must take responsibility for determining how the reliability data would apply to their population.

### THINK ABOUT IT 9.3

a. A standardized test has a reported reliability coefficient of .84 and a standard deviation of 8. Calculate the standard error of measurement for this test.
b. Mary had a score of 100 on this test. Calculate the band within which Mary's true score is likely to fall. (Use the 95 percent confidence level.)

**Answers**

a. $s_M = s_x\sqrt{1 - r_{xx}}$ $\quad s_M = 8\sqrt{1 - .84} = 8\sqrt{.16} = 8(.4) = 3.2$
b. You can state at the 95 percent confidence level that Mary's true score is between 94 and 106 [$100 \pm (1.96)(3) \approx 100 \pm 6 = 94$ and 106].

## RELIABILITY OF CRITERION-REFERENCED TESTS

The traditional methods used to determine the reliability of norm-referenced tests require sets of scores with considerable variability. Thus, these methods are not appropriate for criterion-referenced tests in which the scores are limited to 1, mastery, or 0, nonmastery. Several procedures have been suggested for estimating the reliability of criterion-referenced tests.

### Agreement Coefficient (ρ)

The **agreement coefficient (ρ)** involves administering two equivalent forms of a criterion-referenced test, or the same test on two occasions, and determining the

consistency of the decisions reached. The consistency is expressed as the percentage of people for whom the same decision (mastery or nonmastery) is made on both forms. This index of reliability is referred to as the agreement coefficient (ρ).

For example, the results displayed in Table 9.5 were obtained when two equivalent forms of a criterion-referenced test were administered to a sample of 100 students. In this case, 70 students were consistently classified as masters on both forms and 14 students were consistently classified as nonmasters.

The agreement coefficient (ρ) is the proportion of the total people consistently classified on the two forms, or

$$\rho = \frac{b + c}{N} \tag{9.9}$$

where

$\rho$ = agreement coefficient
$b$ = number classified as masters on both forms
$c$ = number classified as nonmasters on both forms
$N$ = total number of subjects

$$= \frac{70 + 14}{100} = \frac{84}{100} = .84$$

Thus, 84 percent of the subjects were classified consistently, and .84 is the agreement coefficient of this test. If classifications as master or nonmaster are consistent for all examinees on both administrations of the test, the agreement coefficient equals 1, the maximum value.

Some agreement in classifications as master or nonmaster between two forms is expected merely by chance; that is, even if classifications were made randomly, some individuals would be expected to fall in cells (b) and (c) in Table 9.5. Therefore, we suggest using a statistic proposed by Cohen (1960) that takes *chance agreement* into consideration.

**Table 9.5** Decisions Based on Forms 1 and 2 of a Criterion-Referenced Test

| | | Form 1 | | |
|---|---|---|---|---|
| | | Nonmaster | Master | |
| Form 2 | Master | (a) 10 | (b) 70 | 80 |
| | Nonmaster | (c) 14 | (d) 6 | 20 |
| | | 24 | 76 | 100 ($N$) |

$b$ = number classified as masters on both forms
$c$ = number classified as nonmasters on both forms
$a$ = number classified as nonmasters on form 1 but masters on form 2
$d$ = number classified as masters on form 1 but nonmasters on form 2
$N$ = total number of students who have taken both form.

### Kappa Coefficient

Cohen's **kappa coefficient**, κ, refers to the proportion of consistent classifications observed *beyond* that expected by chance alone. The rationale of the kappa coefficient is straightforward. First, calculate the percentage of cases expected to have consistent classification even if there were no genuine relationship between the forms—that is, if the classification on the two forms were completely independent. This index is referred to as the *expected chance agreement* ($\rho_c$). The expected chance agreement is subtracted from the observed agreement ($\rho_o - \rho_c$) to obtain the actual increase over chance consistency; this quantity is then divided by $1 - \rho_o - \rho_c$, the maximum possible increase in decision consistency beyond chance, to yield κ, the kappa coefficient.

Thus, the expected chance agreement is shown by the following formula:

$$\rho_c = \frac{(a + b)(a + c) + (c + d)(b + d)}{N^2} \tag{9.10}$$

where

$\rho_c$ = proportion of agreement expected by chance

$$\kappa = \frac{\rho_o - \rho_c}{1 - \rho_c} \tag{9.11}$$

where

κ = proportion of agreement *above* that expected by chance
$\rho_o$ = observed agreement coefficient
$\rho_c$ = expected chance agreement

Using the data in the precedin(g example,

$$\rho_c = \frac{(80)(24) + (20)(76)}{100^2} = \frac{1920 + 1520}{10{,}000} = .34$$

$$\kappa = \frac{.84 - .34}{1 - .34} = \frac{.50}{.66} = .76$$

You can see that the kappa coefficient (.76), which adjusts for expected chance agreement, provides a lower estimate of reliability than the agreement coefficient (.84). This is always the case, except when agreement is perfect ($\rho = 1.00$), because kappa begins with the observed agreement and then adjusts it for expected chance agreement. Because kappa is not inflated by chance agreements, it is considered a better indicator of reliability than the agreement coefficient. The agreement coefficient and kappa require two administrations of a test. There are techniques available for estimating the reliability of a criterion-referenced test from a single test administration, but we do not discuss them in this text.

### Phi Coefficient

Another coefficient that is not inflated by chance agreement and thus yields results similar to kappa is phi (ϕ): The **phi coefficient,** a mathematical simplification of the Pearson *r* when all scores are either 1 or 0, is a useful measure of reliability for criterion-referenced measures.

$$\phi = \frac{bc - ad}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \tag{9.12}$$

Using the data in Table 9.5,

$$\phi = \frac{(70)(14) - (10)(6)}{\sqrt{(10+70)(14+6)(10+14)(10+6)}}$$

$$= \frac{980 - 60}{\sqrt{(80)(20)(24)(76)}} = \frac{920}{\sqrt{2{,}918{,}400}}$$

$$= \frac{920}{1708.33} = .54$$

Note how close phi (.54) and kappa (.53) are. The phi coefficient is interpreted in the same way as the other forms of the Pearson *r*. It ranges from −1.00 (all disagreement) through 0 (no consistency) to +1.00 (all agreement). Note that phi (.54) is near the square of kappa $(.76^2 = .57)$.

### RELIABILITY OF OBSERVATIONAL DATA

Reliability is also important in measuring instruments that require ratings or observations of individuals by other individuals. The researcher in these cases must determine the reliability of the ratings—whether different judges/observers have given similar scores or ratings to the same behaviors. A simple way to determine the reliability of ratings is to have two or more observers independently rate the same behaviors and then correlate the observers' ratings. The resulting correlation is called the **interrater** or **interobserver reliability.** If the behaviors to be observed are well defined and the observers well trained, the reliability of the observations should be positive and quite high (approximately .90).

Take the case of two individuals who have rated several students in a performance assessment in which the ratings range from 1 (very poor) to 10 (excellent). Here, reliability can be assessed through correlational procedures in the same way these procedures are used in test–retest or alternate-forms reliability. The second observer serves the same function as a retest or an alternate form in a paper-and-pencil test. When the scores are only 1 or 0 (behavior occurred versus behavior did not occur), the kappa (Formula 9.12) can be used to assess the reliability of the observers' scores. These procedures are also useful when training observers. Trainees watch and score a videotape that has been scored by an experienced observer, and the agreement coefficient, or kappa, indicates the correspondence between a trainee and the experienced observer. The trainer can go through the tape with the trainee to determine when and why the trainee misclassified observations.

The phi coefficient may also be used to assess the agreement of observers scoring 0 and 1. Suen and Ary (1989) provide an extensive discussion of reliability procedures in behavioral observations.

## VALIDITY AND RELIABILITY COMPARED

Validity is a more important and comprehensive characteristic than reliability. Because it is more difficult to measure systematic error than random error, evaluating validity is more challenging. Validity is not obtained as directly as reliability. Assessing validity involves accumulating a great deal of evidence to

support the proposed interpretations of scores. The conceptual framework indicates the kinds of evidence that you need to collect to support the meaning and interpretation of test scores. You must answer questions about the appropriateness of test content, the adequacy of criteria, the definitions of human traits, the specification of the behavioral domain, the theory behind the test content, and so forth. All these matters involve judgment and the gathering of data from many sources. You will find that published research studies typically report much more reliability data than validity data.

Reliability, in contrast, can be investigated directly from the test data; no data external to the measure are required. The basic issues of reliability lend themselves easily to mathematical analysis, and reasonable conclusions about the amount of error can be stated in mathematical terms. Figure 9.4 illustrates the difference between reliability and validity.

If a measure is to yield valid score-based interpretations, it must first be reliable. The reliability of an instrument determines the upper limit of its validity. Scores on a test with zero reliability are entirely random and therefore cannot correlate with any criterion. The possible correlation of an instrument with a criterion (validity coefficient) increases as the reliability of the instrument increases.

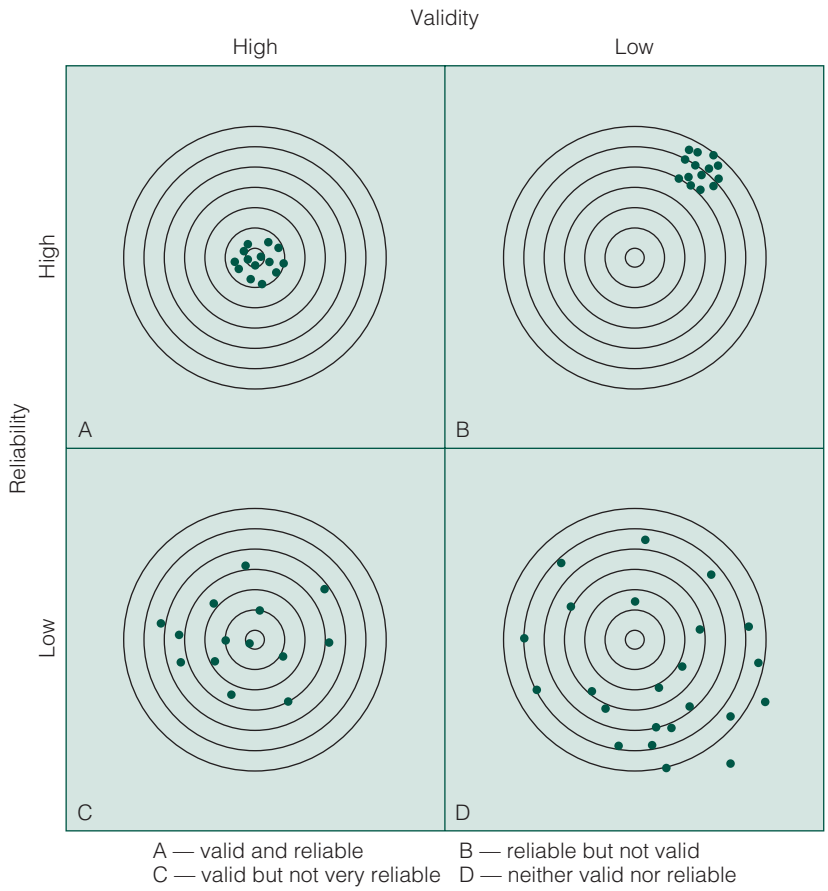



**Figure 9.4** Four Rifles Tested by Aiming at Bull's-Eye and Pulling Trigger

*Source*: From *Contemporary Advertising*, 8th edition, by William Arens. Copyright © 2002 by McGraw-Hill Companies, Inc. Reproduced by permission.

**Table 9.6** Example of a Test Review from Mental Measurements Yearbook

TECHNICAL. The norming sample included $n = 1{,}898$ students from 23 states. For the most part, the sample appears to be representative of nationwide statistics as reported in U.S. Census information with regard to geographic region, gender, family income, educational level of parents, and exceptionality status. The test yields a single raw score. Standard scores, percentile ranks, age, and grade equivalents are provided on easy-to-interpret tables.

RELIABILITY. Evidence of reliability is provided using alternate forms (immediate and delayed), test–retest, and interrater scoring. The reliability coefficients for alternate forms (immediate administration) ranged from .82 to .89 by age level and from .76 to .96 for selected subgroups. The reliability coefficients for test–retest with a 2-week interval ranged from .82 to .95. Evidence of interrater reliability for scorers was high.

VALIDITY. In terms of content validity, the format of the test is more analogous to games like hidden puzzles than reading actual text. As such, it is not a strong reflection of what the authors intend. Some evidence of content validity might be inferred in that the test uses sentences adapted from two well-established reading tests.

For evidence of criterion-related validity, the TOSCRF was compared to archival scores on the Woodcock–Johnson III, the Gray Oral Reading Test (GORT-4), and the Stanford Achievement Test Series 9, and with the Test of Silent Word Reading Fluency (TOSWRF) administered at the time. Average uncorrected correlations across all forms of the TOSCRF ranged from .48 with the GORT-4 to .76 with the TOSWRF. The authors also compare standardized scores from the TOSCRF to a global measure generated from a combination of the other measures, via an independent samples *t*-test. The findings show that the means of the standard scores are similar. The authors interpret this as evidence of validity, but this support seems weak at best.

COMMENTARY. The TOSCRF is acceptable as a quick screening measure for students, as one part of a testing program. Its evidence of reliability and validity need more substantiation. Interpretation of scores for some areas of the U.S. and some subgroups not represented in the norming group should be made with caution.

*Source*: Geisinger, K., Spies, R., Carlson, J., & Plake, B. (Eds.) (2007) ***The seventeenth mental measurements yearbook*** (pp. 797–800). Lincoln: University of Nebraska, Buros Institute of Mental Measurements.

Remember, however, that a measure can have reliability without providing valid interpretations; it can consistently measure the wrong thing. Feldt and Brennan (1989) emphasize the primacy of validity in evaluating the adequacy of an educational measure by stating, "No body of reliability data, regardless of the elegance of the methods used to analyze it, is worth very much if the measure to which it applies is irrelevant or redundant" (p. 143).

Table 9.6 is an excerpt from the *Seventeenth Mental Measurements Yearbook* (Geisinger, Spies, Carlson, & Plake, 2007) showing the kind of validity and reliability data available on published tests. In this case, the instrument is the Test of Silent Contextual Reading Fluency (TOSCRF) designed as "a quick and accurate method of assessing silent reading ability" for individuals from ages 7 to 18 years.

## SUMMARY

Choosing from the multiplicity of measuring instruments available to the researcher requires the use of criteria for the evaluation of these instruments. The two most important criteria for measuring devices are validity and reliability. Validity is the extent to which theory and evidence support the proposed interpretations of test scores for an intended purpose. In the process of assessing validity, the researcher gathers various types of supporting evidence from many sources. Three types of evidence are gathered: (1) content-related evidence, which assesses how well the instrument samples the content domain being measured; (2) criterion-related evidence, which assesses how well the instrument correlates with other measures of the

variable of interest; and (3) construct-related evidence, which assesses how well the instrument represents the construct of interest.

The researcher must also ask, How consistently does the test measure whatever it does measure? This is the issue of reliability. No test can permit meaningful interpretations unless it measures consistently—that is, unless it is reliable. Reliability refers to the extent to which the test is consistent in measuring whatever it does measure. Specifically, reliability refers to the extent to which an individual scores nearly the same in repeated measurements, as indicated by a high reliability coefficient. Reliability coefficients can be computed in various ways, depending on the source of error being considered. The reliability coefficient shows the extent to which random errors of measurement influence scores on the test. The standard error of measurement, another index of reliability, enables researchers to employ the normal curve to estimate the limits within which a subject's true score can be expected to lie.

Validity and reliability procedures appropriate for criterion-referenced tests were discussed in this chapter. Procedures are also available for determining the reliability of observations.

## KEY CONCEPTS

agreement coefficient
alternate-forms technique
coefficient (Cronbach) alpha
coefficient of equivalence
coefficient of reliability
coefficient of stability
coefficient of stability and equivalence
concurrent validity evidence
confidence band
construct-irrelevant variance
construct-related evidence of validity
construct underrepresentation
content-related validity evidence
convergent evidence of validity
criterion-related validity evidence
discriminant evidence of validity
divergent evidence
equivalent-forms technique
evidence based on internal structure
evidence based on response processes
evidence based on test content
face validity
factor analysis
homogeneity measures
internal-consistency procedures
interobserver reliability
interrater reliability
kappa coefficient
known-groups technique
Kuder–Richardson formulas
multitrait–multimethod matrix
observed score
parallel-forms technique
phi coefficient
predictive validity evidence
random errors of measurement
reliability
reliability coefficient
Spearman–Brown prophecy formula
split-half reliability coefficient
standard error of measurement
systematic errors of measurement
test–retest reliability coefficient
true score
validity
validity coefficient

## EXERCISES

1. Compare *validity* and *reliability* with respect to the following:
   a. The meaning of each concept
   b. The relative importance of each concept
   c. The extent to which one depends on the other
2. Explain the following statement: A measuring device may be reliable without being valid, but it cannot be valid without being reliable.
3. How would you propose to gather evidence to support the use of a new scholastic aptitude test that had been developed for use with high school seniors?
4. You have been asked to assess the validity of an instrument designed to measure a student's academic self-concept (i.e., the way he or she views himself or herself as a student). How would you go about this task?

5. What source of evidence supporting the proposed interpretation of test scores is indicated in each of the following situations?
   a. The high school language proficiency test scores of college dropouts and college persisters are compared in order to determine whether the test data correlated with the subjects' college status.
   b. A new scholastic aptitude test is found to have a correlation of .93 with the SAT, which has been used to predict college success.
   c. A new intelligence test has been developed. The author argues that the mental processes required by the test are congruent with the Z theory of intelligence. Furthermore, he shows that among children the average score on the test increases with each year of age.
   d. A teacher carefully examines a standardized achievement test to determine if it covers the knowledge and skills that are emphasized in the class.
   e. The mean difference between the rankings of members of the Ku Klux Klan and members of the Americans for Democratic Action on the liberalism scale was found to be highly significant.
   f. A mathematics test is judged by a group of teachers to be an adequate and representative sample of the universe of test items.
   g. Students are asked to verbalize how they solve mathematics problem-solving items.
6. Identify the type of procedure for estimating reliability that is illustrated in each of the following:
   a. The same test was given twice to a certain group. The correlation between the scores on the two administrations of the test was .90.
   b. The group's scores on the odd items of a test were correlated with their scores on the even items of the same test: $r_{xx} = .95$.
   c. Alternate forms of the test were administered after 1 month, and results of the two administrations were correlated: $r_{xx} = .85$.
   d. The variance, the mean, and the number of items are used to estimate reliability.
7. How would you account for the differences in the reliability coefficients in Exercise 6, assuming that the groups tested were the same?
8. How would you gather evidence for the validity of a reading readiness test?
9. What can you do to increase reliability when constructing a test?
10. Indicate the source of evidence that might be most relevant for assessing validity of the following types of tests:
    a. A classroom history test
    b. An instrument to measure achievement motivation
    c. A measure designed to identify potential dropouts
    d. A group intelligence test
    e. A reading readiness test
11. Explain how a mathematics achievement test could be judged to have high validity in one mathematics class and low validity in another mathematics class.
12. Criticize the following statement: The reliability of the intelligence test is .90. Therefore, you can assume that the test scores can be interpreted as measuring intelligence.
13. Determine the standard error of measurement for a test with a standard deviation of 16 and a reliability coefficient of $r_{xx} = .84$. How would you interpret this standard error of measurement?
14. Select a standardized achievement test that you might use in a research study and obtain the necessary validity data on this test. (You may use *Mental Measurements Yearbook* and the manual that accompanies the test you select.)
15. Check the test manual for the achievement test being used in your school. What type of reliability data are reported there?
16. The following data were obtained when two forms of a criterion-referenced test in mathematics were given to a group of elementary school children. There were 50 items on each form. To pass, a student had to get 80 percent correct on each form. Express the reliability of this test in terms of the kappa coefficient ($\kappa$).

| Examinee | Form 1 | Form 2 |
|---|---|---|
| 1 | 45 | 47 |
| 2 | 43 | 48 |
| 3 | 45 | 31 |
| 4 | 39 | 39 |
| 5 | 39 | 48 |
| 6 | 34 | 37 |
| 7 | 46 | 46 |
| 8 | 48 | 49 |
| 9 | 43 | 38 |
| 10 | 36 | 46 |
| 11 | 45 | 48 |
| 12 | 38 | 39 |
| 13 | 44 | 45 |
| 14 | 31 | 34 |
| 15 | 42 | 48 |

**17.** Criticize the following procedures used to gather validity evidence:

**a.** A high school English teacher developed a writing test for identifying talented high school students and administered the test to her senior English classes. On the basis of high scores, students were permitted to enroll in an English class at the local university. At the end of the semester, the teacher correlated the original test scores with the grades the students earned in the college English class. The teacher was surprised to find a negligible correlation. What was the problem?

**b.** A school counselor developed a scale to measure need for academic achievement in elementary school children. The scale was administered to two classes of elementary school children, and the results were given to the teachers of these children. The teachers were asked to observe these children carefully for one semester, after which they were asked to rate the children on their need for achievement. The teachers' ratings were then correlated with the scores the children received on the scale. The correlation was quite high, so the counselor concluded that the scale had high validity for measuring need for achievement. Do you agree with the counselor's conclusion?

**18.** Assume that you wanted to investigate teacher "burnout." Suggest some indicators of this construct that you might use in developing a scale for this purpose.

**19.** What type of reliability estimate would be most appropriate for the following measuring instruments?

**a.** A multiple-choice achievement test will be used as the dependent variable in an experimental study.

**b.** A researcher will study changes in attitude and will administer one form of an attitude scale as both the premeasure and the postmeasure.

**c.** A researcher has two forms of an achievement test; she administered one form at the beginning of the study and the other at the conclusion of the study. She wants to determine the reliability of the test.

**20.** A 100-item test was split into two halves, and the split-half coefficient of correlation was found to be .60. Calculate the reliability coefficient for the full-length test.

| | Judge 1 | Judge 2 |
|---|---|---|
| Kata | 10 | 9 |
| Ashok | 8 | 7 |
| Mary | 7 | 10 |
| Kwaku | 9 | 8 |
| Anil | 6 | 5 |
| Ester | 4 | 3 |

**21.** Using a 10-point scale, two judges gave the following ratings to the essays written by a group of students. Calculate an index that indicates the reliability of this rating procedure.

**22.** Indicate whether each of the following practices would increase or decrease reliability?

**a.** The teacher decides to give a weekly quiz instead of one major test at the end of the grading period.

**b.** Jane Smith brags about her difficult tests where a large percentage of students fail.

**c.** On Friday afternoon, Miss Jones postponed the major exam until the following Monday after she heard about the football game scheduled for after school that day.

**d.** The teacher decided to add 10 easy test items that everyone could answer correctly.
**e.** The teacher wrote items having a wide range of difficulty, with most items answered correctly by 40 to 70 percent of students.
**f.** To save time, Ms. White had the students do only two of the subtests from a standardized test instead of taking the complete test.
**g.** The teacher decided to give 25 spelling words on the weekly test instead of 10.

**23.** The following are some comments often heard from students following exams. To what test characteristic are the comments most directly related?
**a.** The test measured minute details, not the important concepts emphasized in class.
**b.** The test was too long for the time available.
**c.** That material was not even covered in class.
**d.** The reading level was so complex that the test was really a measure of reading comprehension, not math.
**e.** Many students were observed to be cheating.
**f.** What does this test have to do with choosing students for the advanced chemistry class?

| Test | Mean | Reliability | SD |
|---|---|---|---|
| A | 50 | .75 | 6 |
| B | 50 | .91 | 15 |

**24.** You have the following technical information from two tests:
On which test would a student's score be expected to fluctuate the most on repeated administrations?

**25.** What are the sources of error that affect the reliability of a test? Give an example of each.

**26.** The following types of reliability coefficients were calculated for a test. Which coefficient do you think would be highest? Explain why.
**a.** Test–retest (1 month)
**b.** Parallel forms (1 week)
**c.** Split-half

## ANSWERS

**1.** Validity is the extent to which an instrument measures what it is designed to measure. Reliability is the extent to which an instrument is consistent in measuring whatever it is measuring. Validity is considered a more important aspect than reliability because lack of validity implies lack of meaning. However, an instrument cannot be valid without first being reliable.

**2.** A measure may produce consistent scores (reliability) but may bear no relationship to other accepted measures of the construct or not be able to predict behavior associated with the construct (validity). Scores on a test with zero reliability are entirely random and therefore cannot correlate with any criterion. The extent of reliability sets an upper limit on possible validity.

**3.** You first must define what is meant by *aptitude*. If you wish to measure general academic ability, gather evidence about its content by examining the test items for representativeness. Do they assess the basic academic skills of reading, spelling, math, and so on? Gather evidence about the correlation between the test scores and senior year GPA, college freshman GPA, and other criteria. Correlation with other validated aptitude test scores could also be done.

**4.** The items of the scale or questionnaire would need to cover aspects of the student behavior that would logically be a part of the construct *academic self-concept* (e.g., I intend to go to college). Criterion measures could be personal interviews with students or independent assessment by teachers. Assuming academic self-concept is related to achievement, self-concept scores could be correlated with GPA and/or achievement test scores.

**5. a.** Evidence based on correlation with other variables
**b.** Evidence based on correlation with other variables
**c.** Evidence is construct related
**d.** Evidence based on content
**e.** Evidence based on known-groups technique
**f.** Evidence based on content

**g.** Evidence based on response processes (construct related)

**6. a.** Test–retest reliability
**b.** Split-half reliability
**c.** Alternate forms with time lapse reliability
**d.** Internal consistency (Kuder–Richardson formula 21)

**7.** Split-half reliabilities tend to be higher than test–retest reliabilities because subject variability due to maturation, increase in testing skill, and other random factors is less. Equivalent-forms reliability is lower than same-test reliability because (a) it is impossible to construct exactly equivalent forms and (b) there is an added source of variability when nonidentical forms are used. The internal-consistency reliability will be depressed if the test is not homogeneous.

**8.** You would first identify which specific skills (e.g., letter recognition and left-to-right orientation) comprise reading readiness and then determine if the test incorporated these skills in appropriate proportions. When subjects who have taken the test have begun their reading programs, you would determine how scores on the test and on subtests correlate with reading test scores, teachers' ratings, and other criteria.

**9.** Rewriting ambiguous items, using items of appropriate difficulty, and clarifying instructions will increase reliability. Making a test longer by including additional items drawn from the same universe increases reliability, as does testing on a more heterogeneous group.

**10. a.** Evidence based on content
**b.** Evidence based on internal structure of test, correlation with other criteria of achievement motivation, and performance of contrasted groups
**c.** Evidence based on relationship with some criteria
**d.** Evidence based on internal structure of the test and relationships with appropriate criteria
**e.** Evidence based on internal structure of the test and relationship with appropriate criteria of reading achievement

**11.** A mathematics test that covered only computation would have little validity in a class that stressed concepts and reasoning. If content and emphasis of a different class match the content and emphasis of the test, the test will have high validity in that class.

**12.** A test can be reliable without measuring what it intends to measure. To determine validity, you need to look at content, constructs, and relations with other measures of the same construct as well as relations with measures of behavior assumed to be correlated with the construct.

**13.** By Formula 9.9, you interpret the standard error of measurement as a standard deviation. Thus, you can say that there are two chances in three that the individual's true score will fall in the range of 66.4 score points from the observed score.

$$\begin{aligned} s_M &= s_x\sqrt{1 - r_{xx}} \\ &= 16\sqrt{1 - .84} \\ &= 16\,(.4) \\ &= 6.4 \end{aligned}$$

**14.** Answers will vary.

**15.** Answers will vary.

**16.** A score of 40 represents mastery ($50 \times .80 = 40$).

| | | **Form 1** | | |
|---|---|---|---|---|
| | | Nonmaster | Master | |
| **Form 2** | Master | (b) 2 | (a) 7 | 9 |
| | Nonmaster | (d) 4 | (c) 2 | 6 |
| | | 6 | 9 | 15 |

$$\rho_0 = \frac{7 + 4}{15} = \frac{11}{15} = .73$$

(73% of the students were classified consistently)

$$\rho_c = \frac{(9)(9) + (6)(6)}{15^2} = \frac{81 + 36}{225} = \frac{117}{225} = .52$$

$$\kappa = \frac{.73 - .52}{1 - .52} = \frac{.21}{.48} = .44$$

**17. a.** Selecting just high scorers restricts the variability. The restricted variability lowered the coefficient of correlation.

**b.** There was criterion contamination. Letting the teachers see the results of the original measurement of need for achievement contaminated their ratings of the children on need for achievement.

**18.** There are a number of possible indicators of teacher burnout. You could look at absenteeism from school, lower evaluations by supervisors, incidences of hostility toward students or supervisors, and incidences of challenging of school policies. You might also develop a scale to measure attitudes toward their work; from teachers' own responses to appropriate questions, you might infer the presence of burnout.

**19. a.** You would be interested in the internal consistency of this one form of the test. A split-half, alpha, or Kuder–Richardson reliability coefficient would be appropriate.

**b.** With one form to be used as both a pre- and postmeasure, you would compute a coefficient of stability.

**c.** With two forms and two administrations, you would compute the coefficient of stability and equivalence.

**20.** $r_{xx} = \dfrac{2(.60)}{1 + .60} = .75$

**21.** Pearson $r = .78$

**22. a.** Increase

**b.** Decrease

**c.** Increase

**d.** No effect

**e.** Increase

**f.** Decrease

**g.** Increase

**23. a.** Validity

**b.** Reliability

**c.** Validity

**d.** Validity

**e.** Reliability

**f.** Validity

**24.** Fluctuation would be greater on test B because the standard error of measurement is larger. (Calculate the standard error of measurement for each test.)

**25.** (a) The test itself (too short or ambiguous items); (b) administration of the test (poor directions and distractions during test); and (c) test taker (illness, fatigue, and lack of motivation).

**26.** Split-half because it measures only fluctuation from one-half of the test to another. There is no time lapse; thus, there is only one source of error.

## REFERENCES


American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for educational and psychological tests*. Washington, DC: Author.

Brennan, R. L. (2001). Some problems, pitfalls, and paradoxes in educational measurement. *Educational Measurement: Issues and Practice, 20*(4), 6–18.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin*, *56*, 81–105.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Crocker, L. (2003). Teaching for the test: Validity, fairness, and moral action. *Educational Measurement: Issues and Practice, 22*(3), 5–11.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

Doll, E. A. (1935, 1949, 1965). *Vineland social maturity scale*. Circle Pines, MN: American Guidance Service.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. Brennan (Ed.), *Educational measurement*. New York: American Council on Education and Measurement. [Reprinted 2006 by Greenwood Publishing]

Geisinger, K., Spies, R., Carlson, J., & Plake, B. (Eds.). (2007). *The Seventeenth Mental Measurements Yearbook*. Lincoln: University of Nebraska, Buros Institute of Mental Measurements.

Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika*, *2*, 151–160.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, 50(9), 741–749.

Suen, H. K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data*. Hillsdale, NJ: Erlbaum.

Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education*. Upper Saddle River, NJ: Pearson.

CHAPTER 10

# Experimental Research

Observation is a passive science, experimentation an active science.

Claude Bernard

## INSTRUCTIONAL OBJECTIVES

After studying this chapter, the student will be able to:

1. Describe characteristics of experimental research.
2. Define levels of the independent variable.
3. Explain the concept of control and its importance in experimental research.
4. Distinguish between true control groups and comparison groups in experimental research.
5. Define internal and external validity.
6. Identify threats to internal validity and strategies for avoiding or minimizing them.
7. Identify internal validity problems in research proposals and reports.
8. Describe the difference between random selection and random assignment and show how they are related to internal and external validity.

An **experiment** is a scientific investigation in which the researcher manipulates one or more independent variables, controls any other relevant variables, and observes the effect of the manipulations on the dependent variable(s). An experimenter deliberately and systematically introduces change and then observes the consequences of that change. Only research problems that permit a researcher to manipulate conditions are appropriate for experimental research. The goal of experimental research is to determine whether a causal relationship exists between two or more variables. Because the experiment involves control and careful observation and measurement, this research method provides the most convincing evidence of the effect that one variable has on another.

Early scientists learned the value of observation in the study of our environment but soon realized that nature's complexity could not always be understood through simply observing its many events. They found that events occurring in their "natural" state were often so complicated by irrelevant factors that the operation of the factor they wished to study was obscured. They solved the difficulty by controlling the conditions under which it occurred so that the irrelevant factors were eliminated. Then they could deliberately manipulate the independent variables of interest and measure the changes in the dependent variables that resulted from changes in the independent variables. In other words, they began to perform experiments.

Because the application of experimental methods proved fruitful in investigating the physical world, investigators in other fields began conducting experiments. In the 19th century, great advances were made in zoology, physiology, and medicine through experimentation. Toward the end of the 19th century, scholars began to apply these same methods to psychological research, thus beginning experimental psychology. Rice's (1897) investigation of spelling achievement in the schools marks the first attempt at educational field experimentation. Thorndike (1924) and other early investigators extended the experimental method to education.

In its simplest form, an experiment has three characteristics: (1) An independent variable is manipulated; (2) all other variables that might affect the dependent variable are held constant; and (3) the effect of the manipulation of the independent variable on the dependent variable is observed. Thus, in an experiment the two variables of major interest are the independent variable and the dependent variable. The *independent variable* is manipulated (changed) by the experimenter. The variable on which the effects of the changes are observed is called the *dependent variable*, which is observed but not manipulated by the experimenter. The dependent variable is so named because its value is hypothesized to depend on, and vary with, the value of the independent variable. For example, to examine the effect of different teaching methods on achievement in reading, an investigator would manipulate method (the independent variable) by using different teaching methods in order to assess their effect on reading achievement (the dependent variable).

## THINK ABOUT IT 10.1

Which of the following questions would be appropriate for experimental research?

1. Do high school students who participate in an educational program designed to reduce smoking report smoking fewer cigarettes after 1 year?
2. What do elementary teachers think about retaining low achievers?
3. Are first-born children higher achievers than their younger siblings?
4. Do new teachers who are assigned a mentor report greater satisfaction with teaching than do new teachers not given a mentor?

**Answers**

Questions 1 and 4 because the independent variables can be manipulated.

# CHARACTERISTICS OF EXPERIMENTAL RESEARCH

The essential requirements for experimental research are control, manipulation of the independent variable, and observation and measurement.

## CONTROL

**Control of variables** is the essence of the experimental method. When a study is completed, researchers want to attribute the outcome to the experimental treatment. To do this, they must eliminate all other possible explanations by controlling the influence of irrelevant variables. Without control it is impossible to evaluate unambiguously the effects of an independent variable or to make inferences about causality.

Basically, the experimental method of science rests on two assumptions regarding variables (Mill, 1986/1846):

1. If two situations are equal in every respect except for a variable that is added to or deleted from one of the situations, any difference appearing between the two situations can be attributed to that variable. This statement is called the **law of the single independent variable.**
2. If two situations are not equal, but it can be demonstrated that none of the variables except the independent variable is significant in producing the phenomenon under investigation, or if significant variables other than the independent variable are made equal, then any difference occurring between the two situations after introducing a new variable (independent variable) to one of the systems can be attributed to the new variable. This statement is called the **law of the single significant variable.**

The purpose of control in an experiment is to arrange a situation in which the effect of a manipulated variable on a dependent variable can be investigated. The conditions for applying the law of the single variable are more likely to be fulfilled in the physical sciences than in the behavioral sciences. For example, Robert Boyle applied this principle in formulating his law about the effect of pressure on the volume of a gas: When temperature is held constant, the volume of a gas varies inversely with the pressure on it. Likewise, Jacques Charles formulated a law dealing with the effect of temperature: When pressure is held constant, the volume of a gas varies directly with the temperature.

Because educational research is concerned with human beings, many variables are always present. To attempt to reduce educational problems to the operation of a single variable is not only unrealistic but also perhaps even impossible. Fortunately, such rigorous control is not absolutely essential because many aspects in which situations differ are irrelevant to the purpose of the study and thus can be ignored. It is sufficient to apply the law of the single *significant* independent variable. For example, in a study of the different effects of two methods of teaching arithmetic, an experimenter would wish to have two groups of children who are identical in every respect except the way in which they are taught arithmetic. Because it is impossible to have two absolutely identical groups of children, the experimenter seeks to establish two groups that are as similar as possible with

respect to the variables that are related to arithmetic achievement, such as reading ability, motivation, and general intelligence. Other variables that are highly unlikely to be related to arithmetic, such as athletic ability, height, shoe size, or color of hair, are ignored. Although the law of the single independent variable cannot be followed absolutely, educational experimenters approximate it as closely as possible. Therefore, in experimental studies in education you need procedures that permit you to compare groups on the basis of significant variables. A number of methods of control have been devised to make such comparisons possible.

Assume that you wish to test the hypothesis that children taught by the inductive method (group A) show greater gains in learning scientific concepts than children taught by the deductive method (group B). To draw a conclusion concerning the relationship between teaching method (independent variable) and the learning of scientific concepts (dependent variable), you must rule out the possibility that the outcome is due to some extraneous, usually unmeasured variable(s). An **extraneous variable** is a variable that is not related to the purpose of the study but may affect the dependent variable. In this experiment, aptitude is a factor that certainly affects the learning of scientific concepts; therefore, it would be considered a relevant extraneous variable that you must control. Otherwise, if the children in group A had more aptitude than those in group B, the greater gains in learning by group A could be attributed to aptitude and therefore you could not properly evaluate the effects of the teaching method on learning. Aptitude has confounded the relationship between the variables in which you are interested. The term **confounding** refers to the "mixing" of the variables extraneous to the research problem with the independent variable(s) in such a way that their effects cannot be separated. It could not be determined whether the relation found is (1) between the independent variable and the dependent variable of the study, (2) between the extraneous variables and the dependent variable, or (3) a combination of (1) and (2). Eliminating confounding by controlling for the effect of extraneous variables enables the experimenter to rule out other possible explanations of any observed changes.

In the preceding experiment, the best way to control for aptitude is to randomly assign subjects to the two groups. You could put the names of all the available subjects in a hat and randomly draw one name at a time, assigning the first to group A and the second to group B, the third to A, the fourth to B, and so on. Random assignment increases the probability that the two groups of subjects will be similar on any relevant extraneous variables, such as aptitude, gender composition, socioeconomic level, motivation, and other relevant variables you had not even imagined.

## MANIPULATION

The **manipulation of an independent variable** is a deliberate operation performed by the experimenter. In educational research and other behavioral sciences, the manipulation of an independent variable involves setting up different *treatment* conditions. Treatment is another word for the experimental manipulation of the independent variable. The different treatment conditions administered to the subjects in the experiment are the *levels* of the independent variable. In a study

on the effect of using computer simulations on the learning of science concepts, you have one independent variable with two levels: computer simulation and no computer simulation. Do not confuse one independent variable having two levels with two independent variables. The levels represent two or more values of an independent variable and may involve differences in *degree* or differences in *kind*, depending on the nature of the manipulation. An independent variable whose levels differ in degree or amount involves manipulation in the extent of the independent variable. In contrast, manipulation may involve independent variables with levels that differ in kind. For example, assume an experimenter is interested in the effects of a stimulant on college students' learning of nonsense syllables. The researcher would begin by specifying the stimulant to be used and the amount to be administered. If the researcher is interested in the effect of the stimulant amount on learning, he or she would perhaps set up three levels of the independent variable: high, medium, and low dosage. Or the researcher could compare the effects of one stimulant with another stimulant, or with nothing at all. In this case, the levels differ in kind.

Other examples of independent variables where the levels might differ in amount are sleep deprivation or money used as an incentive. Independent variables, such as different teaching methods (lecture versus discussion) or different instructions given to subjects, have levels differing in kind. A researcher may manipulate more than one independent variable in a single study.

## OBSERVATION AND MEASUREMENT

After applying the experimental treatment, the researcher observes to determine if the hypothesized change has occurred. Some changes can be observed directly, whereas other changes are measured indirectly. Learning, for example, is often the dependent variable in educational research. Researchers cannot measure learning directly. They can only estimate learning through scores on an achievement test or other measures chosen according to the operational definition. Therefore, strictly speaking, the dependent variable is observed scores rather than learning per se. Figure 10.1 illustrates the basic design of an experiment.

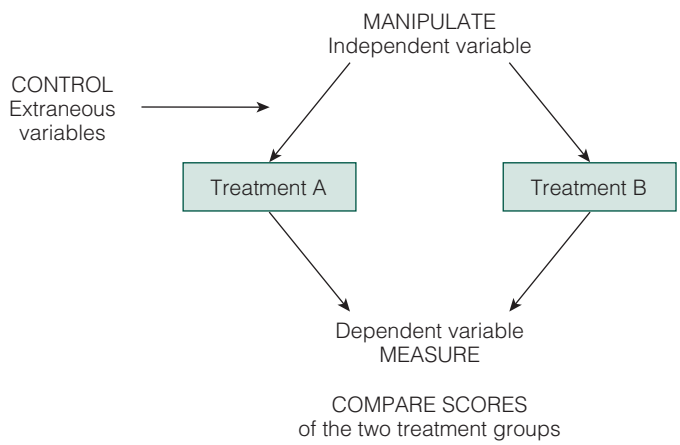



**Figure 10.1** Basic Elements of an Experiment

## EXPERIMENTAL COMPARISON

An experiment begins with an experimental hypothesis, a prediction that the treatment will have a certain effect. The research hypothesis expresses expectations as to results from the changes introduced—that the treatment and no-treatment groups will differ because of the treatment's effects. Recall from Chapter 5 that a null hypothesis must also be developed before the research begins. For the simplest experiment, you need two groups of subjects: the **experimental group** and the **control group.** The experimental group receives a specific treatment; the control group receives no treatment. Using a control group enables the researcher to discount many alternative explanations for the effect of treatment. For example, assume a researcher randomly assigns students to two groups. The experimental group receives a cash reward for successfully completing an object assembly task. If the experimental group does better on the task than the equivalent no-cash-reward group (control group), the researcher has evidence of the relationship between the cash reward and performance on the object assembly task.

More common than comparing a treatment group to a group receiving no treatment (*true control group*) is the situation in which researchers compare groups receiving different treatments. These are called **comparison groups.** The majority of educational experiments study the difference in the results of two or more treatments rather than the difference in the results of one treatment versus no treatment at all. For example, it would be pointless to compare the spelling achievement of an experimental group taught by method A with a control group that had no spelling instruction at all. Instead, researchers compare groups receiving method A and method B treatments. Comparison of groups receiving different treatments provides the same control over alternative explanations, as does comparison of treated and untreated groups. To simplify subsequent discussions, we use the term *control group* to refer both to groups with no treatment and to groups with alternative treatments. Comparisons are essential in scientific investigation. Comparing a group receiving treatment with either an equivalent group receiving no treatment or an equivalent group or groups receiving alternative treatment makes it possible to draw well-founded conclusions from the results.

### THINK ABOUT IT 10.2

Match the term on the left with the appropriate term on the right:

| | |
|---|---|
| 1. Outcome variable | a. Independent variable |
| 2. Treatment variable | b. Extraneous variable |
| 3. Confounding variable | c. Dependent variable |

**Answers**

1. c; 2. a; 3. b

Evaluation follows measurement. Do dependent variable scores differ for the two groups? The experimenter compares the measures of the dependent variable in the first group with the measures of the dependent variable in the other group. The comparison tells the experimenter whether differences on the dependent variable are associated with differences on the independent variable.

## EXPERIMENTAL DESIGN

The term **experimental design** refers to the conceptual framework within which the experiment is conducted. The experimental design sets up the conditions required for demonstrating cause-and-effect relationships. These conditions are as follows: (1) Cause precedes effect in time, (2) the cause variable covaries (occurs together) with the effect, and (3) alternative explanations for the causal relationship can be ruled out.

Experimental designs differ in the number of independent variables that are manipulated. Some experimental designs have only one independent variable; other designs have two or more. Designs differ in the method of assigning subjects to different treatments. In randomized experiments, subjects are randomly assigned to the groups; in other cases, the design uses preexisting groups, or each subject may receive all the treatments. Designs also differ in how often dependent variable measures are made and whether all subjects receive all treatments or not. An experimental design serves two functions: (1) It establishes the conditions for the comparisons required to test the hypotheses of the experiment, and (2) it enables the experimenter, through statistical analysis of the data, to make a meaningful interpretation of the results of the study.

The most important requirement is that the design *must be appropriate* for testing the previously stated hypotheses of the study. The mark of a sophisticated experiment is neither complexity nor simplicity but, rather, appropriateness. A design that will do the job it is supposed to do is the correct design. The hypothesis may state the expected effect of a single independent variable or the effect of two or more variables and the interaction among them. The experimenter's task is to select the design that best arranges the experimental conditions to test the stated hypotheses of the study.

A second requirement is that the design must *provide adequate control* so that the effects of the independent variable can be evaluated as unambiguously as possible. Unless the design controls extraneous variables, you can never be confident of the apparent relationship between the variables of the study. **Randomization** is the single best way to achieve the necessary control. Experimental studies utilizing randomization provide the best evidence for determining the effectiveness of educational practices and programs, and they are considered the gold standard for determining "what works" in educational research. Therefore, the best advice is to select a design that uses randomization in as many aspects as possible.

## VALIDITY OF RESEARCH DESIGNS

Researchers must ask if the inferences drawn about the relationship between the variables of a study are valid or not. A very significant contribution to an understanding of the validity of experimental research designs was made by Campbell and Stanley (1963). They defined two general categories of validity of research designs: *internal validity* and *external validity*. Cook and Campbell (1979) elaborated this previous classification to four types of validity: *internal validity*, *external validity*, *construct validity*, and *statistical conclusion validity*. For a more recent discussion of these four categories of experimental validity, see Shadish, Cook, and Campbell (2002). Table 10.1 presents their classification of the types of validity.

**Table 10.1** Four Types of Validity of Research Designs

| |
|---|
| *Internal validity*: The validity of the inferences about whether the effect of variable A (the treatment) on variable B (the outcome) reflects a causal relationship |
| *Statistical conclusion validity*: The validity of the inferences about the covariation between treatment and outcome |
| *Construct validity*: The validity of the inferences about psychological constructs involved in the subjects, settings, treatments, and observations used in the experiment |
| *External validity*: The validity of the inference about whether the cause–effect relationship holds up with other subjects, settings, and measurements |

*Source*: From Shadish et al., *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, second edition. 

Validity is not a property of an experimental design but, rather, refers to the validity of the inferences. Thus, it is not strictly correct to say that an experiment has internal validity or is internally valid, but we may continue to state it that way because of habit.

## INTERNAL VALIDITY

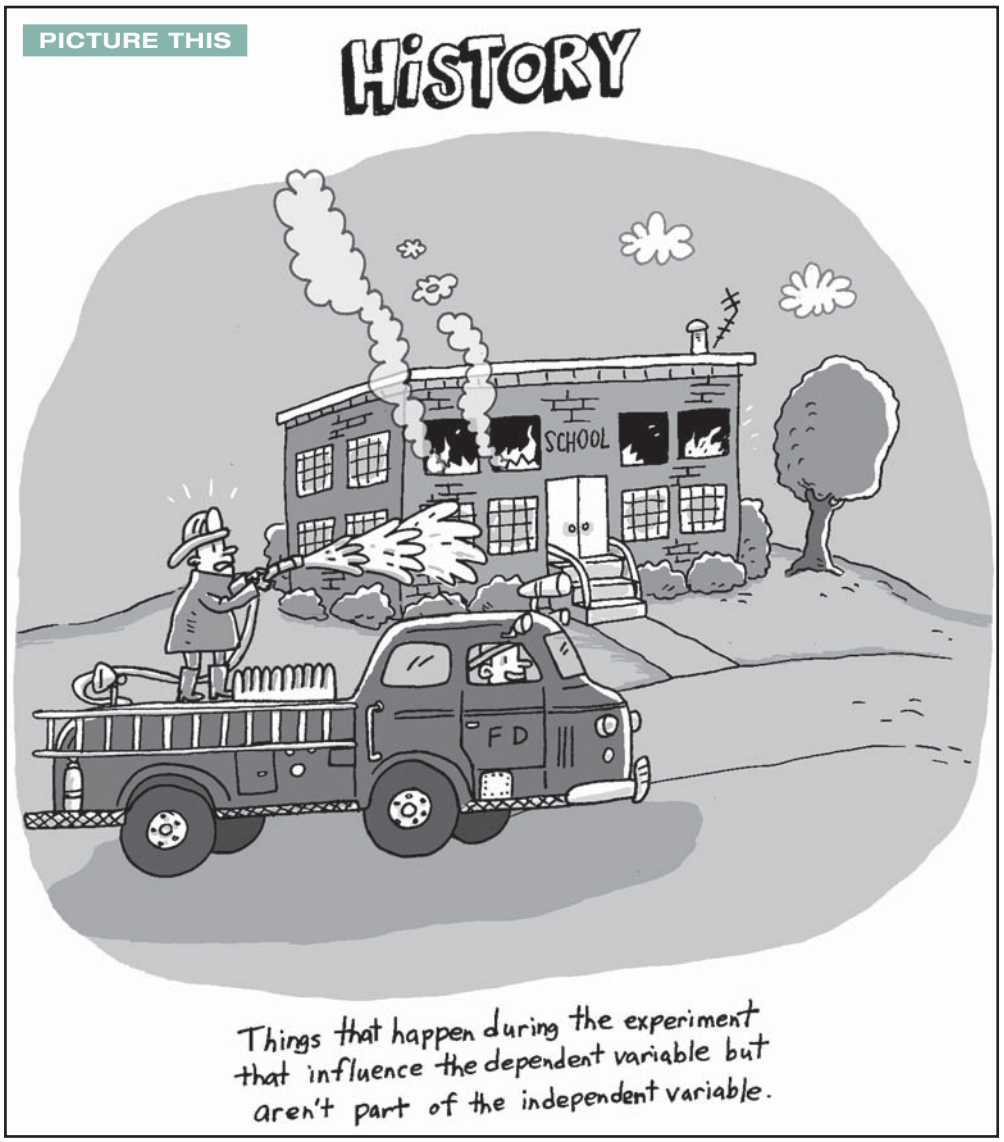
Campbell and Stanley (1963) stated that internal validity is the basic requirement if one is to draw correct conclusions from an experiment. **Internal validity** refers to the inferences about whether the changes observed in a dependent variable are, in fact, caused by the independent variable(s) in a particular research study rather than by some extraneous factors. Internal validity is concerned with such questions as Did the experimental treatment cause the observed change in the dependent variable or was some spurious factor working? and Are the findings accurate? These questions of internal validity cannot be answered positively by the experimenter unless the design provides adequate control of extraneous variables. If the design provides control of variables, you can eliminate alternative explanations of the observed outcome and interpret it as showing an intrinsic relationship between variables. Internal validity is essentially a problem of control. The design of appropriate controls is a matter of finding ways to eliminate extraneous variables that could lead to alternative interpretations and hence lower internal validity. Anything that contributes to the control of a design contributes to internal validity.

Campbell and Stanley (1963) identified eight extraneous variables that frequently represent threats to the internal validity of a research design. These variables are called *threats* because unless they are controlled, they may very well produce an effect that could be mistaken for the effect of the experimental treatment. If uncontrolled, these extraneous variables raise doubts about the accuracy of the experiment because they permit an alternative explanation of the experimental findings.

### Threats to Internal Validity

1. *History*. Specific events or conditions, other than the experimental treatment, may occur between the beginning of the treatment and the posttest measurement and may produce changes in the dependent variable. Such events are referred to as the **history effect.** In this case, history does not

PICTURE THIS




Joe Rocco

refer to past events but to extraneous events occurring *at the same time* that the experimental treatment is being applied and that could produce the observed outcome even without any treatment. These may be major political, economic, or cultural events or some rather minor disruptive factors that occur during the conduct of the experiment. The longer the period of time between the pre- and postmeasurements on the subjects, the greater the history threat becomes.

Assume that a study was undertaken to investigate the use of films to change high school students' attitudes toward the use of selected drugs, including prescription drugs. The research plan called for a pretest to be administered to gauge initial attitudes, then a series of films to be shown over a 2-week period, and finally a posttest to determine what changes may have occurred. During those 2 weeks, the country was shocked by the tragic death of a talented, young

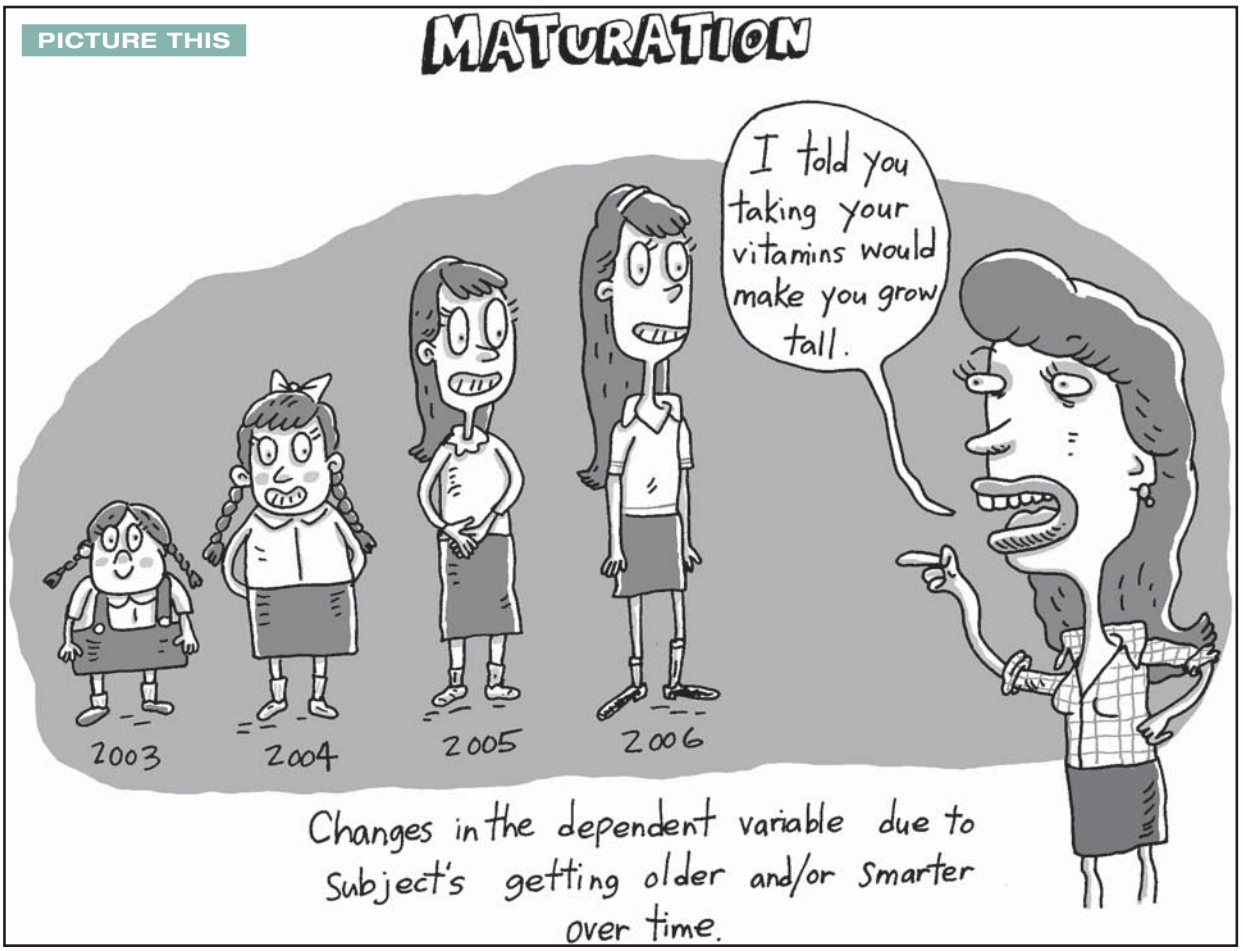
movie actor whose death was believed to be due to an overdose of prescription drugs. The effect of the films on students' attitudes would be confounded with the effects of this tragic event so that the researcher would not know whether any change in attitudes occurred as a result of the films or because of the event and the subsequent extensive media coverage. As another example, during an experiment to measure the effectiveness of an instructional unit on how the stock market works, let's say the stock market drops precipitously and the media devote considerable attention to the stock market. If the students show greater knowledge of the stock market at the conclusion of the unit, the investigators would like to conclude that the instructional unit resulted in greater knowledge. They cannot reach this conclusion, however, because of the events that took place along with the experimental treatment that might have caused the students' greater knowledge. The effects of the unit and of the students' exposure to the media coverage are confounded, and it is impossible to know how much of the students' learning is caused by the unit and how much by events outside the experiment. In this case, history threatens the internal validity of the research.

In general, the use of a control group would eliminate the effects of history but only if both experimental and control groups are affected equally by the event. If one group was affected more than the other, internal validity would be threatened.

2. *Maturation*. The term **maturation** refers to changes (biological or psychological) that may occur within the subjects simply as a function of the passage of time. These changes threaten internal validity because they may produce effects that could mistakenly be attributed to the experimental treatment. Subjects may perform differently on the dependent variable measure simply because they are older, wiser, hungrier, more fatigued, or less motivated than they were at the time of the first measurements. Maturation is especially a threat in research on children because they are naturally changing so quickly. For example, it can be difficult to assess the effects of treatments for articulation problems among preschoolers because young children often naturally outgrow such problems. It has been difficult to assess the effects of compensatory programs such as Head Start on children's cognitive development because normal development ensures that children's cognitive skills will naturally improve over time.

3. *Testing*. Taking a test once may affect the subjects' performance when the test is taken again, regardless of any treatment. This is called the **testing effect.** In designs using a pretest, subjects may do better on the posttest because they have learned subject matter from a pretest, have become familiar with the format of the test and the testing environment, have developed a strategy for doing well on the test, or are less anxious about the test the second time. When an achievement test is used in the research, pretesting is a problem if the same form is used for both the pre- and posttest. We recommend using equivalent forms rather than the same test. Pretesting effects are less threatening in designs in which the interval between tests is large.

With attitude and personality inventories, taking a pretest may sensitize the subjects so that they think about the questions and issues




raised and subsequently give different responses on the posttest (**pretest sensitization**). For example, assume a researcher administers an attitude scale toward an ethnic group, introduces a diversity awareness program, and then gives a posttest to determine whether attitudes changed. The attitude scale itself may stimulate subjects to think about their attitudes; this self-examination, rather than the program itself, may lead to improvements in attitudes.

4. *Instrumentation*. The **instrumentation** threat to internal validity is a result of a change in the instruments used during the study. The change in the way the dependent variable was measured from the first time to the second time, rather than the treatment, may bring about the observed outcome. Changes may involve the type of measuring instrument, the difficulty level, the scorers, the way the tests are administered, using different observers for pre- and postmeasures, and so on. The best advice is to avoid any changes in the measuring instruments during a study. In classroom research, for example, a teacher should not use a multiple-choice pretest and an essay posttest, and the posttest should not be easier or more difficult than the

pretest. Instrumentation is a problem in longitudinal research because the way measures are made may change over a period of time.

5. *Statistical regression*. The term **statistical regression** refers to the well-known tendency for subjects who score extremely high or extremely low on a pretest to score closer to the mean (regression toward the mean) on a posttest. Statistical regression is a threat to internal validity when a subgroup is selected from a larger group on the basis of the subgroup's extreme scores (high or low) on a measure. When tested on subsequent measures, the subgroup will show a tendency to score less extremely on another measure, even a retest on the original measure. The subgroup will have a mean score closer to the mean of the original group. For example, let us assume that the lowest fourth of the scorers on an English proficiency test are selected for a special experimental program in English. The mean of this group will tend to move up toward the mean of the original population on a second test whether or not an experimental treatment is applied. Similarly, high initial scorers would tend to go down toward the population mean on a second testing.

   Let us illustrate regression with a scattergram (Figure 10.2) that shows the pattern we would get if the correlation of fourth-grade reading test scores and fifth-grade reading test scores is $r = .7$. Each dot represents both $z$ scores for an individual. If you select individuals with a particular $z$ score

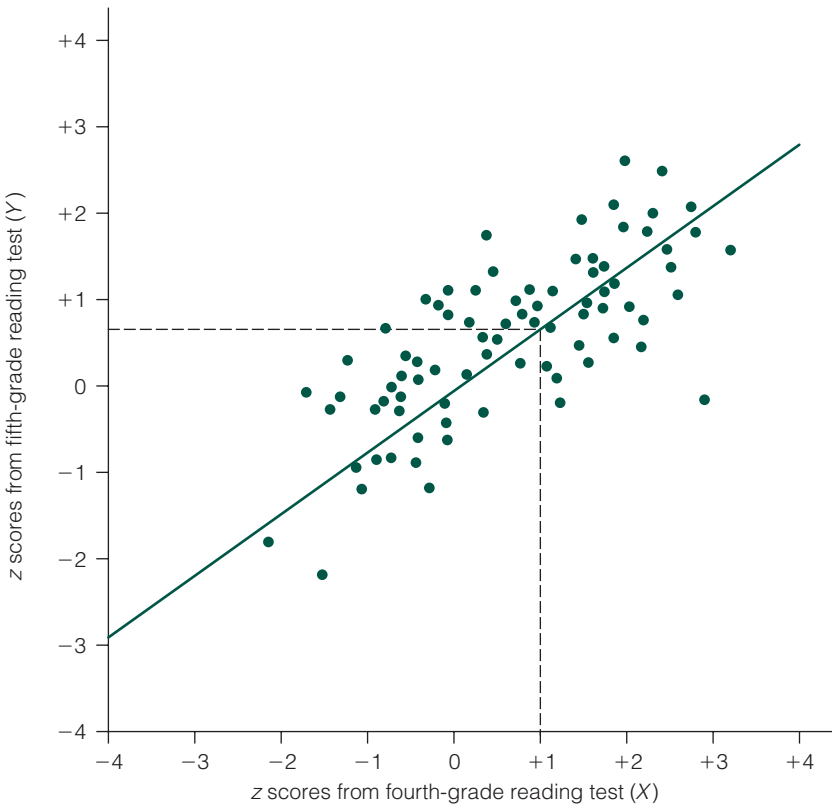
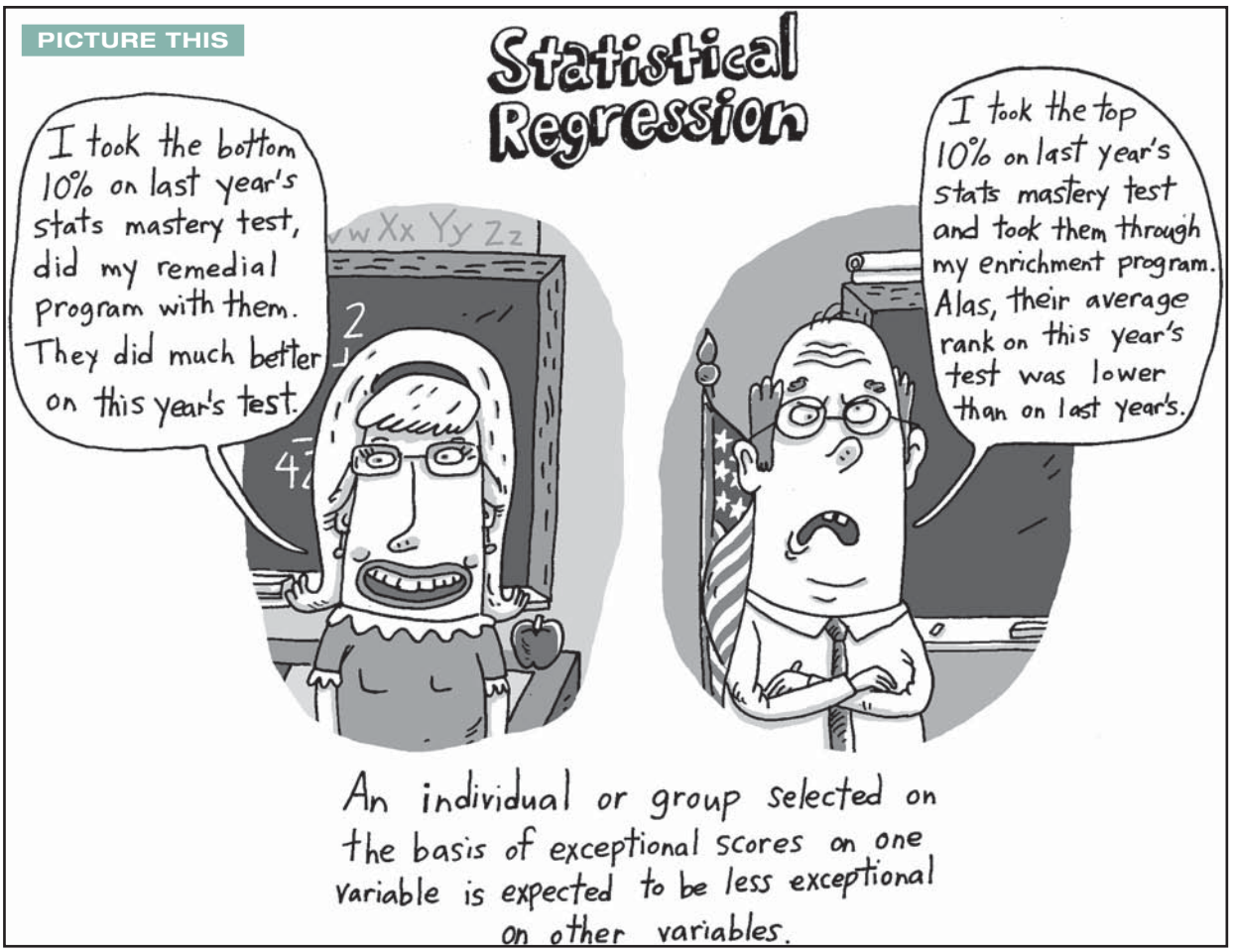



**Figure 10.2** Scattergram for Fourth- and Fifth-Grade Reading Test Scores




on the fourth-grade reading test and look at their scores on the fifth-grade reading test, you find that few have the same $z$ score on both tests. If the fourth-grade $z$ score for this subgroup is above the mean, we will find that a few of the students score farther above the mean in fifth grade than they did in fourth grade, but the majority of them have $z$ scores closer to the mean and some even fall below the mean.

For the subgroup with a $z$ score one standard deviation above the mean on $X$, the mean on $Y$ is $+0.7$. For the subgroup with $z$ scores two standard deviations below the mean on $X$, the mean on $Y$ is $-1.4$. The group with a $z$ score of $+2$ on $X$ has a mean $z$ score of $+1.4$ on $Y$, and so forth. The slanted line on the scattergram connects these means on $Y$ for various scores on $X$. Note that for each group with scores above the mean on the fourth-grade reading test, the mean $z$ score on $Y$ is *lower* on the fifth-grade reading test. For each group with scores below the mean on the fourth-grade reading test, the mean $z$ score is *higher* on the fifth-grade reading test. In other words, the average scores for each group move closer to the mean because of regression.

**THINK ABOUT IT 10.3**

A school established a remedial mathematics program and assigned those middle school students who scored 2 years or more below grade level on a standardized mathematics test to the program. After a semester, the students were given an equivalent form of the standardized test; the majority of the students scored nearer their grade equivalent. The principal was delighted and encouraged other schools to use the program. What would you say to the principal?

**Answer**

You might say "Beware of the regression effect." Before attributing the gains to the program, remember that when one is at the bottom, one has nowhere to go but up. The lowest scorers on the pretest include those whose scores were depressed because of random errors of measurement. The majority of these cases would be expected to score better in a subsequent test.

You must always be aware of the effect of regression in designing your experiments. If dealing with extreme scores is an essential part of your research question, the best solution is to select a large group of extreme scorers and then randomly assign these individuals to two different treatments. Regression will occur equally for each group, so you can determine the effect of the treatment unconfounded with regression. You can reduce regression by using more reliable measures, which are less influenced by random error of measurement.

6. *Selection bias*. **Selection** is a threat when there are important differences between the experimental and control groups even before the experiment begins. A selection bias is a nonrandom factor that might influence the selection of subjects into the experimental or the control group. As a result, there is no assurance that the groups are equivalent. If they are not equivalent before the study, we cannot know whether any difference observed later is due to the treatment or to the pretreatment difference.

   In a learning experiment, for example, if more capable students are in the experimental group than in the control group, the former would be expected to perform better on the dependent variable measure even without the experimental treatment. The best way to control selection bias is to use random assignment of subjects to groups. With random assignment, you cannot determine in advance who will be in each group; randomly assigned groups differ only by chance. We discuss random assignment later in this chapter.

   Selection bias is most likely to occur when the researcher cannot assign subjects randomly but must use intact groups (quasi-experiment). An intact group is a preexisting group such as a class or a group set up independently of the planned experiment.

   Selection bias is also a threat when volunteers are used. People who volunteer for a study may differ in some important respects from nonvolunteers. If the researcher then compares volunteers with nonvolunteers following the experimental treatment, the researcher does not know if the differences are caused by the treatment or by preexisting differences between the two groups.

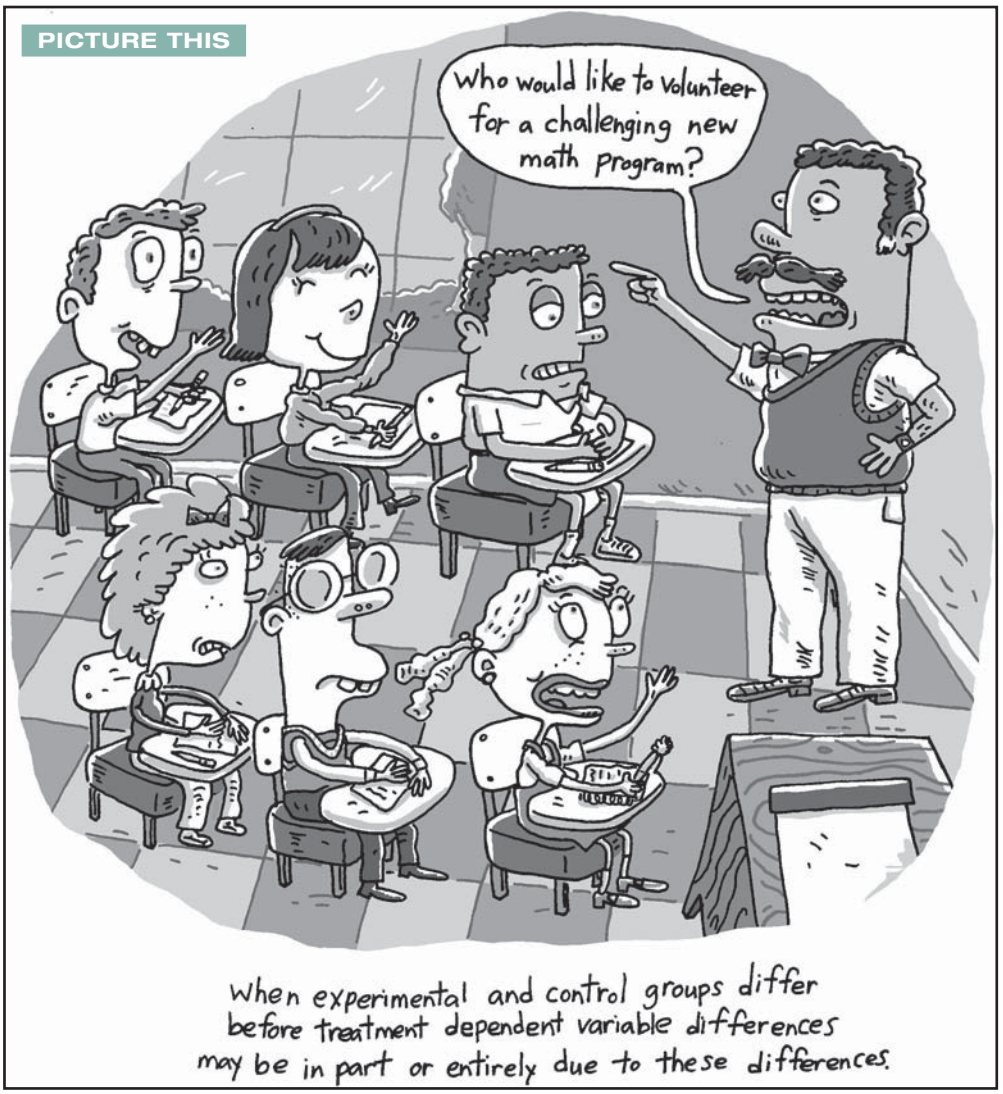
7. *Experimental mortality (attrition).* The **experimental mortality (attrition)** threat occurs when there is differential loss of participants from the comparison groups. This differential loss may result in differences on the outcome measure even in the absence of treatment. If, for example, several of the lowest scorers on a pretest gradually drop out of the experimental group, the remaining subjects will have a higher mean performance on the final measure because the lowest scoring subjects are underrepresented when the posttest is administered.

   Assume a counselor wanted to compare the effectiveness of two different marital counseling procedures. She administered a pretreatment measure of marital adjustment to couples who had enrolled in her program, and then she randomly assigned each couple to procedure A or procedure B. With procedure A, which involved considerable soul searching and confrontation, only 11 of 20 couples persisted in the program and were available for the posttreatment measure. With procedure B, which was relaxed and less demanding, 18 of 20 couples persisted and were available for the posttreatment measure. The posttreatment mean of procedure A was higher than the mean of the procedure B group. This may be evidence that procedure A is more effective, but it may also be evidence that only those couples who were highly motivated to improve their marriage persisted in the program, whereas the posttreatment scores for procedure B included less motivated couples.

   Attrition is not usually a serious threat unless the study goes on for a long time or unless the treatment is so demanding that it results in low-performing participants dropping out.

8. *Selection–maturation interaction*. Some of these threats may interact to affect internal validity. For example, selection and maturation may interact in such a way that the combination results in an effect on the dependent variable that is mistakenly attributed to the effect of the experimental treatment. Such interaction may occur in a quasi-experimental design in which the experimental and control groups are not randomly selected but instead are preexisting intact groups, such as classrooms. Although a pretest may indicate that the groups are equivalent at the beginning of the experiment, the experimental group may have a higher rate of maturation than the control group, and the increased rate of maturation accounts for the observed effect. If more rapidly maturing students are "selected" into the experimental group, the **selection–maturation interaction** may be mistaken for the effect of the experimental variable.

   A group of students who elected to take a freshman honors English class, for instance, may show more vocabulary growth than a group in a regular freshman English class with the same pretest mean because their vocabulary growth was at a higher rate both before and during their freshman year. A comparison of pretest and posttest differences of the honors and the regular classes might lead you to conclude erroneously that the independent variable was responsible for a difference in gain that is really only caused by selection–maturation interaction. Because they elected to take the honors English class, these students may be more motivated to make gains on the dependent variable than are the students in the regular class. The difference in gains may be mistakenly attributed to the independent variable.




Although Campbell and Stanley (1963) originally listed only eight threats to internal validity, Cook and Campbell (1979) suggested that at least three more should be considered.

9. *Experimenter effect*. **Experimenter effect** refers to unintentional effects that the researcher has on the study. Personal characteristics of the researcher, such as gender, race, age, and position, can affect the performance of subjects. Sometimes the actual implementation of the experiment inadvertently gives the experimental group an unplanned advantage over the control group. For example, in an experiment comparing the effectiveness of two teaching methods, the more capable teacher may be assigned to the experimental group. Internal validity is threatened if the experimenter has expectations or a personal bias in favor of one method over another. These preferences and expectancies on the part of the experimenter may be unconsciously transmitted to subjects in such a way that their behavior is affected.

Assume an experimenter is investigating the effectiveness of a new teaching technique that he or she has developed and believes to be superior. If the experimenter is in a position to influence the experiment, he or she may be friendlier or unintentionally behave in a way that improves the performance of the experimental group but not of the control group. An impartial researcher would not necessarily obtain the same findings. Bias could also result if the experimenter who knows which subjects are in which group is asked to rate the performance of these subjects. The experimenter might unintentionally rate the performance of the experimental group higher. Rosenthal (1966) was a pioneer in investigating the experimenter bias effect. He showed that rats labeled "bright" outperformed "dull" ones simply by convincing experimenters that the labels were correct. The higher scores for the bright rats were caused by experimenter bias rather than by any genuine differences between the rats. Experimenter expectancies in research are similar to the Pygmalion effect in education, where teachers' expectancies about student achievement become self-fulfilling prophecies (Rosenthal, 1973).

Note that this bias effect does not result from any deliberate action on the part of the experimenter to alter data or to get the experiment to result a certain way. Instead, the effect comes from subtle, nonverbal cues of which the experimenter may not be aware but that can be detected by the participants, thus influencing their behavior. The best way to reduce experimenter effect is to standardize all procedures or to let other trained individuals (rather than the investigator) work directly with the participants in the study. He or she should also refrain from communicating to the people who administer the experimental treatment any information about the hypotheses of the study. For more information on the experimenter effect, see Rosenthal and Rosnow (1991).

10. *Subject effects*. Subjects' attitudes developed in response to the research situation called **subject effects** can be a threat to internal validity. For instance, in a classic study of the effects of various levels of lighting on worker productivity at the Hawthorne, Illinois, plant of the Western Electric Company, researchers observed that both increases and decreases in light intensity resulted in increased productivity. The researchers concluded that the attention given to the employees and the employees' knowledge that they were participating in an experiment—rather than any changes in lighting—were the major factors leading to the production gains. This tendency for subjects to change their behavior just because of the attention gained from participating in an experiment has subsequently been referred to as the **Hawthorne effect.** This effect can be a problem in educational research that compares exciting new teaching methods with conventional methods. Sometimes subjects may react to what they perceive to be the special *demands* of an experimental situation. That is, subjects react not as they normally might but as they think the more "important" researcher wants them to act. Research has shown, for instance, that subjects who know they are in an experiment tolerate more stress or administer more stress to others than they normally would.

The opposite of the Hawthorne effect is the **John Henry effect.*** This effect, also called **compensatory rivalry,** refers to the tendency of control group subjects who know they are in an experiment to exert extra effort and hence to perform above their typical or expected average. They may perceive that they are in competition with the experimental group and they want to do just as well or better. Thus, the difference (or lack of difference) between the groups may be caused by the control subjects' increased motivation rather than by the experimental treatment. This effect is likely to occur in classroom research in which a new teaching technique is being compared to a conventional method that may be replaced by the new method. The students in the conventional classroom may want to show that they can do just as well as the students being taught by the new method. Furthermore, the teacher in the control classroom may want to show that the old method is just as good and he or she is motivated to make a special effort. For example, early research on televised classroom instruction showed that the teachers in the regular classroom (control group) made a special effort so that their students' performance matched or exceeded the performance of students receiving televised instruction (experimental group). Teachers in the conventional classroom probably felt threatened by this innovation and wanted to show that they could teach better than television.

Another subject effect, called **compensatory demoralization,** occurs when subjects believe they are receiving less desirable treatment or are being neglected. Consequently, they may become resentful or demoralized and put forth less effort than the members of the other group.

11. *Diffusion*. **Diffusion** occurs when participants in one group (typically the experimental group) communicate information about the treatment to subjects in the control group in such a way as to influence the latter's behavior on the dependent variable. Also, teachers involved with the experimental group may share information about methods and materials with teachers of the control group. Assume the subjects in the experimental group being taught math by an innovative method get so excited about the project that they share information with their friends in the control group. Later, the groups may perform similarly on the dependent variable not because the new method was ineffective but because its effects were disseminated to the control group as well. Deemphasizing the fact that an experiment is going on can lessen the likelihood of diffusion problems.

## THINK ABOUT IT 10.4

Identify the potential threat to internal validity in each of the following:

1. A teacher wanted to investigate the effectiveness of new teaching materials designed to increase students' math skills. One of her classes used the new materials, and another class used the conventional materials. Both classes took a math skills test at the end of the semester.
2. Three seventh-grade science teachers used different teaching materials (A, B, and C) for a unit on geology. At the end of the unit, all three classes took the same achievement test.

*This effect is named for John Henry, the legendary railroad worker who died trying to show that he could drive steel railroad spikes as fast as a machine. The experiment was testing the machine as a replacement for human workers.

The class that used the A material had significantly higher scores on the test. The teachers concluded that A material was the best one to use in the future.

3. A teacher administered an attitude scale toward underage drinking among high school students. She then showed a video and used other materials that showed the dangers associated with underage drinking. In order to determine the effectiveness of the video and other materials, she later administered another attitude scale to determine if students changed their attitudes toward drinking.
4. After administering a pretest and then implementing a new instructional method designed for low performers, a teacher noted that posttest results showed improvement for the low performers.
5. A high school started a successful dropout prevention program. The principal was shocked to find that the average annual achievement test scores declined.

**Answers**

1. There may be several threats to internal validity: selection bias (the groups were not equivalent prior to the study), history, experimenter effect (teacher was more enthusiastic about the new materials), and diffusion.
2. Selection bias, history, and diffusion.
3. Pretesting (students may have been sensitized by the initial attitude scale, which influenced the way they responded to the video and other instructional materials).
4. Regression (students who scored low on test 1 scored higher on test 2).
5. Attrition (the low achievers who are most likely to drop out stayed in school and took the achievement test, thus lowering the mean for all students).

In summary, the preceding threats to internal validity represent specific reasons why a researcher's conclusions about a causal relationship between variables may be completely wrong. Researchers must systematically examine how each of the threats may have influenced the results of a study. If the threats can be ruled out, researchers can have more confidence that the observed results were caused by the different treatments. Table 10.2 summarizes the threats to internal validity.

## DEALING WITH THREATS TO INTERNAL VALIDITY

An experiment should be designed to avoid or at least minimize the effect of threats to internal validity.* The researcher's first efforts must be directed toward controlling for any relevant preexisting differences between subjects in the comparison groups. Only in this way can you be fairly confident that any postexperimental differences can be attributed to the experimental treatment rather than to preexisting subject differences. Six basic procedures are commonly used to control intersubject differences and increase equivalence among the groups that are to be exposed to the various experimental situations: (1) random assignment, (2) randomized matching, (3) homogeneous selection, (4) building variables

*Reichardt (2000) stated that it is better to speak of "taking account of threats to validity" than "ruling out threats" because the latter implies a finality that can rarely be achieved in practice. "Ruling out" threats implies an all-or-none quality in which threats either do or do not apply. In many cases, however, threats are a matter of degree rather than being absolute.

**Table 10.2** Threats to Internal Validity

| | |
|---|---|
| History | Unrelated events that occur during the study affect the dependent variable. |
| Maturation | Changes occur within the participants just as a function of time. |
| Testing effect | Exposure to prior test affects posttest. |
| Instrumentation | Unreliability or a change in the measuring instrument affects result. |
| Regression | Extremely high or low scorers on a pretest regress toward mean on a posttest. |
| Selection bias | Because of selection methods, subjects in the comparison groups are not equivalent prior to study. |
| Mortality | A differential loss of participants from the groups affects dependent variable. |
| Selection–maturation interaction | Subjects with different maturation rates are selected into treatment groups. |
| Experimenter effect | Unintentional bias or behavior of experimenter affects results. |
| Subject effect | Attitudes developed during the study affect performance on dependent variable. |
| Diffusion | Participants in experimental group communicate information about treatment to control group, which may affect the latter's performance. |

into the design, (5) statistical control (analysis of covariance), and (6) use of subjects as their own controls.

## RANDOM ASSIGNMENT

The experimenter with an available supply of subjects has the task of dividing them into two groups that will be treated differently and then compared on the dependent variable. In assigning subjects to groups for the study, the experimenter needs a system that operates independently of personal judgment and of the characteristics of the subjects. For example, the known high scorers must not all be assigned to group A and the low scorers to group B. A system that satisfies this requirement is **random assignment.** Random assignment is the assignment of subjects to groups in such a way that for any given placement, every member of the population has an equal probability of being assigned to any of the groups. Chance alone determines whether subjects are placed in the experimental or the control group, thus eliminating selection bias. The term *randomization* is often used as a synonym for random assignment. Randomization is the most powerful method of control because only chance would cause the groups to be unequal with respect to any potential extraneous variables. For example, Ms. Brown has 200 subjects available for an experiment but only has sufficient facilities to include 100 in her experiment. She would first randomly select the 100 for the experiment by numbering all the available subjects and using a table of random numbers to select 100 participants. Thus, she avoids creating systematic differences between the subjects in the experiment and the population from which they were selected. She would then use a random assignment procedure such as a coin toss to put subjects into groups for the experiment. If heads comes up for a subject, then that subject goes into one group; if tails comes up, the subject goes into the other group. A coin toss


would then determine which of the two groups is to get treatment A and which is to get treatment B or the control condition.

Note that random assignment is not the same thing as random selection. Random selection (see Chapter 7) is the use of a chance procedure to select a sample from a population. Random assignment is the use of a chance procedure to assign subjects to treatments.

When subjects have been randomly assigned to groups, the groups can be considered statistically equivalent. **Statistical equivalence** does not mean the groups are absolutely equal, but it does mean that any difference between the groups

is a function of chance alone and not a function of experimenter bias, subjects' choices, or any other factor. A subject with high aptitude is as likely to be assigned to treatment A as to treatment B. The same is true for a subject with low aptitude. For the entire sample, the effects of aptitude on the dependent variable will tend to balance or randomize out. In the same manner, subjects' differences in political viewpoints, temperament, achievement motivation, socioeconomic level, and other characteristics will tend to be approximately equally distributed between the two groups. The more subjects in the original sample, the more likely that randomization will result in approximately equivalent groups.

When random assignment has been employed, any pretreatment differences between groups are nonsystematic—that is, a function of chance alone. Because these differences fall within the field of expected statistical variation, the researcher can use inferential statistics to determine how likely it is that posttreatment differences are due to chance alone.

## RANDOMIZED MATCHING

When random assignment is not feasible, researchers sometimes select pairs of individuals with identical or almost identical characteristics and randomly assign one member of the matched pair to treatment A and the other to treatment B. This procedure is called **randomized matching.** Note that randomized matching requires that the subjects be matched on relevant variables first and then randomly assigned to treatments. The researcher first decides what variables to use for matching. These may be IQ, mental age, socioeconomic status, age, gender, reading, pretest score, or other variables known to be related to the dependent variable of the study. If the groups are adequately matched on the selected variable(s), the resulting groups are reasonably equivalent. The major limitation of matching is that it is almost impossible to find subjects who match on more than one variable. Subjects are lost to the experiment when no match can be found for them. This loss, of course, reduces the sample size and introduces sampling bias into the study. Subjects for whom matches cannot be found are usually those with high or low scores. Therefore, these subjects would be underrepresented.

## HOMOGENEOUS SELECTION

Another method that can make groups reasonably comparable on an extraneous variable is to select samples that are as homogeneous as possible on that variable. This is called **homogeneous selection.** If the experimenter suspects that age is a variable that might affect the dependent variable, he or she would select only children of a particular age. By selecting only 6-year-old children, the experimenter would control for the effects of age as an extraneous independent variable. Similarly, if intelligence is likely to be a variable affecting the dependent variable of the study, then subjects would be selected from children whose IQ scores are within a restricted range—for example, 100 to 110. This procedure has thus controlled the effects of IQ. From this resulting homogeneous population, the experimenter randomly assigns individuals to groups and can assume that they are comparable on IQ. Beginning with a group that is homogeneous on the relevant variable eliminates the difficulty of trying to match subjects on that variable.

Although homogeneous selection is an effective way of controlling extraneous variables, it has the disadvantage of decreasing the extent to which the findings

can be generalized to other populations. If a researcher investigates the effectiveness of a particular method with such a homogeneous sample, such as children with average IQs, the results could not be generalized to children in other IQ ranges. The effectiveness of the method with children of low intelligence or very high intelligence would not be known.

As with matching, a true experiment requires that the subjects be selected first and then assigned randomly to treatments.

## BUILDING VARIABLES INTO THE DESIGN

Some variables associated with the subjects can be built into the experimental design and thus controlled. For example, if you want to control gender in an experiment and you choose not to use the homogeneous selection technique just discussed, you could add gender as another independent variable. You would include both males and females in the study and then use analysis of variance to determine the effects of both gender and the main independent variable on the dependent variable. This method not only controls the extraneous gender variable but also yields information about its effect on the dependent variable, as well as its possible interaction with the other independent variable(s). See Chapter 7.

## STATISTICAL CONTROL

**Analysis of covariance (ANCOVA)** is a statistical technique used to control for the effect of an extraneous variable known to be correlated with the dependent variable. For example, consider an experiment to study the effects of two methods of teaching reading on reading achievement, the dependent variable. Subjects' reading ability before the experiment would be a variable that would certainly be related to the dependent variable of the study. You would expect that those who are good readers to begin with would score well on the reading posttest, whereas those who are poor readers would tend to score more poorly. After randomly assigning half of the subjects to method A and half to method B, you would administer a reading pretest to both groups. At the end of the experiment, ANCOVA would statistically adjust the mean reading posttest scores for any initial differences between the groups on the pretest. The ANCOVA technique removes the portion of each subject's posttest score that is in common with his or her pretest score. The resulting $F$ value can then be checked for statistical significance. The variable used in ANCOVA to adjust scores (in this case, the reading pretest) is called the **covariate.**

Using this technique, you are not considering a subject's posttest score per se. Instead, you analyze the difference between posttest scores and what you would expect the posttest score to be, given the score on the pretest and the correlation between pretest and posttest.

By removing that portion of the dependent variable score variance that is systematically associated with pretest variance, the precision of the experiment improves. With part of the variance in the posttest scores that is *not* caused by treatment removed, any difference caused by treatment stands out more clearly. Using a covariate that is related to the dependent variable reduces the likelihood of Type II error.

Achievement pretest scores are often used as covariates, but other measures such as aptitude or attitude scores can also be used. To be useful, a covariate must

be correlated with the dependent variable. In the example, baseball throw scores or shoe sizes would not be useful covariates because they would be expected to have negligible correlation with the reading posttest.

## USING SUBJECTS AS THEIR OWN CONTROLS

Still another procedure involves **using subjects as their own controls**—assigning the same subjects to all experimental conditions and then obtaining measurements of the subjects, first under one experimental treatment and then under another. For example, the same subjects might be required to learn two different lists of nonsense syllables—one list with high association value and the other with low association value. The difference in learning time between the two lists is found for each subject, and then the average difference in learning time for all subjects can be tested for significance.

This method of control is efficient when feasible, but in some circumstances it cannot be used. In some types of studies, exposure to one experimental condition would make it impossible to use the subjects for the other experimental condition. You cannot, for example, teach children how to divide fractions one way and then erase their memory and teach it another way.

In the foregoing experiment in which one group of subjects was used to investigate the relative ease of learning high-association and low-association nonsense syllables, there could be a "learning to learn" effect, and thus whichever list appeared second would have an advantage over the first. Conversely, fatigue or interference effects might result in poorer performance on the second list. In either case, you cannot separate the effect of the independent variable and the effect of order on the dependent variable. A useful strategy for this experiment would be to randomly divide the subjects into two groups—one group learning the high-association syllables first, and the other learning the low-association syllables first. This would "balance out" the effects of learning to learn or fatigue. However, if learning high-association syllables first helps subjects to learn low-association syllables later and the reverse is not true, this can confound the interpretation of the results.

## CONTROLLING SITUATIONAL DIFFERENCES

Extraneous variables may operate in the experimental setting to create **situational differences** that can also threaten internal validity. Three methods are commonly used to control potentially contaminating situational variables: (1) hold them constant, (2) randomize them, or (3) manipulate them systematically and separately from the main independent variable.

Holding situational variables constant means that all conditions in the groups are exactly alike except for the exposure to the independent variable. For instance, in a reading experiment you would have the same number of subjects in each group because size of group is known to be a factor affecting reading achievement. Another variable to control is the teacher variable because teacher efficiency and enthusiasm may affect the outcome of any learning experiment. Thus, the same teacher should be used for the two teaching methods that are to be compared. In an experiment, the various assistants must follow the same procedures: use the same instructions, apparatus, and tests, and try to assume the same attitudes with all groups. All groups should meet at approximately the same

time of day and in the same type of room. You would not want the experimental group to meet during the first period in the morning of a school day and the control group during the last period of the day. Environmental conditions—such as temperature, light intensity, humidity, furnishings in the room, and the presence or absence of distracting noises—should be the same for all groups. If you cannot hold conditions constant, as the experimenter you must attempt to randomize or balance out certain situational variables. For instance, if it is not possible to have the same teacher for both treatments, you might randomly assign half of the available teachers to use method A and half to use method B. You could do the same with other experimental conditions, such as apparatus.

It is important to control subjects' attitudes as well. For example, consider a study designed to test the effect of a film on attitudes of middle school students. One randomly selected group of children from a classroom is selected to view the film, leaving the control group in the classroom. Unknowingly, the experimenter may have set in motion a large number of forces. The children in the control group may be resentful or feel rejected or inferior to the others, and these factors could affect the outcome of the study. The researcher must take steps to ensure that the control group also views a film at the same time. Both groups, or neither group, should know that they are taking part in an experiment.

It is very important to control subjects' attitudes in experiments on drugs. For example, in studying the effect of a drug on the performance of a manipulative skill, all subjects must think they are taking the drug. This is managed by using a placebo, which is a chemically inert or neutral substance given to subjects to make them believe they are receiving the actual experimental drug. The experimenter gives every subject a substance that appears to be the same. For some subjects this is the drug under investigation, and for the remainder it is the placebo. Thus, the expectation of a drug effect is held constant between experimental and control subjects and is therefore eliminated as a confounding variable. Otherwise, just the knowledge that they had been given a drug might lead the experimental subjects to be either extra cautious or quite reckless—and the experimenter would not know whether it was the effect of the drug or the subjects' behavior or both that produced the result.

The use of a placebo as just described illustrates what is called a **single-blind experiment.** The subjects are unaware of the treatment condition they are in, although the researcher knows. Sometimes, however, it is necessary to hold the attitudes of the researcher constant for different independent variable levels. This is done by using a **double-blind experimental** procedure in which neither the experimenter nor the subjects know which kind of treatment the subjects are getting. Double-blind experiments are more difficult to carry out in educational research because the nature of the treatment may make it obvious which is the experimental group. Another consideration is that in a double-blind situation the experimenter must depend on other people to set up the groups, administer the treatment, and record results.

Another way to control extraneous situational variables is by manipulating them systematically. Many educational experiments must use a sequence of experimental and control conditions to control progressive effects, such as practice and fatigue effects. Experimenters do this by controlling the order in which experimental conditions are presented through counterbalancing; half the

subjects may receive an AB order and the other half a BA order. In this case, an extraneous variable is being systematically manipulated. This procedure not only controls the potentially contaminating effect of order but also can estimate the size of the order effect by determining whether the average A and B values obtained in the two sequences are different.

If a researcher believes that class size may influence the effectiveness of a new instructional method, he or she can control that variable by building it into the design as a second independent variable. The design would provide for two levels of the method variable and two levels of class size, as shown here:

| | | **Class Size** | |
|---|---|---|---|
| | | Small | Large |
| **Method** | A | | |
| | B | | |

The researcher could then determine the effect of the instructional method, the effect of class size, and any interaction effect between method and class size. Adding more independent variables to the experiment increases the complexity of the study, but it has the advantage of furnishing additional information about the effect of relevant variables on the dependent variable and their interaction with the main independent variables.

## STATISTICAL CONCLUSION VALIDITY

**Statistical conclusion validity** refers to the appropriate use of statistics to infer whether an observed relationship between the independent and dependent variables in a study is a true cause–effect relationship or whether it is just due to chance. Any inappropriate use of statistics is thus a threat because it may result in an erroneous conclusion about the effect of the independent variable on the dependent variable. Threats to statistical conclusion validity include using tests with low power, which may fail to detect a relationship between variables; violating the assumptions of statistical tests, which can lead to over- or underestimating the size and significance of an effect; using measures with low reliability or with restricted range of scores, both of which lessen the probability of detecting a relationship; and using statistical tests that result in over- or underestimations of the size of an effect (Shadish, Cook, & Campbell, 2002, Table 2.2). Such problems with statistical analysis may lead researchers to report incorrectly that a treatment has no effect. "In that case, a difference that is real but small may be lost in the statistical noise" (McBurney & White, 2004, p. 179).

## CONSTRUCT VALIDITY OF EXPERIMENTS

Experimental research deals with psychological constructs such as intelligence, motivation, learning, personality, self-concept, creativity, attitudes, and anxiety. Recall from Chapter 8 that such abstract constructs cannot be measured directly.

We choose indicators (operational definitions) that we assume represent the construct. In Chapter 9, we discussed construct validity in connection with tests that are used to measure such psychological constructs. We defined construct validity of a test as the extent to which a test is measuring the psychological construct it is intended to measure. Shadish et al. (2002) applied the concept of construct validity to experimental research and referred to it as the **construct validity of experiments.** Specifically, construct validity of experiments is defined as the validity of the inferences made about a construct based on the measures, treatment, subjects, and settings used in an experimental study.

As an example of the construct validity of an experiment, assume a researcher wants to investigate the effect of anxiety on learning a verbal task. The researcher manipulates anxiety by telling the experimental group "the results will influence your grade" or some similar verbal instructions. The control group receives some benign instructions. Construct validity is concerned with how well this manipulation represents the construct (anxiety) that the researcher wants. The researcher may not be manipulating anxiety at all but, rather, something else, such as need to achieve. The question is, To what extent are the inferences made from this experiment about the effect of anxiety on verbal learning justified?

To continue with the previous example, the researcher could measure anxiety by asking teachers based on their observations to classify the children in their classrooms as "high anxiety" or "low anxiety." Would the teachers' ratings be an acceptable measure of children's anxiety level? Would the researcher be justified in his or her conclusions about the effect of anxiety on learning? One would probably have more faith in the results of a study in which a standardized and well-established measure of anxiety were used to classify high- and low-anxiety children. Such a study should have more construct validity.

Construct validity is also concerned with the subjects chosen for an experiment. Imagine a research study using disadvantaged children. How does the researcher define the term "disadvantaged"? Studies will define it differently so that one might find very different kinds of children with that label in different studies. To take another example, assume the subjects of a study are people suffering from depression. People labeled "depressed" might differ greatly depending on whether the diagnosis was made by a competent psychologist or by a counselor using only scores from a personality inventory.

There is also concern about the construct representation of the settings for experimental research. It receives somewhat less attention, however, than the construct representations of treatments and subjects, except for research dealing with the effect of environment and culture. Assume a study wishes to investigate the effect of preschool on first-grade achievement and adjustment. Preschools differ greatly—from Head Start to those in public schools, Montessori, and other private and very expensive preschools. It would be important to know the characteristics of the setting in order to make valid inferences from the study about the effect of preschool.

## THREATS TO CONSTRUCT VALIDITY

The threats to construct validity concern how well the study's operations match the constructs used to describe those operations.

The following are potential threats:

1. *Measure of the construct*. The measures used were not appropriate (poor operational definition), so the construct was not accurately measured.
2. *Manipulation of the construct*. The construct was not properly manipulated in the study; faulty manipulation may lead to incorrect inferences.
3. *Reactivity to the experimental situation*. Subjects' perceptions of the experimental situation become part of the treatment construct actually being tested. Recall the Hawthorne effect from the discussion of internal validity.
4. *Experimenter effect*. The experimenter can convey expectations about desirable responses, and those expectations become part of the treatment construct being studied.

### PROMOTING CONSTRUCT VALIDITY

Shadish et al. (2002) suggested the following ways to improve construct validity of experiments: (1) Start with a clear explanation of the persons, setting, treatment, and outcome constructs of interest; (2) carefully select instances that match those constructs; (3) assess the match between instances and constructs to determine if any slippage between the two occurred; and (4) revise construct descriptions accordingly. The reader is referred to Shadish et al. for a thorough discussion of construct validity.

## EXTERNAL VALIDITY OF EXPERIMENTAL DESIGNS

**External validity** refers to the extent to which the findings of a study can be generalized to other subjects, settings, and treatments. Any single study is necessarily performed on a particular group of subjects, with selected measuring instruments and under conditions that are in some respects unique. Yet researchers want the results of a study to furnish information about a larger realm of subjects, conditions, and operations than were actually investigated. Would a teaching method found to be highly effective in Indianapolis be equally effective in Minneapolis? To make generalizations from the observed to the unobserved, researchers need to assess how well the sample of events actually studied represents the larger population to which results are to be generalized. To the extent that the inferences about a causal relationship hold over changes in subjects, settings, and treatments, the experiment has external validity.

### THREATS TO EXTERNAL VALIDITY

1. *Selection–treatment interaction* (nonrepresentativeness). A major threat to external validity of experiments is the possibility of interaction between subject characteristics and treatment so that the results found for certain kinds of subjects may not hold for different subjects. This interaction occurs when the subjects in a study are not representative of the larger population to which one may want to generalize. The results of a reading study that used first-graders enrolled in an affluent suburban school district as

subjects might not be the same if first-graders in a rural school district had been the subjects. When two experimentally accessible populations are not representative of the same target population, seemingly similar studies can lead to entirely different results. Counseling method A may produce better results than method B in inner-city schools, whereas method B is superior in affluent suburban schools. The best method for teaching quantitative reasoning among eighth-graders may be the worst method among second-graders. As the old saying goes, "One man's meat is another man's poison." Again, a thorough description of the accessible population will help other educators judge whether a particular treatment is likely to be "meat or poison" for their populations of interest.

Using volunteers also presents an external validity problem because volunteers are known to have special characteristics that may not be typical of the population to which generalizations are to be made. No one knows how nonvolunteers would be affected by the experimental treatment. Critics have also questioned the use of college students in so much of the experimental research in psychology and the social sciences. How representative are college students? Whether we can generalize the findings from research with college students to larger groups partially depends on the type of research. We might generalize the findings from research on reaction time or the learning of nonsense syllables but not the findings from research on attitudes.

2. *Setting–treatment interaction* (artificiality). Artificiality in the setting may limit the generalizability of the results. The findings of a contrived lab study of motivation may not be the same as one would obtain in a study conducted in a public school setting.

3. *Pretest–treatment interaction*. Using a pretest may increase or decrease the experimental subjects' sensitivity or responsiveness to the experimental variable and thus make the results obtained for this pretested population unrepresentative of effects of the experimental variable on the unpretested population from which the experimental subjects are selected. In this case, you could generalize to pretested groups but not to unpretested ones. Assume that you give a group of seventh-graders a questionnaire concerning their dietary habits and randomly divide the group into experimental and control groups. You expose the experimental group to a series of film presentations concerning good eating habits, whereas the control group views a series of health films unrelated to eating habits (placebo). The dependent variable is derived by observing the children's food selections in an actual free-choice situation. If the experimental group shows a significantly greater preference for healthful foods, you would like to conclude that the films are effective. Before reaching a conclusion, you must consider the possibility that the pretest caused the students to think about their eating habits and "set them up" to respond to the films. The same effect might not have been observed in an unpretested group.

4. *Subject effects*. Attitudes and feelings of the participants that develop during a study may influence the generalizability of the findings to other settings. This threat is also called the **reactive effect** because subjects are reacting to the experience of participating in an experiment. For example, the Hawthorne

effect described previously as an internal validity problem can also be an external validity problem. Subjects' knowledge that they have been selected for an experiment and are being treated in a special way may affect the way they respond to the treatment. Thus, the treatment could appear to be more effective than it might be in the long term. This effect weakens generalization to situations in which people do not regard themselves as special. Closely related is a novelty effect that may happen in research that compares groups using innovative new methods to untreated control groups. A new instructional method may appear to be successful because it leads to excitement and enthusiasm among subjects that may affect the application of results to other groups. Likewise, the John Henry effect may occur when subjects in the untreated control group are determined to do as well as or better than the subjects in the experimental group. The teachers in the control group may feel threatened and may exert extra effort so that they and their students will not look bad compared to the experimental group.

5. *Experimenter effects.* Another threat to external validity is the experimenter effect, which occurs when the experimenter consciously or unconsciously provides cues to subjects that influence their performance. The results of the study could be specific to an experimenter with a certain personality or other characteristics. Sometimes the presence of observers during an experiment may so alter the normal responses of the participating subjects that the findings from one group may not be valid for another group or for the broader population, and it would be hazardous to generalize the findings.

Table 10.3 summarizes the threats to external validity.

## DEALING WITH THREATS TO EXTERNAL VALIDITY

Controlling the threats to external validity is not as straightforward as with internal validity. With the latter, the research design is the significant factor. Before you can assume external validity, you need to examine carefully and logically the similarities and differences between the experimental setting and the target setting with respect to subjects and treatments. A review of the literature would reveal if other research on the same question had used different kinds of subjects, settings, or methodology.

**Table 10.3** Threats to External Validity

| | |
|---|---|
| Selection–treatment interaction | An effect found with certain kinds of subjects might not apply if other kinds of subjects were used. Researcher should use a large, random sample of participants. |
| Setting–treatment interaction | An effect found in one kind of setting may not hold if other kinds of settings were used. |
| Pretest–treatment interaction | Pretest may sensitize subjects to treatment to produce an effect not generalizable to an unpretested population. |
| Subject effects | Subjects' attitudes developed during study may affect the generalizability of the results. Examples are the Hawthorne and the John Henry effects. |
| Experimenter effects | Characteristics unique to a specific experimenter may limit generalizability to situations with a different experimenter. |

The following suggestions can help control threats to external validity:

1. Randomly sample the target population to select subjects for the study and then randomly assign them to treatment groups. If this is not possible because of the population size, then select subjects randomly from the experimentally accessible population and show the similarity of the experimentally accessible population and the target population.
2. Identify the relevant characteristics of subjects in the target population, and determine the impact of these characteristics by incorporating them into the research study. For example, if you want to generalize to ethnically diverse urban high schools, you could include different ethnic groups in the study and examine the performance of each group separately to determine if the experimental treatment worked equally well with all groups or if there were differences. This kind of information would help determine the groups of students to whom the results could be generalized. The same could be done with gender, age, educational levels, and other characteristics. Factorial designs enable researchers to assess the effectiveness of the treatment at different levels of other variables such as race and gender.
3. You can control problems arising from a pretest–treatment interaction by choosing a design that does not use a pretest.
4. You can control the reactive effects by arranging for a second control group to experience an interesting interaction with the researcher just like the experimental group. The interaction for the second control group, however, is totally unrelated to the dependent variable of the study. If there is a reactive effect due to the subjects' feeling special that they are in a research study (Hawthorne effect), it will show up in the second control group just as it would in the experimental group. If the experimental group performs better on the dependent variable, then you have confidence that the effect is due to the treatment and not the Hawthorne effect. This is similar to the use of placebos in medical research. In research on a new treatment or new drug, the experimental group receives the real treatment or drug and the control group receives a placebo. Because neither group knows whether it is the experimental or the control group, any expectations or psychological effects of a new drug are controlled. The experimental and control groups are later compared to determine the effectiveness of the real medication.
5. Replicate the research study in a new setting. This is a good way to determine if similar results will be found. If you find the same results with other populations and in other settings, you can have reasonable confidence that generalizations are valid.

## RELATIONSHIPS AMONG THE TYPES OF VALIDITY

We have discussed four types of validity of concern in experimental research. You may be wondering how one ever designs a study that has appropriate validity. In practice, researchers prioritize and make trade-offs among the validity types. The most important and most discussed trade-off is that between internal validity

and external validity. Internal validity has long been considered the sine qua non ("without which, nothing") of experimentation. As an experiment becomes more rigorously controlled (internally valid), its artificiality tends to increase, and it becomes less generalizable and less externally valid. An experiment conducted in a more natural environment such as a classroom may have greater external validity, but its internal validity may be less. Fortunately, internal and external validity are not incompatible. Most researchers would try first to establish internal validity because unless one can show that a treatment has an unambiguous effect in one setting, it is almost pointless to ask if the treatment would cause the same effect in other settings. Researchers attempt to reach a compromise, which amounts to choosing a design that provides sufficient control to make the results interpretable, while preserving some realism so that the findings will generalize.

External validity is also related to construct validity because valid knowledge about the constructs that are involved in a study is important to generalizing the results. You should notice that some of the same threats to internal validity are also threats to construct validity (subject effects and experimenter effects).

Lastly, internal validity and statistical conclusion validity are related because both are concerned with the relationship between treatment and outcome. A study may be very carefully designed and well controlled (internal validity), but a statistical error can occur and lead to incorrect conclusions about statistical significance and effect sizes. "Thus, in quantitative experiments, internal validity depends substantially on statistical conclusion validity" (Shadish et al., 2002, p. 63).

We have pointed out a list of threats that may lead researchers to make incorrect inferences from their studies. We hope that beginning researchers will not find this discussion too discouraging. Cooper (2006) wrote,

> Instead, researchers should use the list of pitfalls to ensure that they take their task seriously. No research design, in theory or practice, leads to flawless inferences be they about descriptions, relations, or causal connections, quantitative or qualitative, about individual or group-level phenomena. Ultimately, the veracity of scientific claims rests not on the results of a single study but on an accumulation of evidence, conducted by multiple researchers. (p. 876)

## SUMMARY

Experimentation is the most rigorous and the most desirable form of scientific inquiry. The controlled conditions that characterize the experiment make it possible to identify verified functional relationships among the phenomena of interest to educators. Experimenters who control the conditions under which an event occurs have distinct advantages over observers who simply watch or study an event without control: (1) They can manipulate or vary the conditions systematically and note the variations in results; (2) they can make the event occur at a time when they are prepared to make accurate observations and measurements; and (3) they can repeat their observations under the same conditions, for verification, and can describe these conditions so that other experimenters can duplicate them and make an independent check on the results.

We discussed four types of validity important in experimental research. These four types of validity correspond to four questions that people ask when interpreting experimental results:

1. Is there a causal relationship between the variables, or would the outcome have been obtained without the treatment? (internal validity)

2. How large and reliable is the covariation between the presumed cause and effect? (statistical conclusion validity)
3. What constructs are involved in the persons, settings, treatments, and measures used in the experiment? (construct validity)
4. How generalizable is the observed causal relationship to other persons, treatments, and settings? (external validity)

We also discussed the factors that threaten the different types of validity of experimental research and how to minimize their influence.

## KEY CONCEPTS

analysis of covariance (ANCOVA)
comparison group
compensatory demoralization
compensatory rivalry
confounding
construct validity of experiments
control group
control of variables
covariate
diffusion
double-blind experiment
experiment
experimental design
experimental group
experimental mortality (attrition)
experimenter effect
external validity
extraneous variable
Hawthorne effect
history effect
homogeneous selection
implementation threat
instrumentation
internal validity
John Henry effect
law of the single independent variable
law of the single significant variable
manipulation of independent variable
maturation
novelty effect
population external validity
pretest sensitization
random assignment (randomization)
randomization
randomized matching
reactive effect
selection
selection–maturation interaction
single-blind experiment
situational differences
statistical conclusion validity
statistical equivalence
statistical regression (as internal-validity problem)
subject effects
testing effect
using subjects as their own controls

## EXERCISES

1. Which type of validity is being referred to in the following questions?
   - **a.** Is the statistical test appropriate to ensure a correct conclusion about a causal relationship between the independent variable and the dependent variable?
   - **b.** Did variable A cause variable B or is the observed relationship due to some uncontrolled factor(s)?
   - **c.** If it is a causal relationship, what constructs are involved in the cause–effect relationship?
   - **d.** Can we generalize this causal relationship to other persons, settings, and times?
2. What is the difference between random selection and random assignment? How are they related to internal and external validity?
3. Evaluate the following research designs with respect to methods used and the control provided. Make suggestions for improvements if needed.
   - **a.** A researcher wanted to ascertain if homogeneous grouping improves learning in a first course in biology. The researcher designated one of two high schools in a small city to serve as the experimental school and the other as the control. Both schools had approximately the same number of students in each of four sections of science. In the experimental school, pupils were grouped homogeneously on the basis of aptitude and scores on achievement tests in science. In the control school, pupils were placed in sections at random. At the end of the year, all pupils were given a standardized test in biology. Statistical tests showed the experimental group to be superior on the test. The researcher

concluded that homogeneous grouping results in greater learning in biology.

**b.** A history teacher was concerned about her students' lack of knowledge of their state and national governments and of current events. She decided to experiment with some new materials and methods to determine if she could obtain improvement. In classes A and B, she introduced the new materials and methods. In classes C and D, she used the traditional methods. Classes A and B were administered both the pretest and the posttest; classes C and D were administered only the posttest. When comparisons were made on the posttest, classes A and B were found to be superior. Their superior performance was attributed to the new materials and methods.

**4.** Identify the internal validity threat defined in each of the following:

**a.** The experimental group performs better because its members they know they are participating in an experiment.

**b.** Changes occur within subjects over time that may affect the dependent variable.

**c.** Extraneous events occur during the research that may influence the dependent variable.

**d.** Subjects differ on the dependent variable even before the experimental treatment.

**e.** Control group perceives itself in competition with the experimental group and tries harder.

**f.** Learning about the experiment from people in the experimental group may affect the control group's performance on the dependent variable measure.

**5.** Identify the threats to internal validity in each of the following experimental studies:

**a.** At the beginning of the school year, an elementary teacher set up reading stations in her room with supplementary reading materials in each. At the end of the year, she reported that the interest of her class in reading was higher than in any of her previous classes. She recommended that all teachers set up reading stations.

**b.** A researcher wanted to investigate the effectiveness of new teaching materials designed to increase students' verbal skills. One class used the new materials, and another class used conventional workbooks. Both classes were given a test of verbal skills at the end of the semester.

**c.** A study was conducted in three third-grade classrooms investigating the effect of three methods of teaching math. The teacher could choose the method she wanted to use but could use only one method with her students. At the end of the semester, all the third-graders were given the same math achievement test.

**6.** Identify the threat to external validity in the following examples:

**a.** A teacher read research reporting great results from using a particular remedial reading program. She tried the program in her classroom but did not get the same results.

**b.** A teacher read about a new method of instruction that was very successful with regular classes. The method did not work when she tried it with her special education students.

**c.** A research journal reported that a new antismoking film had brought about a significant change in students' attitudes. After viewing the film, students expressed many more negative attitudes toward smoking than they had on a prefilm measure. A teacher decided to use the film with his middle school students but did not want to take time to administer the "pretest." He administered an attitude scale following the film but did not find marked negative attitudes toward smoking.

**7.** What are the threats to internal validity in the following examples?

**a.** Very obese individuals who go to a weight control clinic will likely show some loss of weight even if the weight control treatment has no effect.

**b.** A longitudinal study of the effectiveness of a study skills treatment followed a group of college students. The results showed that the seniors had a final mean grade point average higher than they had as freshmen.

**8.** Distinguish the instrumentation threat to internal validity from the testing threat.

**9.** Interpret the following statement: For a research study to have external validity, it must first have internal validity.

## ANSWERS

1. **a.** Statistical conclusion validity
   **b.** Internal validity
   **c.** Construct validity
   **d.** External validity
2. Random selection is using a chance procedure to draw a sample from a population. Because it addresses the question of how well results drawn from a sample can be generalized to the population from which the sample was drawn, it is a strategy for increasing external validity. Random assignment is using a chance procedure to assign the subjects available for an experiment to treatment. It is a strategy for increasing internal validity.
3. **a.** Because the researcher could not assign students randomly to the high schools, there are several threats to internal validity. Students in the experimental school may have been brighter or have had more background in science than students in the control school. Differences in the quality of teaching biology in the schools have not been controlled. Because the researcher used only schools in a small city, the results of the study could not be generalized to other high schools in different settings. To improve the study, the researcher could compare initial science achievement and aptitude scores for the schools to determine if the groups were equivalent before treatment. Using several high schools, with classes within each high school being randomly assigned to experimental conditions, would control for factors specific to a given school.
   **b.** Classes not randomly assigned may not be equivalent. Pretesting could have been used to determine equivalence but was only given to the experimental groups. The pretesting of the experimental groups alone may have sensitized the groups and influenced the differences found.
4. **a.** Hawthorne effect (or subject effect)
   **b.** Maturation
   **c.** History
   **d.** Selection
   **e.** John Henry effect
   **f.** Diffusion
5. **a.** Selection, history, and subjects' attitudes
   **b.** Selection, history, and subjects' attitudes
   **c.** Without random assignment, groups may not have been equivalent at the beginning. Teachers chose the method they wanted to use; their enthusiasm, rather than the method, may have been the important factor. If one of the methods was very novel, diffusion may have occurred.
6. **a.** The Hawthorne effect could be a threat in this situation. The students in the original study may have felt special because they were taking part in an experiment and thus put forth extra effort. The findings did not generalize to a nonexperimental setting.
   **b.** The threat in this case is an interaction between selection and treatment. Certain characteristics of the students interacted with the treatment to produce an effect that did not generalize to a group with different characteristics.
   **c.** The threat is the interaction effect of pretesting. The prefilm measure may have sensitized students so that they responded to the film differently than would a group that did not have the "pretest." It was the interaction that brought about the observed change in attitudes. Thus, the results would probably generalize to other pretested populations but not to unpretested ones. The teacher should have used the pretest.
7. **a.** Regression
   **b.** Attrition; the low performers initially may have dropped out of school.
8. Instrumentation involves a change in the measuring instrument; testing involves a change in the subject because of exposure to a prior test (pretest).
9. If we are not sure that the independent variable caused the change in the dependent variable (internal validity), there would be no point in generalizing the findings to other populations or settings (external validity).

## REFERENCES


Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Cook, T. D., & Campbell, D. T. (1979). Q*uasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Cooper, H. (2006). Research questions and research designs. In P. Alexander & P. Winne (Eds.), *Handbook of educational psychology* (2nd ed.). Mahwah, NJ: Erlbaum.

Fisher, R. A. (1925). *Statistical methods for research workers*. London: Oliver & Boyd.

McBurney, D. H., & White, T. L. (2004). *Research methods*. Belmont, CA: Thomson.

Mill, J. S. (1986). *A system of logic*. Charlottesville, VA: Lincoln-Rembrandt. [Original work published 1846]

Reichardt, C. S. (2000). A typology of strategies for ruling out threats to validity. In L. Bickman (Ed.), *Research design: Donald Campbell's legacy* (Vol. 2, pp. 89–115). Thousand Oaks, CA: Sage.

Rice, J. M. (1897). The futility of the spelling grind. *Forum, 23*, 163–172, 409–419.

Rosenthal, R. (1966). *Experimenter effects in behavioral research*. New York: Appleton-Century-Crofts.

Rosenthal, R. (1973). The mediation of Pygmalion effects: A four-factor theory. *Papua New Guinea Journal of Education, 9*, 1–12.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis*. New York: McGraw-Hill.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.

Thorndike, E. L. (1924). Mental discipline in high school subjects. *Journal of Educational Psychology, 15*, 1–22, 83–98.

CHAPTER 11

# Experimental Research Designs

Statistical designs always involve compromises between the desirable and the possible.

Leslie Kish

## INSTRUCTIONAL OBJECTIVES

After studying this chapter, the student will be able to:

1. Explain the function of a research design.
2. Explain the relationship between research design and internal validity.
3. Define preexperimental design and discuss its limitations.
4. Describe the most common randomized experimental designs and state the advantages of each.
5. Define factorial design and state its advantages.
6. Distinguish between-groups and within-groups experimental designs.
7. Define quasi-experimental design and discuss its limitations.
8. Describe single-subject experimental research and its uses.

An **experimental design** is the general plan for carrying out a study with an active independent variable. The design is important because it determines the study's internal validity, which is the ability to reach valid conclusions about the effect of the experimental treatment on the dependent variable. Designs differ in their efficiency and their demands in terms of time and resources, but the major difference is in how effectively they rule out threats to internal validity. Obviously, one first chooses the design that is appropriate for testing the hypothesis of the study. From the appropriate designs, one must choose the one that will (1) ensure that the subjects assigned to the treatment and control groups do not differ systematically on any variables except those under consideration and (2) ensure that the outcome is a consequence of the manipulation of the independent variable and not of extraneous variables.

## CLASSIFYING EXPERIMENTAL DESIGNS

Experimental designs may be classified according to the number of independent variables: single-variable designs and factorial designs. A *single-variable design* has one manipulated independent variable; *factorial designs* have two or more independent variables, at least one of which is manipulated. Experimental designs may also be classified according to how well they provide control of the threats to internal validity: preexperimental, true experimental, and quasi-experimental designs. *Preexperimental designs* do not have random assignment of subjects to groups or other strategies to control extraneous variables. *True experimental designs* (also called randomized designs) use randomization and provide maximum control of extraneous variables. *Quasi-experimental designs* lack randomization but employ other strategies to provide some control over extraneous variables. They are used, for instance, when intact classrooms are used as the experimental and control groups. Thus, true experimental designs have the greatest internal validity, quasi-experimental designs have somewhat less internal validity, and the preexperimental designs have the least internal validity.

Before we discuss the different experimental designs, we introduce the terms and symbols that we will use:

1. $X$ represents the independent variable, which is manipulated by the experimenter; we also refer to it as the *experimental variable* or the *treatment*.
2. $Y$ represents the measure of the dependent variable. $Y_1$ represents the dependent variable *before* the manipulation of the independent variable $X$; it is usually a pretest of some type administered before the experimental treatment. $Y_2$ represents the dependent variable *after* the manipulation of the independent variable $X$; it is usually a posttest administered to subjects after the experimental treatment.
3. $S$ represents the subject or participant used in the experiment; the plural is $S$s.
4. E group is the experimental group—the group that is given the independent variable treatment.
5. C group is the control group—the group that does not receive the experimental treatment. It receives a different treatment or no treatment at all.
6. $R$ indicates random assignment of subjects to the experimental groups and the random assignment of treatments to the groups.
7. $M_r$ indicates that the subjects are matched and then members of each pair are assigned to the comparison groups at random.

In the paradigms for the various designs, the $X$s and $Y$s across a given row are applied to the same people.

The left-to-right dimension indicates the temporal order, and the $X$s and $Y$s vertical to one another are given simultaneously. A dash (—) indicates that the control group does *not* receive the $X$ treatment or receives an alternative treatment.

**Think About It 11.1**

An investigator conducted a study to assess the effectiveness of interactive videotape instruction on learning of high school economics. Forty-five students were chosen at random and assigned to one of two groups. Group E received interactive videotape instruction; group C received conventional instruction. It was found that the groups did not differ significantly in knowledge of economics on a pretest. An achievement test was given immediately after completion of the study; four weeks later a retention test was given.

| Achievement | Group | Mean | *t* | *p* |
|---|---|---|---|---|
| test | C | 97.55 | | |
| | E | 125.73 | 4.02 | .01 |
| Retention test | C | 90.52 | | |
| | E | 112.91 | 2.19 | .05 |

Match the terms on the right with the description given on the left.

1. Pretest knowledge of economics
2. $p = .01$
3. Random assignment to classes
4. Retention of material over a 4-week period showed a significant difference in favor of group E
5. Use of interactive videotape versus conventional instruction

a. Independent variable
b. Control
c. Finding
d. Level of significance

**Answers**

1. b; 2. d; 3. b; 4. c; 5. a

## PREEXPERIMENTAL DESIGNS

Following are two designs that are classified as **preexperimental** because they provide little or no control of extraneous variables. We do not recommend these designs; however, we realize they are still sometimes used in educational research. We include these weak designs in our discussion simply because they illustrate quite well the way that extraneous variables may operate to jeopardize the internal validity of a design. If you become aware of these sources of weakness in a design, you should be able to avoid them.

### Design 1: One-Group Pretest–Posttest Design

The **one-group pretest–posttest design** usually involves three steps: (1) administering a pretest measuring the dependent variable; (2) applying the experimental treatment $X$ to the subjects; and (3) administering a posttest, again measuring the dependent variable. Differences attributed to application of the experimental treatment are then evaluated by comparing the pretest and posttest scores.

Design 1: One-Group Pretest–Posttest Design

| Pretest | Independent | Posttest |
|---|---|---|
| $Y_1$ | $X$ | $Y_2$ |

To illustrate the use of this design, assume that an elementary teacher wants to evaluate the effectiveness of a new technique for teaching fourth-grade math. At the beginning of the school year, the students are given a standardized test (pretest) that appears to be a good measure of the achievement of the objectives of fourth-grade math. The teacher then introduces the new teaching technique and at the end of the semester administers the same standardized test (posttest), comparing students' scores from the pretest and posttest in order to determine if exposure to the new teaching technique made any difference. The limitation of this design is that because no control group is used, the experimenter cannot assume that any improvement in scores is due to the new technique.

Two obvious extraneous variables not controlled in this design are history and maturation. Things happen between pretest and posttest, other than the experimental treatment, that could affect learning. In the math example, widespread media interest in math education, increased emphasis on math in the school, or the introduction of a particularly effective teacher could increase student achievement in this area. Or an epidemic causing increased absences could depress achievement. Between pretest and posttest, children are growing mentally and physically, and they may have learning experiences that could affect their achievement. History and maturation become more threatening to internal validity as the time between pre- and posttest increases. Instrumentation and regression also present uncontrolled threats to internal validity of this design.

Another weakness is that Design 1 affords no way to assess the effect of the pretest. We know there is a practice effect when subjects take a test a second time or even take an alternate form of the test—or they may learn something just from taking the test and will do better the second time. To deal with this problem, some researchers have used Design 1 without the pretest. However, eliminating the pretest would only make a poor design worse.

The best advice is to avoid using Design 1. Without a control group to make a comparison possible, the results obtained in a one-group design are basically uninterpretable.

### Design 2: Static Group Comparison

The **static group comparison** uses two or more preexisting or intact (static) groups, only one of which is exposed to the experimental treatment. Although this design uses two groups for comparison, it is flawed because the subjects are not randomly assigned to the groups and no pretest is used. The researcher makes the assumption that the groups are equivalent in all relevant aspects before the study begins and that they differ only in their exposure to $X$. To attempt to assess the effects of the $X$ treatment, the researcher compares the groups on the dependent variable measure.

Design 2: Static Group Comparison

| Group | Independent Variable | Posttest |
|---|---|---|
| E | $X$ | $Y_2$ |
| C | — | $Y_2$ |

Although this design has sometimes been used in educational research, it is basically worthless. Because neither randomization nor even matching on a pretest is used, we cannot assume that the groups are equivalent prior to the experimental treatment. Because of the possibility of initial differences between the groups, one could not conclude that the outcome is a result of the experimental treatment. In addition to selection bias, maturation and mortality are threats to the internal validity of this design.

## TRUE EXPERIMENTAL DESIGNS

The designs in this category are called *true experiments* because subjects are randomly assigned to groups. Because of the control they provide, they are the most highly recommended designs for experimentation in education.

### Design 3: Randomized Subjects, Posttest-Only Control Group Design

**Randomized subjects, posttest-only control group design** is one of the simplest yet one of the most powerful of all experimental designs. It has the two essential elements necessary for maximum control of the threats to internal validity: randomization and a control group. No pretest is used; the randomization controls for all possible extraneous variables and ensures that any initial differences between the groups are attributable only to chance and therefore will follow the laws of probability. After the subjects are randomly assigned to groups, only the experimental group is exposed to the treatment. In all other respects, the two groups are treated alike. Members of both groups are then measured on the dependent variable $Y_2$, and the scores are compared to determine the effect of $X$. If the obtained means of the two groups differ significantly (i.e., more than would be expected on the basis of chance alone), the experimenter can be reasonably confident that the experimental treatment is responsible for the observed result.

Design 3: Randomized Subjects, Posttest-Only Control Group Design

| | Group | Independent Variable | Posttest |
|---|---|---|---|
| ($R$) | E | $X$ | $Y_2$ |
| ($R$) | C | — | $Y_2$ |

The main advantage of Design 3 is randomization, which ensures statistical equivalence of the groups before introduction of the independent variable. Recall that as the number of subjects increases, the likelihood that randomization will produce equivalent groups increases. We recommend at least 30 subjects in each group. Design 3 controls for the main effects of history, maturation, regression, and pretesting; because no pretest is used, there can be no interaction effect

of pretest and *X*. Thus, this design is especially recommended for research on changing attitudes. It is also useful in studies in which a pretest is either not available or not appropriate, such as in studies with kindergarten or primary grades, where it is impossible to administer a pretest because the learning is not yet manifest. Another advantage of this design is that it can be extended to include more than two groups if necessary. Possible threats to internal validity are subject effects and experimenter effects.

Design 3 does not permit the investigator to assess change. If such an assessment is desired, then a design that uses both a pretest and a posttest should be chosen (e.g., Design 5).

Because of the lack of a pretest, mortality could be a threat. Without having pretest information, preferably on the same dependent variable used as the posttest, the researcher has no way of knowing if those who dropped out of the study were different from those who continued (Shadish, Cook, & Campbell, 2002).

### Design 4: Randomized Matched Subjects, Posttest-Only Control Group Design

**Randomized matched subjects, posttest-only control group design** is similar to Design 3, except that it uses a matching technique to form equivalent groups. Subjects are matched on one or more variables that can be measured conveniently, such as IQ or reading score. Of course, the matching variables used are those that presumably have a significant correlation with the dependent variable. Although a pretest is not included in Design 4, if pretest scores on the dependent variable are available, they could be used very effectively for the matching procedure. The measures are paired so that opposite members' scores are as close together as possible. The flip of a coin can be used to assign one member of each pair to the treatment group and the other to the control group.

Design 4: Randomized Matched Subjects, Posttest-Only Control Group Design

| | Group | Independent Variable | Posttest |
|---|---|---|---|
| ($M_r$) | E | $X$ | $Y_2$ |
| | C | — | $Y_2$ |

Matching is most useful in studies in which small samples are to be used and where Design 3 is not appropriate. Design 3 depends completely on random assignment to obtain equivalent groups. With small samples the influence of chance alone may result in a situation in which random groups are initially very different from each other. Design 3 provides no assurance that small groups are really comparable before the treatments are applied. The matched-subjects design, however, serves to reduce the extent to which experimental differences can be accounted for by initial differences between the groups; that is, it controls preexisting intersubject differences on variables highly related to the dependent variable that the experiment is designed to affect. The random procedure used to assign the matched pairs to groups adds to the strength of this design.

Design 4 is subject to the difficulties mentioned previously in connection with matching as a means of control. The matching of all potential subjects must

be complete, and the members of each pair must be assigned randomly to the groups. If one or more subjects were excluded because an appropriate match could not be found, this would bias the sample. When using Design 4, it is essential to match every subject, even if only approximately, before random assignment. Design 4 can be used with more than two groups by creating matched sets and randomly assigning one member of each set to each group.

### Design 5: Randomized Subjects, Pretest–Posttest Control Group Design

Design 5 is one of the most widely used true (randomized) experiments. In the **randomized subjects, pretest–posttest control group design,** one randomly assigns subjects to the experimental and control groups and administers a pretest on the dependent variable *Y*. The treatment is introduced only to the experimental subjects (unless two different treatments are being compared), after which the two groups are measured on the dependent variable. The researcher then compares the two groups' scores on the posttest. If there are no differences between the groups on the posttest, the researcher can then look at the average change between pretest and posttest ($Y_2 - Y_1$) scores for each group to determine if the treatment produced a greater change (gain) than the control situation. The significance of the difference in the average pretest–posttest change for the two groups could be determined by a *t* test or *F* test. For reasons beyond the scope of this discussion, measurement experts have pointed out that technical problems arise when comparing gain scores.

Design 5: Randomized Subjects, Pretest–Posttest Control Group Design

| | Group | Pretest | Independent Variable | Posttest |
|---|---|---|---|---|
| (*R*) | E | $Y_1$ | *X* | $Y_2$ |
| (*R*) | C | $Y_1$ | — | $Y_2$ |

The recommended statistical procedure to use with Design 5 is an analysis of covariance (ANCOVA) with posttest scores as the dependent variable and pretest scores as the covariate to control for initial differences on the pretest. ANCOVA is a more powerful test and gives more interpretable results than does the comparison of gain scores for the two groups.

The main strength of this design is the initial randomization, which ensures statistical equivalence between the groups prior to experimentation; also, the fact that the experimenter has control of the pretest can provide an additional check on the equality of the two groups on the pretest, $Y_1$. Design 5 thus controls most of the extraneous variables that pose a threat to internal validity. For example, the effects of history and maturation are experienced in both groups; therefore, any difference between the groups on the posttest measure could probably not be attributed to these factors. Differential selection of subjects and statistical regression are also controlled through the randomization procedure. There is one internal validity issue, however. Although both E and C groups take the pretest and may experience the sensitizing effect, the pretest can cause the experimental subjects to respond to the *X* treatment in a particular way just because of their increased sensitivity.

The result is a difference on the posttest that could mistakenly be attributed to the effect of the treatment alone. The crucial question is, Would the effect of *X* on the experimental subjects be the same without the exposure to the pretest? This problem has been particularly evident in studies of attitude change. When the first attitude scale is administered as the pretest in a study, it can arouse interest or sensitize subjects to the issues or material included in the scale. Then, when the experimental treatment (a lecture, film, or the like) is administered, the subjects may be responding not so much to the *X* treatment as to a combination of their aroused sensitivity to the issues and the experimental treatment.

The main concern in using Design 5 is external validity. Ironically, the problem stems from the use of the pretest, an essential feature of the design. As mentioned previously, there may be an interaction between the pretest and the treatment so that the results are generalizable only to other pretested groups. The responses to the posttest may not be representative of how individuals would respond if they had not been given a pretest.

Let us consider an example. Suppose that one criterion for the success of a new teaching method in high school social studies is the number of students who report that they read newspapers such as the *New York Times*. During the course, no special emphasis is placed on this particular source; but it, along with several other papers of somewhat less repute, is made available to students. If the study uses a pretest–posttest design, the pretest questionnaire might include an item such as "Do you read the *New York Times* for daily news?" This question alone may be enough to sensitize the experimental students to that newspaper, so when it becomes available during the course, they will be more likely to pick it out from the others. As a result, the experimental group may show greater use of the *New York Times* on the posttest than does the control group—not because of the course content only but because of the combined effect of course content and pretest. A new class taught by the same method, but not pretested and hence not sensitized, may show no greater attentiveness to the *New York Times* than the control group.

Despite this shortcoming, Design 5 is widely used because the interaction between pretest and treatment is not a serious problem in most educational research. The pretests used are often achievement tests of some type and therefore do not significantly sensitize subjects who are accustomed to such testing. However, if the testing procedures are somewhat novel or motivating in their effect, then it is recommended that the experimenter choose a design not involving a pretest. Alternatively, whenever you suspect that the effect of the pretest might be reactive, it is possible to add a new group or groups to the study—a group that is not pretested. Solomon (1949) suggested two designs that overcome the weakness of Design 5 by adding an unpretested group or groups. We present these designs next.

### Design 6: Solomon Three-Group Design

The first of the Solomon designs uses three groups, with random assignment of subjects to groups. Note that the first two lines of this design are identical to Design 5. However, the **Solomon three-group design** has the advantage of employing a second control group labeled $C_2$ that is *not* pretested but is exposed to the treatment *X*. This group, despite receiving the experimental treatment, is functioning as a control and is thus labeled control group. This design overcomes the difficulty inherent in Design 5—namely, the interactive effect of pretesting

and the experimental treatment. The posttest scores for the three groups are compared to assess the interaction effect.

Design 6: Solomon Three-Group Design

| | Group | Pretest | Independent | Posttest |
|---|---|---|---|---|
| (*R*) | E | $Y_1$ | $X$ | $Y_2$ |
| (*R*) | $C_1$ | $Y_1$ | — | $Y_2$ |
| (*R*) | $C_2$ | — | $X$ | $Y_2$ |

If the experimental group has a significantly higher mean on the posttest $Y_2$ than does the first control group ($C_1$), the researcher cannot be confident that this difference is caused by $X$. It might have occurred because of the subjects' increased sensitization after the pretest and the interaction of their sensitization and $X$. However, if the posttest mean ($Y_2$) of the second control group ($C_2$) is also significantly higher than that of the first control group, then one can conclude that the experimental treatment, rather than the pretest–$X$ interaction effect, has produced the difference because the second control group is not pretested.

### Design 7: Solomon Four-Group Design

**The Solomon four-group design** provides still more rigorous control by extending Design 6 to include one more control group that receives neither pretest nor treatment.

Design 7: Solomon Four-Group Design

| | Group | Pretest | Independent Variable | Posttest |
|---|---|---|---|---|
| (*R*) | E | $Y_1$ | $X$ | $Y_2$ |
| (*R*) | $C_1$ | $Y_1$ | — | $Y_2$ |
| (*R*) | $C_2$ | — | $X$ | $Y_2$ |
| (*R*) | $C_3$ | — | — | $Y_2$ |

Design 7, with its four groups, has strength because it incorporates the advantages of several other designs. It provides good control of the threats to internal validity. Design 7 has two pretested groups and two without a pretest; one of the pretested groups and one of the nonpretested groups receive the experimental treatment, and then all four groups take the posttest. The first two lines (as in Design 5) control extraneous factors such as history and maturation, and the third line (as in Design 6) controls the pretest–$X$ interaction effect. When the fourth line is added to make Design 7, you have control over any possible contemporary effects that may occur between $Y_1$ and $Y_2$. The last two lines represent Design 3, so actually you have a combination of the randomized subjects pretest–posttest control group design and the randomized subjects posttest-only control group design. In addition to the strengths of each design taken separately, you also have the replication feature provided by the two experiments. This combination takes advantage of the information provided by the pretest–posttest

procedure and at the same time shows how the experimental condition affects an unpretested group of *S*'s.

In Design 7, you can make several comparisons to determine the effect of the experimental *X* treatment. If the posttest mean of the E group is significantly greater than the mean of the first control group, $C_1$, and if the $C_2$ posttest mean is significantly greater than that of $C_3$, you have evidence for the effectiveness of the experimental treatment. You can determine the influence of the experimental conditions on a pretested group by comparing the posttests of E and $C_1$ or the pre–post changes of E and $C_1$. You can find the effect of the experiment on an unpretested group by comparing $C_2$ and $C_3$. If the average differences between posttest scores, $E-C_1$ and $C_2-C_3$, are approximately the same, then the experiment must have had a comparable effect on pretested and unpretested groups.

Design 7 actually involves conducting two experiments, one with pretests and one without pretests. If the results of these two experiments agree, as indicated previously, the investigator can have much greater confidence in the findings.

The main disadvantage of this design is the difficulty involved in carrying it out in a practical situation. More time and effort are required to conduct two experiments simultaneously, and there is the problem of locating the increased number of subjects of the same kind that would be needed for the four groups.

Another difficulty is with the statistical analysis. There are not four complete sets of measures for the four groups. As noted, you can compare E and $C_1$, and $C_2$ and $C_3$, but no single statistical procedure would use the six available measures simultaneously. Campbell and Stanley (1966) suggested working only with posttest scores in a two-way analysis of variance design. The pretest is considered as a second independent variable, along with *X*. The design is as follows:

| | **No *X*** | ***X*** |
|---|---|---|
| Pretested | $Y_2$, control 1 | $Y_2$, experimental |
| Unpretested | $Y_2$, control 3 | $Y_2$, control 2 |

From the column means, you can determine the main effect of *X*; from row means, the main effect of pretesting; and from cell means, the interaction of testing with *X*.

## FACTORIAL DESIGNS

The designs presented thus far have been the classical single-variable designs in which the experimenter manipulates one independent variable *X* to determine its effect on a dependent variable *Y*. However, in complex social phenomena several variables often interact simultaneously, and restricting a study to one independent variable may impose an artificial simplicity on a complex situation. The *X* variable alone may not produce the same effect as it might in interaction with another, so the findings from one-variable designs may be misleading. For instance, we might ask about the effectiveness of a particular method of teaching on students' learning. The answer may well be that the effectiveness depends on a number of variables, such as the age and ability level of the students, the personality of the teacher, the subject matter, and so on. Computer-assisted instruction,

for example, may be more effective with below-average students than with bright ones. A classical one-variable design would not reveal this interactive effect of method and intelligence level. The information yield of an experiment can be markedly increased by using a factorial design. A **factorial design** is one in which the researcher manipulates two or more variables simultaneously in order to study the independent effect of each variable on the dependent variable, as well as the effects caused by interactions among the several variables. Some have said that the real breakthrough in educational research came with Fisher's (1925) development of factorial designs.

The independent variables in factorial designs are referred to as *factors*. Factors might be categorical variables such as gender, ethnicity, social class, and type of school, or they might be continuous variables such as aptitude or achievement. The researcher identifies the levels of each of these factors to be investigated. For example, aptitude might have two levels (high and low) or three levels (high, average, and low). Gender would have two levels (male and female), as would method of instruction (lecture and discussion).

### Design 8: Simple Factorial Design

Factorial designs have been developed at varying levels of complexity. The simplest factorial design is the $2 \times 2$, which is read as "2 by 2." This design has two factors, and each factor has two levels.

Design 8: Simple Factorial Design

| **Variable 2 ($X_2$)** | **Variable 1 ($X_1$)** | |
|---|---|---|
| | **Treatment A** | **Treatment B** |
| Level 1 | Cell 1 | Cell 3 |
| Level 2 | Cell 2 | Cell 4 |

To illustrate, let us assume that an experimenter is interested in comparing the effectiveness of two types of teaching methods—methods A and B—on the achievement of ninth-grade science students, believing there may be a differential effect of these methods based on the students' level of science aptitude. Table 11.1 shows the $2 \times 2$ factorial design. The aptitude factor has two levels—high and low; the other factor (instructional method) also has two levels (A and B). The researcher randomly selects 60 *S*s from the high-aptitude group and assigns 30 Ss to method A and 30 Ss to method B. This process is repeated for the low-aptitude

**Table 11.1** Example of a Factorial Design

| | **Instructional Method ($X_1$)** | | |
|---|---|---|---|
| **Aptitude ($X_2$)** | **Method A** | **Method B** | **Mean** |
| High | 75.0 | 73.0 | 74 |
| Low | 60.0 | 64.0 | 62 |
| Mean | 67.5 | 68.5 | |

group. Teachers are also randomly assigned to the groups. Note that a $2 \times 2$ design requires four groups of subjects; each group represents a combination of a level of one factor and a level of the other factor.

The scores in the four cells represent the mean scores of the four groups on the dependent variable, the science achievement test. In addition to the four cell scores representing the various combinations of treatments and levels, there are four marginal mean scores: two for the columns and two for the rows. The marginal column means are for the two methods, or treatments, and the marginal row means are for the two levels of aptitude.

From the data given, you can first determine the *main effects* for the two independent variables. The main effect for treatments refers to the treatment mean scores without regard to aptitude level. If you compare the mean score of the two method A groups, 67.5, with that of the two method B groups, 68.5, you find that the difference between these means is only 1 point. Therefore, you might be tempted to conclude that the method used has little effect on the achievement scores, the dependent variable.

Now examine the mean scores for the levels to determine the main effect of $X_2$, aptitude level, on achievement scores. The main effect for levels does not take into account any differential effect caused by treatments. The mean score for the two high-aptitude groups is 74, and the mean score for the two low-aptitude groups is 62; this difference, 12 points, is the effect attributable to aptitude level. The high-aptitude group has a markedly higher mean score; thus, regardless of treatment, the high-aptitude groups perform better than the low-aptitude groups. Note that the term *main effects* does not mean the most important effect but, rather, the effect of one independent variable (factor) ignoring the other factor. In the example, main effect for teaching method refers to the difference between method A and method B (column means) for all students regardless of aptitude. The main effect for aptitude is the difference between all high- and low-aptitude students (row means) regardless of teaching method.

A factorial design also permits the investigator to assess the **interaction** between the two independent variables—that is, the different effects of one of them at different levels of the other. If there is an interaction, the effect that the treatment has on learning will differ for the two aptitude levels. If there is no interaction, the effect of the treatment will be the same for both levels of aptitude. Looking at Table 11.1, you can see that the method A mean is higher than the method B mean for the high-aptitude group, and the method B mean is higher for the low-aptitude group. Thus, some particular combinations of treatment and level of aptitude interact to produce greater gains than do some other combinations. This interaction effect between method and aptitude levels is shown graphically in Figure 11.1. If this interaction is statistically significant, you conclude that the effectiveness of the method depends on aptitude. Method A is more effective with the high-aptitude students; method B is more effective with the low-aptitude group.

Now examine another set of data obtained in a hypothetical $2 \times 2$ factorial study. Table 11.2 shows the results of a study designed to investigate the effect of two methods of instruction on achievement. Because the investigator anticipates that the method may be differentially effective depending on the aptitude of the subject, the first step is to use an aptitude measure and place participants into

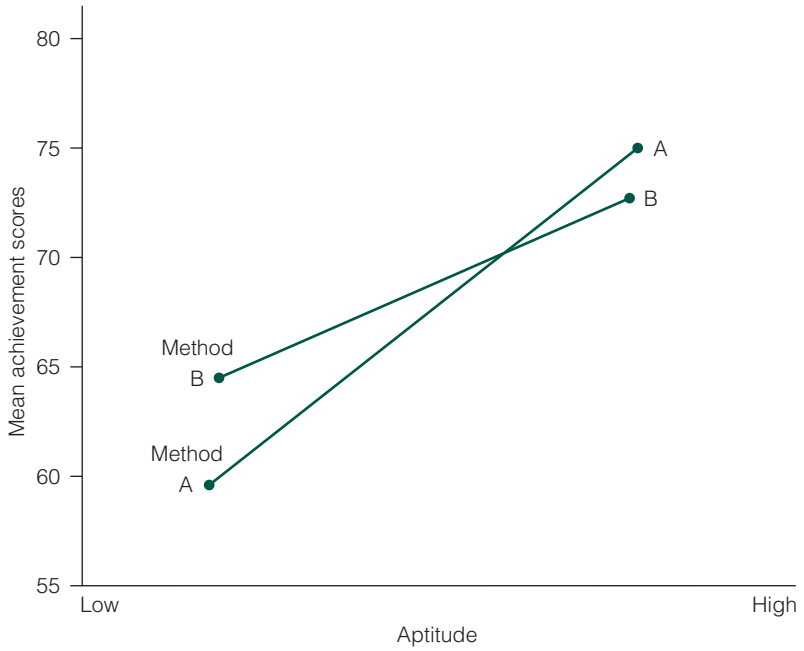



**Figure 11.1** Illustration of Interaction between Method and Aptitude Level

one of two levels: high aptitude or low aptitude. The researcher randomly assigns subjects within each aptitude level to one of the two methods.

After the experiment, the researcher administers achievement tests and records the scores for every subject. If you compare the mean score of the two groups taught by method B, 53, with that of the two groups taught by method A, 45, the former is somewhat higher and method B appears to be more effective than method A. The difference between the means for the two aptitude levels, or the main effects for aptitude, is 10 (54 − 44). Regardless of treatment, the high-aptitude group performs better than the low-aptitude group. The data reveal no interaction between treatment and levels. Method B appears to be more effective regardless of the aptitude level. In other words, treatments and levels are independent of each other. It would not be possible to demonstrate either the presence or the absence of interaction without using a factorial design. The lack of interaction is illustrated graphically in Figure 11.2.

**Table 11.2** Example of a Factorial Design

| | Treatment ($X_1$) | | |
|---|---|---|---|
| **Aptitude ($X_2$)** | **Method A** | **Method B** | **Mean** |
| High | 50 | 58 | 54 |
| Low | 40 | 48 | 44 |
| Mean | 45 | 53 | |

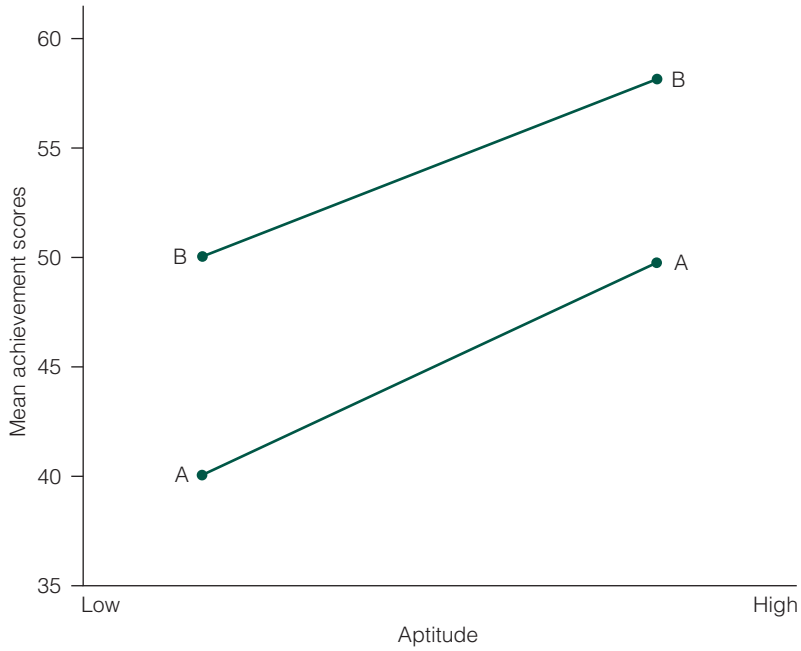



**Figure 11.2** Illustration of Lack of Interaction between Method and Aptitude Level

The factorial design can be extended to more complex experiments in which there are a number of independent variables; the numeric values of the digits indicate the number of levels for the specific independent variables. For instance, in a $2 \times 3 \times 4$ factorial design, there are three independent variables with two, three, and four levels, respectively. Such an experiment might use two teaching methods, three ability levels, and four grades. Theoretically, a factorial design may include any number of independent variables with any number of levels of each. However, when too many factors are manipulated or controlled simultaneously, the study and the statistical analysis become unwieldy and some of the combinations may be artificial. The number of groups required for a factorial design is the product of the digits that indicate the factorial design. In the $2 \times 3 \times 4$ design, 24 groups would be required to represent all combinations of the different levels of the multiple independent variables. The mere thought of the complexities involved in arranging for large numbers of subjects under large numbers of conditions will perhaps help the reader to understand why most educational researchers attempt to answer their questions with the simplest possible designs, even though the statistical analysis can be easily handled by computer. The advantages of the factorial design are that it accomplishes in one experiment what otherwise might require two or more separate studies, provides an opportunity to study interactions that are often very important in educational research, and provides a more powerful test of hypotheses.

**THINK ABOUT IT 11.2**

Consider the following statements and indicate whether each is true or false:

1. If there is a significant interaction in a factorial design, then at least one of the two main effects also must be significant.
2. A factorial design with two levels of factor A and four levels of factor B would require exactly eight separate groups of subjects.
3. A researcher believes that a special training course will be equally effective for males and females. This researcher is predicting an interaction between training and gender.
4. In a two-factor experimental design, the researcher obtains two separate scores for each subject.

**Answers**

1. False; 2. True; 3. False; 4. False

## OTHER RANDOMIZED EXPERIMENTAL DESIGNS

The experimental designs we have discussed so far use at least two groups of subjects, one of which is exposed to the treatment (independent variable) and the other that does not receive the treatment or is exposed to another level of the treatment. The researcher then compares the dependent variable scores for the different treatment groups. The essential feature of these designs is that they compare separate groups of subjects in order to determine the effect of the treatment. When the independent variable is manipulated in this way, we have what is called a **between-subjects design.** For example, a researcher who compares reading achievement scores for students taught by one method with scores for an equivalent group of students taught by a different method is using a between-subjects design.

However, the manipulation of an independent variable does not have to involve different groups of subjects. It is possible to use experimental designs in which the same participants are exposed to different levels of the independent variable at different times. For example, a researcher might measure the learning of nonsense syllables by one group of students under different levels of anxiety or the math performance scores of a group of students when music is played in the classroom versus no music. This type of design in which a researcher observes each individual in all of the different treatments is called a **within-subjects design.** It is also called a **repeated-measures design** because the research repeats measurements of the same individuals under different treatment conditions. The main advantage of a within-subjects design is that it eliminates the problem of differences in the groups that can confound the findings in between-subjects research. Remember that one is not comparing one group of subjects to another; one is comparing each individual's score under one treatment with the same individual's score under another treatment. Each subject serves as his or her own control. Another advantage of within-subjects designs is that they

can be conducted with fewer subjects. The disadvantage of these designs is the carryover effect that may occur from one treatment to another. To deal with this problem, researchers typically arrange for the participants to experience the different treatments in random or counterbalanced order.

Discussion of the different designs for within-subjects experimental research is beyond the scope of this text. Interested readers should consult advanced texts.

## QUASI-EXPERIMENTAL DESIGNS

We have discussed various true experimental designs, all of which require random assignment of subjects. In many situations in educational research, however, it is not possible to randomly assign subjects to treatment groups. Neither the school system nor the parents would want a researcher to decide to which classrooms students were assigned. In this case, researchers turn to quasi-experiments in which random assignment to treatment groups is not used. **Quasi-experimental designs** are similar to randomized experimental designs in that they involve manipulation of an independent variable but differ in that subjects are not randomly assigned to treatment groups. Because the quasi-experimental design does not provide full control, it is extremely important that researchers be aware of the threats to both internal and external validity and consider these factors in their interpretation. Although true experiments are preferred, quasi-experimental designs are considered worthwhile because they permit researchers to reach reasonable conclusions even though full control is not possible.

### Design 9: Nonrandomized Control Group, Pretest–Posttest Design

In a typical school situation, schedules cannot be disrupted nor classes reorganized to accommodate a research study. In such a case, one uses groups already organized into classes or other preexisting intact groups.

The **nonrandomized control group, pretest–posttest design** is one of the most widely used quasi-experimental designs in educational research. You can see that it is similar to Design 5 but with one important difference: Design 9 does *not* permit random assignment of subjects to the experimental and control groups.

Design 9: Nonrandomized Control Group, Pretest–Posttest Design

| Group | Pretest | Independent Variable | Posttest |
|---|---|---|---|
| E | $Y_1$ | $X$ | $Y_2$ |
| C | $Y_1$ | — | $Y_2$ |

A researcher might be allowed to use two sections of freshman English at a high school for a study on vocabulary development. The researcher should select two sections that at least appear to be similar; for example, one should not choose a remedial class and an advanced class. Although subjects cannot be randomly assigned, one can flip a coin to determine which of the two intact groups will be the experimental group and which will be the control group. The researcher would give a vocabulary pretest to both classes, administer a program designed to improve vocabulary to the experimental group only, and then give a vocabulary posttest to

both groups. If the experimental group shows significantly greater achievement on the posttest, can the researcher conclude that the new program was effective?

Without random assignment of subjects, you do not know if the groups were equivalent before the study began. Perhaps the class designated the experimental group would have done better on the posttest without the experimental treatment. Thus, there is an initial *selection bias* that can seriously threaten the internal validity of this design. The pretest, the design's most important feature, provides a way to deal with this threat. The pretest enables you to check on the equivalence of the groups on the dependent variable before the experiment begins. If there are no significant differences on the pretest, you can discount selection bias as a serious threat to internal validity and proceed with the study. If there are some differences, the investigator can use ANCOVA to statistically adjust the posttest scores for the pretest differences.

Because both experimental and control groups take the same pretest and posttest, and the study occupies the same period of time, other threats to internal validity, such as maturation, instrumentation, pretesting, history, and regression (if groups are not selected on the basis of extreme scores), should not be serious threats to internal validity. Having the same person teach both English classes would be recommended.

There are some possible internal validity threats, however, that this design does not control, namely threats resulting from an interaction of selection and some of the other common threats.

*Interaction of Selection and Maturation* We have stated that maturation per se is not a serious threat in this design because both groups would mature during the course of the experiment. The problem arises when the two groups differ in their propensity to maturation. If one of the selected groups is more subject to maturation than the other, you have an *interaction of selection and maturation* to threaten internal validity. Suppose Section 1 of freshman English meets at the same hour that the remedial mathematics class meets, whereas Section 3 meets at the same hour as the advanced algebra class. You would expect Section 1 as a group to show the most gain on academic performance on the posttest because many of the poorer students are in the remedial mathematics class. Section 3 would be expected to show the least gain because many of the best students are in the advanced algebra class at the same time. The timing of the mathematics classes influences the makeup of the English sections, and because of the interaction between selection and maturation, the gain to be expected in the sections is affected.

Selection–maturation interaction can be a particularly difficult problem when volunteers are compared with nonvolunteers. For example, let us say you offer an after-school reading improvement program to those who wish it. Reading pretest means show no difference between those who volunteer for the after-school program and those who do not. If the posttreatment scores show greater gain for the treatment group than for the control group, you cannot confidently attribute the greater gain to the treatment. It is quite possible that students who were willing to participate in the after-school program were more concerned about their reading or their parents were more concerned about their reading and they were therefore more likely to show greater gain in reading whether they received treatment or not.

*Interaction of Selection and Regression* A selection–regression interaction could occur in this design if you drew the groups used in the study from populations having different means. Even though the groups are equivalent on a pretest, regression would occur if the experimental group's mean of 75 on the pretest is below the mean of its parent population, whereas the control group with a pretest mean of 75 is above the mean of its parent population. Because each group will regress toward the mean of the parent population when retested, the experimental group will be expected to have a higher mean on the posttest, whether or not a treatment is introduced; on the other hand, the mean of the control group will regress downward. The experimental group will appear to have made more progress during the course of the study than the control group, which would most likely be erroneously attributed to the effect of the treatment.

*Interaction of Selection and Instrumentation* As noted in Chapter 8, many educational tests have a ceiling, which means that range of achievement on the test is limited. This characteristic could result in changes occurring for one group that apparently do not occur for the other. The gains are limited by the difference between the posttest's ceiling and the magnitude of the pretest score. If a student answers 92 items correctly on a 100-item pretest, this student can only gain 8 points on the 100-item posttest. However, a student with a score of 42 on the pretest could make a gain of 58 points on the posttest. The ceiling effect thus restricts students with high pretest scores to a low change score. They may have improved greatly, but the instrument cannot show the gain. In contrast, low-scoring students on the pretest show a large gain because they have more room to improve their scores. You might erroneously conclude that the treatment was more effective with the latter group. This illustrates the problem with change or gain scores; an analysis of covariance with the pretest scores as the covariate would be the recommended procedure for analyzing the posttest scores.

In summary, the nonrandomized control group, pretest–posttest design is a good second choice when random assignment of subjects to groups is not possible. The more similar the experimental and the control groups are at the beginning of the experiment, and the more this similarity is confirmed by similar group means on the pretest, the more credible the results of the nonrandomized control group pretest–posttest study become. If the pretest scores are similar and selection–maturation and selection–regression interactions can be shown to be unlikely explanations of posttest differences, the results of this quasi-experimental design are quite credible.

Even if the group means are noticeably different before treatment, a nonequivalent control group is better than a preexperimental design that has no control group at all. Design 9 can be extended to employ more than two groups. The threats to external validity in Design 9 are similar to those encountered with Design 5. An advantage of Design 9, however, is that the reactive effects of experimentation are more easily controlled than they are in Design 5. When intact classes are used, subjects are probably less aware of an experiment being conducted than when subjects are drawn from classes and put into experimental sessions. This contributes to the generalizability of the findings.

### Design 10: Counterbalanced Design

A **counterbalanced design,** another design that can be used with intact class groups, rotates the groups at intervals during the experimentation. For example, groups 1 and 2 might use methods A and B, respectively, for the first half of the experiment and then exchange methods for the second half. The distinctive feature of Design 10 is that all groups receive all experimental treatments but in a different order. In effect, this design involves a series of replications; in each replication the groups are shifted so that at the end of the experiment each group has been exposed to each *X*. The order of exposure to the experimental situation differs for each group. The following shows a counterbalanced design used to compare the effects of two treatments on a dependent variable:

Design 10: A Counterbalanced Design with Two Treatments

| | Experimental Treatments | |
|---|---|---|
| **Replication** | $X_1$ | $X_2$ |
| 1 | Group 1 | Group 2 |
| 2 | Group 2 | Group 1 |
| | Column mean | Column mean |

A classroom teacher could use a counterbalanced study to compare the effectiveness of two methods of instruction on learning in science. The teacher could choose two classes and two units of science subject matter comparable in the nature of the concepts, difficulty of concepts, and length. It is essential that the units be equivalent in the complexity and difficulty of the concepts involved. During the first replication of the design, class (group) 1 is taught unit 1 by method $X_1$ and class (group) 2 is taught unit 1 by method $X_2$. An achievement test over unit 1 is administered to both groups. Then class 1 is taught unit 2 by method $X_2$ and class 2 is taught unit 2 by method $X_1$; both are then tested over unit 2. After the study, the column means are computed to indicate the mean achievement for both groups (classes) when taught by method $X_1$ or method $X_2$. A comparison of these column mean scores through an analysis of variance indicates the effectiveness of the methods on achievement in science.

A counterbalanced design may be used when several treatments are to be investigated. Table 11.3 shows a counterbalanced design with four treatments.

Each row in Table 11.3 represents one replication. For each replication, the groups are shifted so that group A first experiences $X_1$, then $X_2$, $X_3$, and finally $X_4$. Each cell in the design would contain the mean scores on the dependent variable for the group, treatment, and replication indicated. The mean score for each column would indicate the performance of all four groups on the dependent variable under the treatment represented by the column.

Design 10 overcomes some of the weaknesses of Design 9; that is, when intact classes must be used, counterbalancing provides an opportunity to rotate out any differences that might exist between the groups. Because all treatments are administered to all groups, the results obtained for each *X* cannot be attributed to preexisting differences in the subjects. If one group should have more aptitude

**Table 11.3** A Counterbalanced Design with Four Treatments

| | Experimental Treatments | | | |
|---|---|---|---|---|
| **Replication** | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
| 1 | Group A | B | C | D |
| 2 | Group C | A | D | B |
| 3 | Group B | D | A | C |
| 4 | Group D | C | B | A |
| | Column mean | Column mean | Column mean | Column mean |

on the average than the other, each $X$ treatment would benefit from this greater aptitude.

The main shortcoming of Design 10 is that there may be a carryover effect from one $X$ to the next. Therefore, this design should be used only when the experimental treatments are such that exposure to one treatment will have no effect on subsequent treatments. This requirement may be difficult to satisfy in much educational research. Furthermore, one must establish the equivalence of learning material used in various replications. It may not always be possible to locate equivalent units of material. Another weakness of the counterbalanced design is the possibility of boring students with the repeated testings this method requires.

## TIME-SERIES DESIGNS

### Design 11: One-Group Time-Series Design

The **one-group time-series design** involves periodic measurement on one group and the introduction of an experimental treatment into this time series of measurements. As the design indicates, a number of measurements on a dependent variable are taken, $X$ is introduced, and additional measurements of $Y$ are made. By comparing the measurements before and after, you can assess the effect of $X$ on the performance of the group on $Y$. A time-series design might be used in a school setting to study the effects of a major change in administrative policy on disciplinary incidents. Or a study might involve repeated measurements of students' attitudes and the effect produced by introducing a documentary film designed to change attitudes.

Design 11: One-Group Time-Series Design

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $Y_1$ | $Y_2$ | $Y_3$ | $Y_4$ | $X$ | $Y_5$ | $Y_6$ | $Y_7$ | $Y_8$ |

Figure 11.3 illustrates some possible patterns from time-series studies into which an experimental treatment is introduced. It shows the series of measurements $Y_1$ through $Y_8$, with the introduction of the experimental treatment at point $X$. You can assess the effect of the $X$ by examining the stability of the repeated measurements.

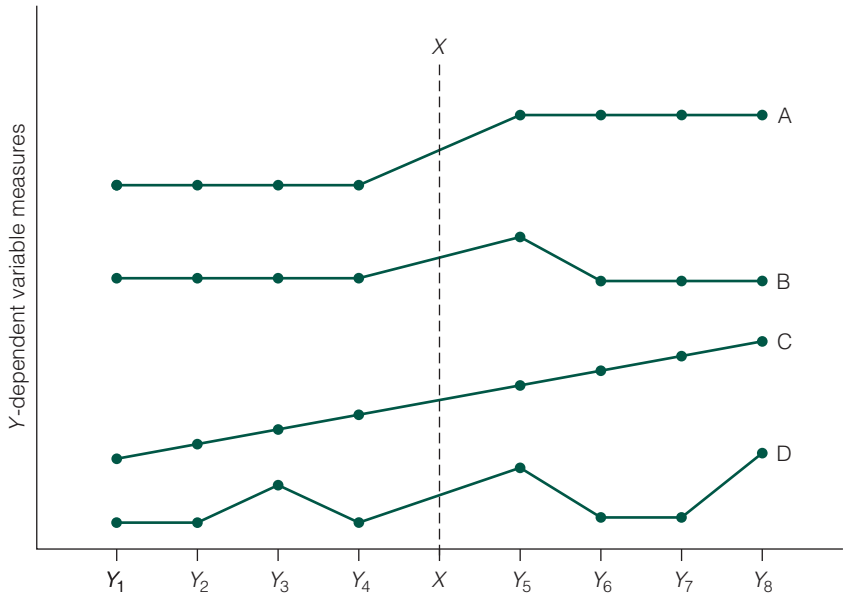



**Figure 11.3** Illustration of Possible Outcome Patterns in a Time Design

From examining the difference between $Y_4$ and $Y_5$ in pattern A in Figure 11.3, perhaps you would be justified in assuming that $X$ affects the dependent variable. Pattern B suggests the possibility of a temporary experimental effect of $X$. However, you could not assume that $X$ produces the change in either pattern C or pattern D. Pattern C appears to result from maturation or a similar influence. The erratic nature of pattern D suggests the operation of extraneous factors.

Design 11 is similar to Design 1 in that it uses before-and-after measures and lacks a control group. However, it has certain advantages over Design 1 that make it more useful in educational research. The repeated testing provides a check on some common threats to internal validity. Maturation, testing, and regression could be ruled out as plausible explanations of the shift occurring between $Y_4$ and $Y_5$ if such shifts do not occur in the previous time periods under observation. It is recommended that no change in measuring instruments be made during the course of the time study. In this way, you eliminate changes in instrumentation as a possible explanation of the $Y_5 - Y_4$ difference.

The major weakness of Design 11 is its failure to control history; that is, you cannot rule out the possibility that it is not $X$ but, rather, some simultaneous event that produces the observed change. Perhaps such factors as seasonal or weather changes or such school events as examinations could account for the change. In a study designed to assess the effect of a lecture–film treatment on student attitudes toward minorities, to what extent would the attitude measurements be affected by a nationally publicized minority riot in a distant city? The extent to which history (uncontrolled contemporary events) is a plausible explanatory factor must be taken into account by the experimenters as they attempt to interpret their findings. You must also consider the external validity of the time design. Because there are repeated tests, perhaps there is a kind of interaction effect of testing that would restrict the findings to those populations subject to repeated testing. However, as

long as the measurements are of a typical, routine type used in school settings, this is not likely to be a serious limitation. Furthermore, a selection–treatment interaction may occur, especially if you select some particular group that may not be typical.

Statistical interpretation can be a particular problem with time data. The usual tests of significance are not appropriate with a time design because they assume that observations are independent of one another; but time-series data are typically correlated with one another. See Chapter 4 of Shadish et al. (2002) for a discussion of the statistical tests that may be used with this design.

### Design 12: Control Group Time-Series Design

The **control group time-series design** is an extension of Design 11 to include a control group. The control group, again representing an intact class, would be measured at the same time as the experimental group but would not experience the $X$ treatment. This design overcomes the weakness of Design 11—that is, failure to control history as a source of extraneous variance. The control group permits the necessary comparison. If the E group shows a gain from $Y_4$ to $Y_5$ but the C group does not show a gain, then the effect must be caused by $X$ rather than by any contemporaneous events, which would have affected both groups.

Design 12: Control Group Time-Series Design

| *Group* | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| E | $Y_1$ | $Y_2$ | $Y_3$ | $Y_4$ | $X$ | $Y_5$ | $Y_6$ | $Y_7$ | $Y_8$ |
| C | $Y_1$ | $Y_2$ | $Y_3$ | $Y_4$ | — | $Y_5$ | $Y_6$ | $Y_7$ | $Y_8$ |

Other variations of the time-series design include adding more control groups, more observations, or more experimental treatments.

## VALIDITY PROBLEMS WITH EXPERIMENTAL DESIGNS

Some sources of invalidity in the one-variable experimental designs are summarized in Table 11.4. This brief summary must not be depended on as the sole guide in selecting a design. It must be accompanied by a thorough consideration of the qualified presentation appearing in the text so that the reader understands the particular strengths and weaknesses that characterize each design.

## SINGLE-SUBJECT EXPERIMENTAL DESIGNS

The single-subject experimental designs are a type of experimental design with a unique feature: the sample size is just one or is composed of a few participants who are treated as one unit. How can an experiment be run with a sample size of one? Obviously, there can be no random assignment or use of control groups. In single-subject experimental designs (also called single-case experimental designs), the participant serves as both the treatment and the control group. The researcher measures participant behavior repeatedly during at least two different points in time, when a treatment is not present and again when a treatment is present. The periods during which the treatment is given are called *treatment periods*, and the periods during which the treatment is not present are called

**Table 11.4** Factors Jeopardizing the Internal Validity of Experimental Designs

| | Designs[a] | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Preexperimental | | True Experimental | | | | | | Quasi-Experimental | | | |
| Sources of Invalidity | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| History[b] | – | + | + | + | + | + | + | + | + | + | – | + |
| Maturation | – | ? | + | + | + | + | + | + | + | + | + | + |
| Pretesting | – | + | + | + | + | + | + | + | + | + | + | + |
| Instrumentation | – | + | + | + | + | + | + | + | + | + | ? | + |
| Statistical regression | ? | + | + | + | + | + | + | + | ? | + | + | + |
| Differential selection | + | – | + | + | + | + | + | + | – | + | + | + |
| Experimental mortality | + | – | + | + | + | + | + | + | + | + | + | + |
| Interaction of selection and other threats | – | – | + | + | + | + | + | + | – | ? | + | + |
| Subject effects | – | – | ? | ? | ? | – | – | – | – | + | – | – |
| Experimenter effect | – | – | ? | ? | ? | – | – | – | – | – | – | – |
| Diffusion | na | – | ? | ? | ? | ? | ? | ? | – | ? | na | ? |

[a]Designs are as follows:

1. One-group pretest–posttest
2. Static group comparison
3. Randomized Ss, posttest-only control group
4. Randomized matched Ss, posttest-only control group
5. Randomized Ss, pretest–posttest control group
6. Solomon, three groups
7. Solomon, four groups
8. Simple factorial
9. Nonrandomized control group, pretest–posttest
10. Counterbalanced
11. One-group time series
12. Control group time series

[b]A plus sign indicates that the factor is controlled; a minus sign indicates lack of control, and a question mark indicates a possible source of concern.

*baseline periods*. The baseline period is usually referred to by the letter A and the treatment period by the letter B.

For example, a teacher might want to know the effect of a certain reward (treatment or independent variable) on the time-on-task behavior (dependent variable) of a child with attention deficit/hyperactivity disorder. The child's behavior would be measured before the treatment (baseline), during the treatment, and after the treatment. The data for the baseline period would serve as the control group data and would be compared with the data during obtained the treatment and after the treatment period. **Single-subject research** has become popular during the past 30 years as proponents of this particular methodology have demonstrated that experimental control can be effectively achieved in other than the traditional ways. After describing the two major approaches to single-subject research and the rationale behind them, we examine the strengths and limitations of this type of research in comparison with the other more conventional designs.

Study of the individual has always had a place in educational and psychological research. Freud's case studies and Piaget's observations of individual children are notable examples. Although case studies (see Chapter 16) and single-subject experiments both study the individual, in a single-subject experiment, the investigator deliberately manipulates one or more independent variables, whereas in

a case study the observer studies the subjects' interaction with events that occur naturally.

Single-case designs have been particularly useful in clinical applications in which the focus is on the therapeutic value of an intervention for the client. A teacher of severely learning-disabled children, for example, would want information regarding the effectiveness of a specific procedure with an individual child. Some believe that studies that report mean or average differences for groups may have little meaning when treating a specific individual. Single-subject designs are basically extensions of the quasi-experimental one-group time-series design (Design 11). The two most common are the ABAB and the multiple-baseline designs.

### ABAB Designs

The **ABAB design** consists of a period of no treatment, or baseline (A), during which the behavior of interest is repeatedly measured. Examples of such behaviors are the number of times a student with cognitive disability leaves her seat or the number of times an autistic child bangs his head. This pretreatment assessment serves as a control period with which treatment effects will be compared. After a stable picture of pretreatment behavior has been established, phase B, the treatment, is initiated. In the treatment phase, the child might be given tokens (exchangeable for desired privileges) for time periods spent seated or time periods with no self-abusive action. The behavior is consistently monitored throughout the treatment phase, usually until the intervention appears to have taken effect and the rate of behavior stabilizes. Further experimental control is achieved by a second A phase. This is usually a withdrawal of treatment, but in some cases the second A phase is actually a reversal of treatment, reinforcing a behavior incompatible with the desired response. Discontinuing the giving of tokens to the child constitutes a withdrawal of treatment, whereas giving her tokens when she leaves her seat is a reversal of treatment. In either case, removal of treatment is expected to cause the behavior to return to the original (first baseline) level. Ending the experiment with the second A phase has the drawback of leaving the client in the same state as before the experiment started. For ethical reasons and to add strength to the design by replicating the procedure, the treatment phase (B) is again instituted. Many variations on the ABAB design are possible. More than one treatment can be tested—for instance, using an ABCACB format in which A is no treatment and B and C are alternate treatments.

Figure 11.4 illustrates an ABAB design. Mornings were often difficult for the entire family of an 8-year-old developmentally disabled boy, Curt, because he took up to 2 hours to get dressed in the morning. Constant reminders, pleadings, and occasional scoldings had not helped. Once during a week of baseline measurement (A), Curt did get dressed in only 4 minutes, so his parents knew he was capable of dressing himself in a reasonable time.

During the treatment (B) phase, each morning Curt's clothes were laid out in the bathroom, he was awakened, and then a kitchen timer was set for 10 minutes. Curt earned a red sticker to put on the chart on his door if he dressed himself before the timer rang. The sticker entitled him to watch TV that evening. If he did not finish within the 10 minutes, he had to stay in the bathroom until dressed and forfeit TV for the day. Curt was praised when he met the criterion, and he was ignored when he did not.

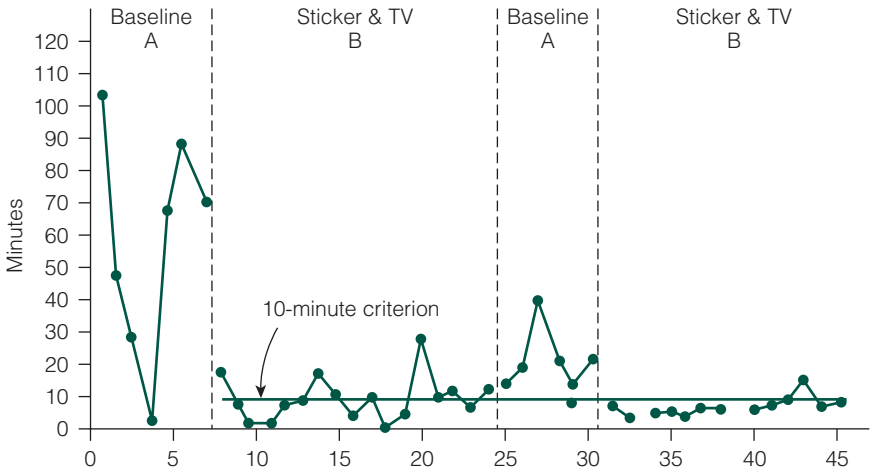



**Figure 11.4** ABAB Design (Use of Differential Reinforcement to Decrease Dressing Time of an 8-Year-Old Boy)

*Source*: From *Working with parents of handicapped children* (p. 59) by W. Heward, J. C. Dardig, & A. Rossett, 1979, Columbus, OH: Merrill. Copyright 1979 by Bell & Howell Company. Reprinted by permission.

During the week of baseline, Curt took an average of 59 minutes to get dressed. During 17 days of the first B phase, Curt met the criterion nine times and his average was 10 minutes. On three occasions, 0 minutes was recorded because he woke and dressed himself before his parents were awake.

During the following 6 days, Curt's clothes were laid out for him, but the timer was omitted and he was allowed to watch TV no matter how long he took to dress. His dressing time averaged 23 minutes during this second A phase. When treatment was reinstituted for 12 of the next 13 days, Curt's dressing time then averaged 8 minutes.

Because treatment was always accompanied by a change in dressing time, a credible relationship between treatment and dressing time was established. The key element in the ABAB design is the return to former levels of behavior when the baseline condition is reinstated. In certain circumstances, the assumption that the intervention is the cause of the change can be quite weakened. If there is considerable variability during the baseline period, you could argue that post-treatment behavior did not differ meaningfully from pretreatment behavior. If the shift from treatment (B) back to baseline (A) is equivocal, much of the power of the ABAB design is gone. The second baseline often does not show the same extreme of behavior as the first, and in some cases there is no return to previous rates. In some cases, other factors, such as history or maturation, could cause the observed effect.

Some researchers argue that returning to baseline conditions may be unethical in certain circumstances. Do you really want the autistic child to return to the previous levels of head-banging just to provide evidence that the treatment caused the reduction? Because of these potential problems with the ABAB design, the multiple-baseline design was developed. This design has the clinical advantage of continuing any improvement made without returning to less desired behavior.

### Multiple-Baseline Designs

In a **multiple-baseline design,** observations are made on several participants, different target behaviors of one or more subjects, or different situations. Thus, there are three basic designs: **multiple-baseline across-participants design, multiple-baseline across-behaviors design,** and **multiple-baseline across-settings design** (Evans & Rooney, 2008). In a multiple-baseline across-behaviors design, the same intervention is applied to similar behaviors to the same individual in the same setting. For example, the researcher might record the number of times a student talked in class without permission, the number of times a student got out of his or her seat without permission, and the number of times a student hit another child. In a multiple-baseline across-participants design, the same intervention is applied to the same or similar behaviors of different individuals in the same setting. For example, the same behavior of several clients, such as amount of eye contact with the teacher, could be recorded for two or more children in a preschool class during the baseline phase. In both cases, the treatment (B) is the same across all conditions or students. In a multiple-baseline across-settings design, the same intervention is applied to the same behavior to the same individual in different settings. For example, one might want to know if a type of reinforcement is as effective with an individual in a math class as it is in a reading class.

Experimental control in the multiple baseline results from starting the treatment at a different point in time for each behavior and/or person involved rather than from returning to baseline. Thus, after the baseline is established, treatment for behavior 1 is instituted and the baseline is continued for behaviors 2 and 3. When treatment for behavior 2 is instituted, treatment for behavior 1 and baseline for behavior 3 are continued. Finally, treatment for behavior 3 is instituted. It is expected that each behavior will change in the desired direction at the point at which treatment is begun, not before or after. If this happens, we have confidence that the treatment effected the change.

Thus, the multiple-baseline design uses two or more basic AB units. If some outside event other than the treatment was the actual cause of the changes, it should affect all children or all behaviors at the same point in time. One assumption of this design is that treatment affects different behaviors specifically. Reinforcing one behavior (completing arithmetic problems) is not expected to increase another response (reading rate). The behaviors, or situations, must be independent (uncorrelated) for the multiple-baseline study to show interpretable effects. In actuality, independence of behavior may be difficult to attain. Modifying one behavior (such as talking in class) may influence other targeted behaviors (completing assignments on time).

Lane, Little, Redding-Rhodes, Phillips, and Welsh (2007) used a multiple-baseline design to investigate the effectiveness of a teacher-implemented reading intervention on the behavior of seven first-grade students (three girls and four boys) at risk for emotional/behavioral disorders and reading difficulties. The participants were in two general education classrooms in a rural elementary school. Researchers measured reading and behavioral performance twice weekly during the 10-week baseline period and once a week during postintervention. The intervention used PALS (Peer-Assisted Learning), in which participants were paired with teacher-selected students who had above-average reading skills. The pairs

who remained together during the entire intervention phase worked on sounds, word speed games, reading aloud, and partner reading. The outcomes measured were decoding, oral reading fluency, and academic engagement. The results revealed increases in oral reading fluency relative to the baseline for all students, accompanied by decreases in variability of academic engagements for four students.

## COMPARISON OF SINGLE-SUBJECT AND GROUP DESIGNS

In both single-subject and group experiments, the goal of the experimenter is to establish as unequivocally as possible the connection between the manipulation of the independent variable (treatment) and its effect on the dependent variable (behavior). In group designs, random assignment of subjects to experimental or control groups eliminates many rival explanations of differences observed after treatment. Treatment effects (between groups) can be assessed relative to intersubject variability effects (within group) by using appropriate statistical tests. These tests determine whether chance alone is a credible explanation for the results. The single-subject design uses other methods to establish credibility. The experimenter controls the amount of time in which baseline and treatment phases are in effect, and the length of the baseline period can be extended until the behavior stabilizes. For unambiguous interpretation, the baseline should be relatively flat or the trend should be in the opposite direction from that expected after treatment. One drawback to experimenter control of the length of the treatment phase is the tendency to continue treatment until "something happens." If behavior change does not closely follow the beginning of treatment, it is possible that another, nonexperimental variable is the cause of the observed change.

Single-subject experimental designs do bypass one source of error of group designs—namely, intersubject variability. Each individual serves as his or her own control, so comparability is not a problem. The major means of control is replication, a feature seldom incorporated into group designs. The ABAB design involves a single replication using the same subject, whereas the multiple-baseline design replicates more than one treatment. Replication of the multiple-baseline design makes it less likely that effects attributed to treatment were in fact caused by extraneous event or subject variables.

Well-designed single-subject research can meet the criteria for internal validity. However, the question of external validity—the generalizability of experimental findings—is not as easily answered by designs that use only one or a few subjects. You can demonstrate that allowing a behaviorally disordered teenager to listen to rock music contingent on completing assignments increases the amount of schoolwork done by that particular teenager, but how can you determine whether this treatment will be successful with other teenagers or all behaviorally disordered teenagers? Although any one particular single-subject study will be low in external validity, a number of similar studies that carefully describe subjects, settings, and treatments will build the case for wide application of particular treatment effects. For further discussion of single-subject research, the reader is referred to Kennedy (2005).

## SUMMARY

The design one chooses to use is very important to the validity of any conclusions that might be drawn from the research. We classified designs as preexperimental, randomized experimental, or quasi-experimental depending on the degree of control provided. The preexperimental designs provide little or no control of extraneous variables and are not recommended.

Randomized experimental designs are the best for investigating causal relationships among variables. As the name indicates, they require random assignment of subjects to levels of the independent variable. Because of the randomization, they provide the best control of the factors that threaten internal and external validity and are recommended for use whenever possible. Experiments may use different groups of subjects for the different experimental conditions, or the experiment may have each subject experience every condition. The former are called between-subjects experiments and the latter are called within-subjects experiments.

Factorial designs use two or more independent variables, each having at least two levels. Factorial designs enable the researcher to investigate the main effects of each of the independent variables on the dependent variable as well as the interaction effect of the independent variables. An interaction exists when variable A has a different effect on the dependent variable when it is combined with one level of variable B than when it is combined with another level of variable B. Thus, one factorial design experiment can accomplish more than two single, independent variable experiments.

Quasi-experimental designs are used when the investigator cannot randomly assign subjects to treatments. There is less control of extraneous variables in quasi-experiments; thus, they are subject to a variety of different threats to internal validity. Quasi-experimental designs that study the effect of treatment on a single subject have proven useful in behavioral research. Single-subject experimental designs use one or a few participants to investigate the effect of a new procedure or an intervention. The basic characteristic is that measures of the dependent variable are repeated before and after an intervention or treatment is implemented. The researcher compares the difference in behavior before and then after the intervention. Single-subject research is widely used in research with exceptional children and in counseling situations.

## KEY CONCEPTS

ABAB design
between-subjects design
control group time-series design
counterbalanced design
experimental design
factorial design
multiple-baseline across-behaviors design
multiple-baseline across-participants design
multiple-baseline across-settings design
multiple-baseline design
nonrandomized control group, pretest–posttest design
one-group pretest–posttest design
one-group time–series design
preexperimental design
quasi-experimental design
randomized experimental design
randomized matched subjects, posttest-only control group design
randomized subjects, posttest-only design
randomized subjects, pretest–posttest control group design
repeated-measures design
single-subject research
Solomon four-group design
Solomon three-group design
static group comparison
within-groups design
within-subjects design

## EXERCISES

1. From a group of students enrolled in social studies in a high school, a researcher randomly selected 60 students. The students were then divided into two groups by random assignment of 30 to group A, the traditional social studies curriculum, and 30 to group B, a new program designed to deal with the history of certain ethnic groups. The two groups were compared at the end of the semester on a scale designed to measure attitudes toward ethnic groups. In this study, identify the following:
   - **a.** Independent variable
   - **b.** Dependent variable
   - **c.** Control group
   - **d.** Experimental group
   - **e.** Method(s) used to control for differences between the groups
   - **f.** Research design used
   - **g.** Any threats to internal validity
2. Consider the following research question: Does teaching the first year of French through an oral–aural approach, rather than the grammar-transformational method, alter pupil performance on a standardized year-end test in grammar, reading, and vocabulary?
   - **a.** Identify the ideal experiment to answer this question, assuming that there are no administrative or other restrictions.
   - **b.** Identify the experiment that would most likely be required in the typical high school setting.
   - **c.** State the relative advantages of the ideal experimental design (Exercise 2a) compared with the design in Exercise 2b.
3. Design the ideal experiment to test the following hypothesis: Children who view films of harmonious racial interaction will show a more positive attitude toward racial minorities than will children shown films that depict racial conflict.
4. Returning to the research problem in Exercise 2, suppose you also want to know if the two methods of teaching French have differential effects for boys and girls. Outline the experimental design that would permit you to answer this question at the same time.
5. Assume an investigator used two methods of instruction ($A_1$ and $A_2$) with two groups of students ($B_1$ and $B_2$) having varying levels of achievement motivation. The groups were compared on an achievement test at the end of the study. The means are presented here. What interpretation would you make of these results?

| | | *Method* | |
|---|---|---|---|
| | | $A_1$ | $A_2$ |
| Motivation | $B_1$ | 35 | 15 |
| | $B_2$ | 15 | 35 |

6. What must occur to establish credibility of results in a single-subject design?
7. A researcher wants to test the effectiveness of two different computer-based instructional methods on teaching a unit on weather in seventh-grade science. The researcher finds a teacher who will permit three of his classes to participate: Two classes use the computer-based instruction, and the third receives conventional classroom instruction. The researcher administers a pretest to all students, each class has a different method of instruction, and then a posttest is administered.
   - **a.** Specify the design of this study and represent it using the notation system used in this chapter.
   - **b.** What statistics would you choose to analyze the data?
8. Indicate whether an ABAB design or a multiple-baseline design would be appropriate for the following:
   - **a.** You have a student with an extreme nail-biting problem and you want to work with her to eliminate the behavior.
   - **b.** You have a student who has several problem behaviors that are disruptive and interfering with learning in the classroom. You have asked the school psychologist to help.

## ANSWERS

**1. a.** Type of social studies curriculum
**b.** Scores on an ethnic attitude scale
**c.** Group A, the present curriculum
**d.** Group B, the curriculum with ethnic history
**e.** Random selection of the sample from the population and random assignment of the sample to the experimental and control groups
**f.** Design 3, the randomized subjects, posttest-only control group design
**g.** In the event that randomization does not control for initial group differences, there is no pretest to use to check if there are differences in attitudes before the study.

**2. a.** Use Design 3—that is, randomly assign first-year French students to either the grammar-transformational (control) or oral–aural (experimental) group. Maintain the same conditions, time spent, teachers, and classroom facilities for both groups so that only the teaching method is different. Administer test at the end of the year and compare group achievement.
**b.** Randomly assign intact classes of first-year French students to the two teaching methods. Each teacher has an equal number of the two types of classes.
**c.** In the ideal design, threats to external and internal validity are better controlled through randomization of individual students. The design in Exercise 2b could have problems with nonequivalence of subjects before treatment is given so that test score differences could be caused by factors other than difference in treatment.

**3.** The ideal experiment would randomly assign students to groups. The results of a posttreatment measure of attitudes toward racial minorities would be used to compare the experimental and control groups.

**4.** This question requires a factorial design with half of the boys and half of the girls assigned randomly to the control and experimental conditions.

**5.** It appears that there is an interaction between achievement motivation and type of instruction. Students with achievement motivation at level $B_1$ did better with method $A_1$, whereas those at level $B_2$ did better with method $A_2$. The significance of this interaction could be tested with an $F$ test. There is no overall effect of motivation or instructional method because the means for $A_1$ and $A_2$, and $B_1$ and $B_2$, are the same.

**6.** There must be an unambiguous change in behavior whenever there is a change in treatment.

**7. a.** Nonrandomized control group, pretest–posttest design:

| | | | |
|---|---|---|---|
| E | $Y_1$ | $X$ | $Y_2$ |
| E | $Y_1$ | $X$ | $Y_2$ |
| C | $Y_1$ | — | $Y_2$ |

**b.** If the groups were not significantly different on the pretest, you could use analysis of variance on the posttest scores. Otherwise, ANCOVA could be used with pretest scores as the covariate.

**8. a.** ABAB
**b.** Multiple baseline

## REFERENCES


Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.

Evans, A. N., & Rooney, B. F. (2008). *Methods in psychological research*. Thousand Oaks, CA: Sage.

Fisher, R. A. (1925). *Statistical methods for research workers*. London: Oliver & Boyd. [Reprinted 1970]

Heward, W., Dardig, J. C., & Rossett, A. (1979). W*orking with parents of handicapped children*. Columbus, OH: Merrill.

Kennedy, C. H. (2005). *Single-case designs for educational research*. Boston: Pearson.

Lane, A., Little, M. A., Redding-Rhodes, J., Phillips, A., & Welsh, M. (2007). Outcomes of a teacher-led reading intervention for elementary students at risk for behavioral disorders. *Exceptional Children*, *74*(1), 47–70.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Solomon, R. L. (1949). On extension of control group design. *Psychological Bulletin*, *46*, 137–150.

CHAPTER 12

# Ex Post Facto Research

Not all important questions in education can be answered with experimental research.

## INSTRUCTIONAL OBJECTIVES

After studying this chapter, the student will be able to:

1. Describe ex post facto research and compare it to experimental research.
2. State conditions needed to infer a causal relationship.
3. Describe alternative explanations in ex post facto research and identify cases in which these are or are not plausible.
4. Describe methods of partial control and identify cases in which they would be useful.
5. Distinguish between proactive and retroactive ex post facto designs.
6. Identify questions for which ex post facto research would be the method of choice.
7. State the major weakness of ex post factor research.

As researchers probe such educational questions as "Why are some children better readers than others?" "What is the effect of single-parent homes on achievement?" and "Why do some youths become delinquent while others do not?" they find that only some questions can be investigated through experimental research. If you want to investigate the influence of such variables as home environment, motivation, intelligence, parental reading habits, age, ethnicity, gender, disabilities, self-concept, and so forth, you cannot randomly assign students to different categories of these variables. Independent variables such as these are called **attribute independent variables.** An attribute variable is a characteristic that a subject has before a study begins.

In contrast, an independent variable that an investigator can directly manipulate is an **active independent variable.** An investigator can determine which students will have access to a computer laboratory and which will not or which will use program A to study a unit in algebra and which will use program B. When active independent variables are involved, an investigator can employ experimental or quasi-experimental research. When an investigation involves attribute independent variables that the researcher cannot manipulate, he or she must turn to *ex post facto* research. Ex post facto research is also appropriate when the variable actually could be manipulated but is not because it would be unethical or irresponsible to do so. For example, it would not be ethical to manipulate illegal drug use or use of alcohol or cigarettes to study their effects on human subjects. To study the effect of retention on subsequent achievement, you would not want to randomly assign some children to be retained and others to be promoted. Also, you would not manipulate the use of a drug such as Ritalin to study its effects on children's problem-solving behavior. In such cases, a researcher could use ex post facto research to compare the subsequent achievement of students who have been retained with the academic achievement of otherwise equivalent students who have been promoted or to compare the problem-solving behavior of a group of children already taking Ritalin with that of a matched group not taking the drug.

The designation *ex post facto*, from Latin for "after the fact," indicates that **ex post facto research** is conducted after variation in the variable of interest has already been determined in the natural course of events. This method is sometimes called *causal comparative* because its purpose is to investigate cause-and-effect relationships between independent and dependent variables. Researchers use it in situations that do not permit the randomization and manipulation of variables characteristic of experimental research. Thus, much of the basic rationale for experimental and ex post facto is the same. They both investigate relationships among variables and test hypotheses.

However, with an experiment it is possible to obtain much more convincing evidence for a **causal** (functional) **relationship** among variables than can be obtained with ex post facto studies. The effects of extraneous variables in an experiment are controlled by the experimental conditions, and the antecedent independent variable is directly manipulated to assess its effect on the dependent variable. If you observe $Y$ to vary concomitantly with the variation in $X$ in this controlled situation, then you have obtained evidence for the validity of the hypothesized antecedent–consequent relationship between $X$ and $Y$. In an ex post facto investigation, in contrast, the researcher cannot control the independent variables by manipulation or by randomization: Changes in the variables have already taken place. Because of this lack of control, in an ex post facto study it is more hazardous to infer a genuine relationship between $X$ and $Y$.

Let us illustrate the difference between ex post facto and an experimental approach by examining these two approaches to the same research question. Consider the question of the effect of students' anxiety in an achievement-testing situation on their examination performance. The ex post facto approach would involve measuring the already existing anxiety level at the time of the examination and then comparing the performance of "high anxious" and "low anxious" students. The weakness of such an approach is that you could not necessarily conclude that the students' anxiety produced the observed difference

in achievement examination performance. Both sets of scores may have been influenced by a third factor, such as knowledge of the subject matter being examined or general self-confidence. Knowledge or aptitude may be the major cause of both the level of anxiety and the achievement test results.

In an experimental approach to the same problem, the investigator could randomly assign subjects to two exam conditions that are identical in every respect except that one is anxiety arousing and the other is neutral. The experimenter can induce anxiety by telling the subjects that their final grade may depend on their performance, that they may find the test to be extremely difficult, or that the test can be used to identify the incompetent. The neutral group would merely be told that their cooperation is needed for the experiment. The investigator could randomly assign subjects to the two conditions. Then, if the anxious group performed better than the neutral group, it could be concluded that the induced anxiety had a facilitating effect on test performance. If the neutral group performed better than the anxious group, it could be concluded that the induced anxiety had a debilitating effect. A conclusion could be legitimately drawn because of the control provided by the random assignment of groups to treatments and by the experimenter's direct manipulation of the independent variable. Anxiety is one of the few variables that can be either an active or an attribute independent variable. You can manipulate it actively, as described (experimental approach), or you can classify subjects on the basis of their scores on an anxiety measure (ex post facto approach).

Ex post facto research, unlike experimental research, does not provide the safeguards that are necessary for making strong inferences about causal relationships. Mistakenly attributing causation based on a relationship between two variables is called the **post hoc fallacy.** An investigator who finds a relationship between the variables in an ex post facto study has secured evidence only of some concomitant variation. Because the investigator has not controlled *X* or other possible variables that may have determined *Y*, there is less basis for inferring a causal relationship between *X* and *Y*.

If you wish to reach a conclusion that one variable (*X*) is the cause of another variable (*Y*), three kinds of evidence are necessary:

1. A statistical relationship between *X* and *Y* has been established.
2. *X* preceded *Y* in time.
3. Other factors did not determine *Y*.

A statistical relationship is one in which a change in one variable can be predicted from a change in the other. However, we must note that such a relationship between two variables, by itself, is not sufficient evidence of cause and effect. You must proceed to look for evidence on the other two criteria.

Therefore, the investigator must also establish the time sequence; that is, you must consider whether *Y* might have occurred before *X* and hence could not be an effect of *X*. If *X* is a cause of *Y*, then a change in *X* must precede a change in *Y*. Decisions about the time relationship between *X* and *Y* can be made either on a logical basis or as a result of measurements that show the groups did not differ on *Y* before exposure to *X*.

The third kind of evidence shows that no other plausible explanation can account for the observed effect. It is extremely important that the investigator

consider whether factors other than *X* might have determined *Y*. This is the most difficult evidence to obtain. You proceed to check this possibility by introducing other relevant variables into the analysis and observing how the relationship between *X* and *Y* is affected by these additional variables. You may find that the relationship between *X* and *Y* holds up even when the other variables are introduced. In this case, you have evidence to support a causal inference. However, you may find that other variables may account for the apparent relationship between *X* and *Y*. In this case, you would conclude that the relationship between *X* and *Y* is spurious. A **spurious relationship** is one in which the two variables really have no effect on each other but are related because some other variable influences both. For example, a positive relationship between the number of churches and the number of armed robberies in the cities within a state does not mean that building more churches will increase armed robberies or that increasing armed robberies will cause more churches to be built. Here, the extraneous variable is city size. Large cities have more churches and more armed robberies, whereas small cities have fewer armed robberies and fewer churches.

### THINK ABOUT IT 12.1

1. Why does the current administration of the U.S. Department of Education prefer randomized experimental research to ex post facto research?
2. Why do researchers conduct ex post facto research?

**Answers**

1. Because active independent variables are factors that one can change; therefore, true experimental research yields information on which proposed programs have sufficient evidence to obtain funding.
2. Because many important independent variables in education cannot be deliberately manipulated; or, in some cases they could be manipulated, but it would be unethical to do so.

## PLANNING AN EX POST FACTO RESEARCH STUDY

1. *The first step in an ex post facto study is to state the research problem, usually in the form of a question.* What is the relationship between variable A and variable B? Or, what is the effect of variable A on variable B? Again, in this type of research the variables involved are those that the researcher does not directly manipulate. The researcher then states a hypothesis about the expected relationship and defines the variables in operational terms.
2. *Next, select two or more groups to be compared.* Recall that investigators doing ex post facto research achieve the variation they want not by directly manipulating the variable but by selecting individuals in whom the variable is present or absent, strong or weak. Thus, these two groups should differ on the variable of interest in that one group should possess the characteristic and the other group should not, but they should be similar on any relevant extraneous variables. Differential (subject) selections pose a major

threat to the internal validity of ex post facto investigations because you have no control over the selection of subjects into the two groups. They are selected because they already possess the variable of interest, for example, smoker/nonsmoker and retained/not retained. Whenever assignment is not random, there is always an opening for other variables to enter to explain the observed difference between the groups. The way to deal with this threat is to collect data to show that the groups are similar on other extraneous variables that might affect the variable of interest. For example, if you were studying the effect of preschool attendance on the social maturity of kindergarteners, you would have to control any other factors that might have been shown to influence social maturity. Some of these might be age, gender, socioeconomic status, and aptitude. You use logic and previous research to determine what factors need to be controlled in an ex post facto study.

3. *Determine whether your question requires a proactive or a retroactive design.* There are two types of ex post facto research designs—proactive and retroactive.

| | Independent Variables | Dependent Variables |
|---|---|---|
| Proactive design | Begin with subjects grouped on the basis of preexisting independent variables. | Test hypothesis concerning possible dependent variables. |
| Retroactive design | Begin with subjects grouped on the basis of preexisting dependent variables. | Test hypothesis concerning possible independent variables. |

   a. The **proactive ex post facto research** design begins with subjects grouped on the basis of an independent variable such as father present/father not present or retained/promoted. The researcher then compares these preexisting groups on measurers of dependent variables such as self-confidence, mental health, and academic performance.
   b. **Retroactive ex post facto research** seeks possible antecedent causes (independent variables) for a preexisting dependent variable. Whether students graduate from high school or drop out is a variable that cannot be manipulated. Therefore, a researcher would use retroactive ex post facto research to investigate hypotheses about possible causes, such as truancy, attitude toward school, ambition, and parents' education.

## ALTERNATIVE EXPLANATIONS IN EX POST FACTO RESEARCH

When investigators can control the treatment ($X$) and then observe the dependent variable ($Y$) as in experimental research, they have reasonable evidence that $X$ influences $Y$. Ex post facto research, on the other hand, lacks control of the independent variable and thus has lower internal validity. If researchers cannot control ($X$), they may be led to inappropriate conclusions. When interpreting ex post facto research, one should consider alternative explanations, such as common cause, reverse causality, and the presence of other independent variables.

### Common Cause

In an ex post facto investigation, you must consider the possibility that both the independent variable and the dependent variable of the study are merely two separate results of a third variable—that they have a **common cause.** For example, if you use a school's total budget as an independent variable and cases of diagnosed learning disability as a dependent variable, you might find a positive correlation between the two variables. Does this mean that an increase in total school budget leads to an increase in cases of learning disability? A more plausible explanation is that the relationship is spurious. An increase in school size/number of children attending could account for both the budget and the cases of diagnosed learning disability because funding is tied to the number of students. It is well established that the average income of private high school graduates is much higher than the average income of public and parochial high school graduates. Does this mean that private schools better prepare students for financial success? Or is the difference due to the fact that those families with enough money to send their children to private schools are also able to finance their children's professional training and set them up in business?

When doing ex post facto research, you must always consider the possibilities of common cause or causes accounting for an observed relationship. In our examples, fairly obvious common causes could be identified. However, in ex post facto research there is always a nagging doubt that maybe common causes no one has thought of explain a relationship. Research has shown that the injury rate of drivers who use safety belts is lower than the injury rate of drivers who do not. Is this because using safety belts reduces injury or because cautious drivers have fewer injury-causing accidents and also are more likely to use safety belts?

### Reverse Causality

In interpreting an observed relationship in an ex post facto study, the researcher must consider the possibility of **reverse causality**—that the reverse of the suggested hypothesis could also account for the finding. Instead of saying that *X* causes *Y*, perhaps it is the case that *Y* causes *X*. For instance, the proportion of Episcopalians listed in *Who's Who in America* is much greater than their proportion in the general population. Does this mean Episcopalianism leads to the kind of success that results in being listed in *Who's Who*? It is just as plausible to hypothesize that successful people tend to gravitate to the Episcopal Church. Similarly, if you find that college students who drink have a lower grade point average (GPA) than nondrinkers, you cannot automatically conclude that alcohol consumption depresses academic performance. Perhaps bad grades drive students to drink. (Or, of course, any number of common causes could lead to both drinking and poor grades.)

Investigations on the effects of child-rearing practices have revealed that children who are frequently punished show more aggressive behavior. Can you conclude that parental punishment leads to aggressive children? Or are aggressive children more likely to be punished? A study conducted at the University of Southern California (Hotz, 2002) reported that children who are outgoing and adventurous as toddlers have substantially higher IQs and reading ability by the time they are preteens (age 11 years). The difference in mental ability was true for boys and girls and across variations in ethnic background and income levels.

The researchers correctly concluded that the research does not show whether the children's outgoing curiosity as toddlers was the cause of their better scores on later IQ tests or whether children who have higher IQs are, by nature, more curious and outgoing.

The hypothesis of reverse causality is easier to deal with than the hypothesis of common cause. With the latter, numerous common causes in each case could produce a spurious relationship. With reverse causality, there is only one possibility in each case: $Y$ caused $X$, instead of $X$ caused $Y$.

In a situation in which $X$ always precedes $Y$ in time, the very nature of the data rules out the possibility of reverse causality. For example, numerous studies have shown that the average annual income of college graduates is higher than the average annual income of nongraduates. You can rule out the hypothesis of reverse causality because graduation or nongraduation precedes the subsequent annual income. You cannot rule out a variety of possible common causes.

A method of establishing the time order of variables is to obtain measurements of the same subjects at different times. Let us assume that you are interested in the relationship between employees' acceptance of the philosophy of a corporation and job promotion within that corporation. If you merely interviewed a sample of the employees and found that those in higher positions held attitudes and opinions more in line with the company's value system, you would not know whether acceptance of company values and objectives led to promotion or whether promotion increased acceptance of the company value system. To rule out reverse causality as an explanation, you could interview a group of new trainees and obtain—by means of a questionnaire, rating scale, or the like—a measure of their acceptance of the corporation philosophy. Then after a period of time, perhaps 18 months, you could determine from company records which employees had been promoted. If the findings showed that a significantly higher proportion of employees who had expressed attitudes and opinions consistent with corporation philosophy had been promoted, compared with those who had not, you would have better evidence that conformity with company philosophy was conducive to promotion. (You are still left with the possibility that some common cause or causes account for differences in both philosophy and promotion.)

### Other Possible Independent Variables

Independent variables other than the one considered in the ex post facto study may bring about the observed effect on the $Y$ variable; that is, in addition to $X_1$, other variables, $X_2$ and $X_3$, may also be antecedent factors for the variation in the dependent variable. The recorded suicide rate in Sweden is among the highest in the world. Does this mean the Swedish environment causes more people to commit suicide? Does it mean that the Swedish people are more suicide prone than others? Perhaps there is truth in one or both of these hypotheses. It is equally possible, however, that the actual independent variable is the honesty of coroners in Sweden compared with the honesty of coroners in other countries. In countries where great social stigma falls on the families of those who commit suicide, coroners may well use every conceivable means to record a death as accidental rather than suicide. Therefore, the difference between reported suicide rates may be a function of coroner behavior and nothing else.

Let us consider some more examples. At a governors' conference, governor *X* points with pride to the low crime rate in his state. Another governor points out that the police forces in governor *X*'s state are seriously undermanned and the low crime rate may indicate only that very few crimes there are ever reported. An industrialist asks his personnel manager why he does not hire more Old Tuephingen University graduates, asserting that because so many of them are rapidly moving up the promotion ladder, they are obviously more competent than other graduates. The personnel manager tactfully points out that the phenomenon may not be explained by competence but, rather, by the fact that the industrialist is himself an Old Tuephingen University graduate and may be subconsciously favoring his fellow alumni in promotion decisions.

### THINK ABOUT IT 12.2

1. The chief of police finds that the proportion of African Americans who are arrested for traffic violations is much higher than the proportion of whites in his city. State a possible explanation the chief should consider before concluding that African Americans are more reckless drivers than others.
2. Why is reverse causality not a credible explanation in this case?

**Answers**

1. His police officers expect African Americans to be more reckless and thus give them greater surveillance than other drivers. The "hood" expression is "arrested for DWB" (Driving While Black).
2. It is not possible for traffic violations to cause skin color.

Research published by Donahoo and Eckel (2004) reported that the arrival of spring can bring about a significant drop in cholesterol levels of people suffering from high cholesterol. Can we conclude that spring weather causes cholesterol levels to fall? A more likely explanation is that with the arrival of spring people get outside to walk and engage in other physical activity and that it is the increased physical activity that results in the drop in cholesterol. We can rule out common cause and reverse causality in this case, but it is plausible that another independent variable (exercise) could account for the finding.

An obvious first task for investigators is to attempt to list all the possible alternative independent variables. Then by holding the others constant, you can test in turn each variable to determine if it is related to *Y*. If you can eliminate the alternative independent variables by showing that they are not related to *Y*, you gain support for the original hypothesis of a relationship between *X* and *Y*.

## AN APPLICATION OF ALTERNATIVE EXPLANATIONS

Is there a relationship between having a car and academic achievement among high school students? A high school principal observes that the GPA of automobile owners is noticeably lower than the GPA of other students. He proposes placing restrictions on students' use of cars at school. Is such a conclusion justified?

Let us consider the possible alternative hypotheses.

1. *Common cause.* Are there variables that may influence both auto use and scholarship? A possible common cause is student employment. Having a car may require students to work in order to afford the car. Working results in less time spent studying and thus lower grades. Differences in student lifestyle or values could also account for the apparent relationship. If some students value driving highly and have little interest in scholarship, denying them access to cars would not necessarily increase their scholarship. You could propose a number of credible common cause hypotheses.
2. *Reverse causality.* Is it possible that poor grades are a cause of car use? You could reasonably hypothesize that students who do poorly in school look for other paths to social acceptance and that car use is one possible path.
3. *Other possible independent variables.* Are some teachers assuming that car owners are not interested in their studies and assigning them lower grades than they deserve?

How could this question be investigated? Instead of considering only present grades and auto use, one could record the grades of freshmen and the grades of the same subjects at the end of the first semester during their senior year. If auto use affects scholarship, then the grades of drivers would be expected to drop between the time they were freshmen—and therefore not eligible for driver's licenses—and the time they were seniors. Thus, the dependent variable would be the difference between freshman and senior GPA. One might also co-vary (see Chapter 11) on IQ.

## THINK ABOUT IT 12.3

A supervisor of student teachers notices that those with pierced tongues get lower ratings from their mentor teachers than do other student teachers. The supervisor might hypothesize the following:

a. The same lifestyle that leads to pierced tongues leads to poor student teaching performance.
b. Pierced tongues lead to poor health, which leads to poor student teaching performance.
c. Student teachers who are not doing well decide to have their tongues pierced.
d. Mentor teachers imagine that pierced tongues indicate antisocial attitudes and give student teachers with pierced tongues lower ratings than they deserve.

Which of the previous hypotheses represent the following possibilities?

1. Reverse causality
2. Cause/effect
3. Common cause
4. Alternate independent variable

**Answers**

1. c; 2. b; 3. a; 4. d

## PARTIAL CONTROL IN EX POST FACTO RESEARCH

There are strategies for improving the credibility of ex post facto research, although none can adequately compensate for the inherent weakness of such research—namely, lack of control over the independent variable. These strategies provide **partial control** of the internal validity problems of common cause and other possible independent variables. Among these strategies are matching, homogeneous groups, building extraneous variables into the design, analysis of covariance, and partial correlation.

### MATCHING

A common method of providing partial control in ex post facto investigations is to match the subjects in the experimental and control groups on as many extraneous variables as possible. This **matching** is usually done on a subject-to-subject basis to form matched pairs. For example, if you are interested in the relationship between scouting experiences and delinquency, you could locate two groups of boys classified as delinquent and nondelinquent according to specified criteria. In such a study, it would be wise to select pairs from these groups matched on the basis of socioeconomic status, family structure, and other variables known to be related to both choosing the scouting experience and delinquency. The data from the matched samples could be analyzed to determine whether the proportion of those who participated in scouting is greater among nondelinquents than among delinquents. The matching procedure in ex post facto research presents some of the difficulties described in our discussion of its use in experimentation (see Chapter 10).

First, using matching in an ex post facto study assumes that you know what the relevant factors are—that is, the factors that may have some correlation with the dependent variable. Even if you are able to match on a few relevant variables, you will leave many other variables unmatched, and these unmatched variables may affect the dependent variable. Furthermore, matching is likely to greatly reduce the number of subjects that can actually be used in the final analysis. Probably there are several variables that need to be controlled. As the number of matching variables increases, it becomes increasingly more difficult to find a match. The loss of cases inherent in the matching process is an even more serious problem in ex post facto research than in experimentation, in which matching precedes the measurement of the independent variable.

A more serious problem than loss of subjects is the role of regression in an ex post facto matched-pairs design. Let us illustrate the point. Principal A has introduced a new reading program in the fifth grade, and after it has been in use for a year he wants to compare its effectiveness with the effectiveness of the reading program it replaced. In the same district, principal B's school is still using the old program. Both schools give the same standardized reading test at the end of each school year. Principal A compares the mean grade-equivalent reading scores for fifth-graders in the two schools. He finds the mean grade level equivalent for his fifth-graders is 6.0, whereas the mean for principal B's fifth-graders is 4.0. Is this dramatic evidence of the effectiveness of the new method? Principal A realizes that the difference between means could be caused by differences between

the pupils when they began the fifth grade in the two schools. He obtains scores for the reading test administered when the pupils were finishing fourth grade. He finds that his students had a mean grade level equivalent of 4.8, whereas principal B's students had a mean grade level equivalent of only 3.2. Therefore, he must make an adjustment for the fact that the two groups were not at the same point when they began fifth grade.

Principal A decided to use matching as a way to control for this difference. He created matched pairs by matching a student from his school with a fourth-grade reading score of 3.1 with a student from principal B's school with a fourth-grade reading score of 3.1, a student from his school with a score of 4.8 with a student from principal B's school with a score of 4.8, and so on. However, because the mean of the B population is lower than the mean of the A population, there will be many low-scoring B students for whom there is no match in the A group and many high-scoring A students for whom there is no match in the B group. The scores of all the unmatchable students will be excluded from the data analysis. For those who could be matched, the mean score for the pretreatment fourth-grade reading test for group A and group B will be identical. Therefore, it appears the researchers have successfully created a group from school B that is the same as the group from school A with regard to reading achievement.

This all sounds very good. Can you now attribute differences in fifth-grade reading scores to a difference in the effectiveness of the old and new methods? Alas, no! The matched pairs are basically those students with poorer fourth-grade reading scores from population A and those students with higher scores from population B. The matched A students' scores will regress *up* toward the total A mean, and the matched B students' scores will regress *down* toward the total B mean. Thus, when you compare the fifth-grade reading scores of the matched groups, you would expect the A mean to be higher than the B mean even if the new method is no more effective than the old method.

Matching looks good because it provides experimental and control groups that are equal on a pretreatment variable or variables. However, when two matched groups are drawn from different populations, regression toward the original population means will be expected to create spurious results whenever the two populations are not equal. Matching pairs from within a *single* population is often a useful strategy. Matching subjects from one population with subjects from another population is a *bad* strategy. As is the case with change scores, matching only partly adjusts for preexisting differences between groups, and this underadjustment can be misleading.

## HOMOGENEOUS GROUPS

You may recall from the discussion of control in experimentation that it is possible to control for the effects of a variable by selecting samples that are as homogeneous as possible on that variable. A similar procedure can be followed in ex post facto research. Instead of taking a heterogeneous sample and comparing matched subgroups within it, an investigator may control a variable by including in the sample only subjects who are homogeneous on that variable. If academic aptitude is a relevant extraneous variable, the investigator could control its effect by using subjects from only one academic aptitude level. Or if gender is a variable to be

controlled, you can use only males or only females in the research. This procedure serves the purpose of disentangling the independent variable in which you may be interested from other variables with which it is commonly associated so that any effects you find can more justifiably be associated with the independent variable. Using **homogeneous groups** restricts the generalizability of the findings only to the specific homogeneous group, thus reducing external validity of the study. Lee and Loeb (2000) used homogeneous groups to study the effect of school size on teacher attitudes and student achievement. They avoided any influence that the dimension of rural, suburban, and urban settings might have on the dependent variables by studying only schools in the city of Chicago. They found that in small schools (those with fewer than 400 students), teachers had a more positive attitude about their responsibility for students' learning and students learned more than in medium-sized or large schools.

## BUILDING EXTRANEOUS VARIABLES INTO THE DESIGN

It may be possible to build relevant extraneous independent variables into the ex post facto design and investigate their effect through the use of two-way and higher order analyses of variance. For example, suppose an investigator wants to compare the number of truancies among students who have been through an attendance-promoting program and among students who have not been in such a program. If the investigator thinks that ethnic membership and gender influence the number of truancies, the subjects could be classified as Anglo, Hispanic, Asian American, or black, and also as male or female. An *F* test would first be used to assess the main effect of program versus no program on number of truancies. An *F* test for gender would assess the truancy differences of males and females. The *F* test for ethnicity by program versus no program would assess whether the effect of the program was consistent among the ethnic groups. The *F* test for program by gender would assess the consistency of program effectiveness between genders. The gender by ethnicity *F* test would assess the consistency of gender differences in truancy among ethnic groups. Finally, the $2 \times 4 \times 2$ higher order interaction *F* test would indicate if the treatment by ethnic group interaction was consistent between males and females. Building other variables into an ex post facto design is a partial solution, but you can never be sure that you have *all* the variables that should have been considered.

## ANALYSIS OF COVARIANCE

**Analysis of covariance** (ANCOVA) is sometimes used to partially adjust for preexisting differences between groups in an ex post facto design. Specifically, it adjusts scores on the dependent variable for any initial differences on the extraneous variable. However, because the adjustment is only partial, ANCOVA does not "solve" the problem of initial differences between groups but only reduces it. When interpreting ex post facto research, it is inappropriate to assume ANCOVA has satisfactorily adjusted for initial differences.

A classic example of the problems inherent in matching and ANCOVA is the Cicirelli (1969) ex post facto investigation of the effects of the Head Start program. This study compared the academic achievement of students who had been in the Head Start program with the achievement of those who had not been in the Head Start program. Children who had been in the program were matched with

non-Head Start children from the same neighborhoods on gender, racial/ethnic groups, and kindergarten attendance. ANCOVA was used to adjust for differences in income per capita, educational level of father, and occupational level of father. The results appeared to indicate that the achievement of the non-Head Start group was greater than that of the Head Start group even when scores were adjusted for initial differences. The authors concluded that the Head Start program was harmful. Researchers have pointed out that because of regression, both matching and ANCOVA underadjust for initial differences between groups. One would therefore expect adjusted posttreatment scores of a disadvantaged group to be less than adjusted posttreatment scores of a less disadvantaged group. Because the extent of the underadjustment is unknown, the Cicirelli study does not determine whether the Head Start experience was harmful or beneficial or had no effect.

## THINK ABOUT IT 12.4

Read the following abstract from a report of a study using ex post facto research that was published in 2000 and answer the following questions:

# *Academic Achievement and Parental School Involvement as a Function of High School Size*

**Pamela W. Gardner**
**Shulamit N. Ritblatt**
**James R. Beatty**
**San Diego State University**

The purpose of this research was to examine academic achievement, absenteeism, dropout rate, and parental school involvement as a function of high school size. Previous studies indicated that, generally, students of small schools exhibited greater extracurricular participation, greater satisfaction, lower absenteeism, and lower dropout rate. In terms of academic achievement, findings have been mixed.

Academic achievement was measured by scores on the Scholastic Aptitude Test (SAT), while parent participation was measured by membership in the Parent–Teacher Association (PTA) or other parent organization(s). Absenteeism and dropout rate were measured using data provided by the California Department of Education.

Sixty-seven randomly selected large California public high schools (enrollments over 2000) were compared to 60 randomly selected small California public high school (enrollments between 200 and 600 students). The t-tests and analyses of covariance were employed to determine the differences between the large and small high schools. Contrary to several hypotheses, the large schools exhibited higher academic achievement on total SAT score, verbal SAT score, and math SAT score. In addition, this cohort had a higher proportion of students taking the SAT. Confirming other hypotheses, the small school cohort displayed lower absenteeism, lower dropout rate, and higher parental school involvement.

The results of this investigation indicate the need for further research, both in size and scope. In addition, the findings carry implications for educational policy decisions. Standard procedures for reporting absenteeism and parental school involvement should be established.

*Source:* Gardner, P. W., Ritblatt, S. N., & Beatty, J. R. (2000). Academic achievement and parental school involvement as a function of high school size. *High School Journal, 83*(2), 21–27.

**Questions**

1. What is the independent variable?
2. What are the dependent variables?
3. How are the dependent variables operationally defined?
4. Is this study proactive or retroactive?
5. What is the population of interest?
6. How was the sample selected?
7. Is reverse causality credible?
8. Is common cause credible?
9. Which partial control strategy was used?

**Answers**

1. High school size
2. a. Academic achievement
   b. Parental participation
   c. Absenteeism
   d. Dropout rate
3. a. Scholastic Aptitude Test scores
   b. Membership in PTA or other parent organization(s)
   c. Data from the California Department of Education
4. Proactive
5. Students and parents of California high schools
6. Random selection of large (over 2000 students) and small (200 to 600 students) California high schools
7. No
8. No
9. Analyses of covariance

## THE ROLE OF EX POST FACTO RESEARCH

Given the hazards involved in ex post facto research, many educational researchers believe they should not engage in this type of research at all. Basically, they contend that it is better to admit ignorance than to risk reaching incorrect conclusions. However, others point out that many variables of great interest are not amenable to experimental research. Researchers cannot randomly assign children to broken or intact homes, to high or low social class, to achievement-oriented or non-achievement-oriented peer groups, to high or low self-esteem groups, and so forth. Therefore, if they want to learn anything about relationships between such attribute variables and other variables, the ex post facto method is their only recourse. An ex post facto study is better than no study at all. If researchers use appropriate methods of partial control and consider alternative hypotheses, perhaps they can be correct more often than wrong.

Certainly, there have been many highly credible ex post facto studies. The U.S. Surgeon General's 1964 study of the relationship between smoking and lung cancer is a well-known example. It is not possible to randomly designate

a group of human subjects who are to smoke for years and a group who are not to smoke, so the study had to be done as an ex post facto investigation. The reversed-causality hypothesis that lung cancer causes people to smoke is not plausible. None of the common-cause hypotheses offered seem very likely: Nervous people are prone to both smoking and lung cancer, some genetic predisposition leads to both, and so forth. The Surgeon General controlled for many alternative independent variables—for example, by analyzing separately samples from areas of high air pollution and low air pollution. Experimental results from studies of animals that were made to inhale or not inhale cigarette smoke produced evidence of a cause–effect relationship. Given all this, despite the dangers inherent in ex post facto research, most would conclude that it is better to bet there *is* a cause-to-effect relationship between smoking and lung cancer among humans than to bet there is *not* such a relationship. In fact, much medical research is ex post facto in design. Researchers study diseases and try to determine the influence of factors such as eating habits, lifestyle, and genes. Ex post facto research in education has permitted investigations of the effects of variables such as home background, father absence, early experiences, disabilities, teacher competence, and others that are beyond the control of educators. In some instances, ex post facto research has discovered relationships or raised questions that can later be investigated more systematically in well-controlled experimental studies. Appropriately used and cautiously interpreted, ex post facto research will continue to provide a valuable methodology for the acquisition of knowledge.

## SUMMARY

Ex post facto research is used to investigate relationships when the researcher cannot randomly assign subjects to different conditions or directly manipulate the independent variable. Ex post facto research begins with subjects who differ on an observed dependent variable and tries to determine the antecedents (cause) of the difference. Or the researcher begins with subjects who differ on an independent variable and tries to determine the consequences of the difference.

Although there are many disadvantages of ex post facto design, it nevertheless is frequently the only method by which educational researchers can obtain necessary information about characteristics of defined groups of students or information needed for the intelligent formulation of programs in the school. It permits researchers to investigate situations in which controlled variation is impossible to introduce. Attributes such as academic aptitude, creativity, self-esteem, socioeconomic status, and teacher personality cannot be manipulated and hence must be investigated through ex post facto research rather than through the more rigorous experimental approach.

The possibility of spurious relationships is always present in ex post facto research. Considering the possibilities of common cause, reversed causality, and possible alternate independent variables can help educators evaluate such research more realistically. Several partial control strategies can help researchers avoid gross errors in ex post facto designs, but none can entirely solve the problems inherent in those designs. Always exercise caution when interpreting ex post facto results.

## KEY CONCEPTS

active independent variable
analysis of covariance (ANCOVA)
attribute independent variable
causal relationship
common cause
ex post facto research
extraneous variables
homogeneous group
matching
partial control
post hoc fallacy
proactive ex post facto research
retroactive ex post facto research
reverse causality
spurious relationship

## EXERCISES

1. What are the two major limitations of ex post facto or causal-comparative designs?
2. Classify the following independent variables as active or attribute:
   - **a.** Kindergarten entrance age
   - **b.** Method of teaching reading
   - **c.** Montessori school attendance versus attendance at public school
   - **d.** Teaching experience
   - **e.** Use of drug A in treating diabetes versus drug B
   - **f.** Learning style
3. What method would you use to investigate the following research hypotheses?
   - **a.** High school students who work during the school year will have lower grades and lower standardized test scores than students who do not work.
   - **b.** First-grade students who are taught reading using phonics will read at a higher level than first-grade students taught by another method.
   - **c.** Beginning teachers who are mentored will be more effective and express more satisfaction with teaching than beginning teachers who are not mentored.
   - **d.** Students who are in small classes (less than 15) in grades kindergarten to 3 will perform at a higher level in grades 4 to 6 than will students in regular-sized classes (more than 20).
4. Interpret this statement: "A statistical relationship between two variables is a necessary but not a sufficient condition for inferring a causal relationship between the variables."
5. Show the design you would use to test the hypothesis stated in Exercise 3(a).
6. What alternative hypotheses would you want to explore if your causal-comparative research indicates that there is a relationship between self-concept and achievement in school?
7. What steps would you take to avoid the post hoc fallacy in your ex post facto research?
8. For the following examples, suggest an alternative explanation for the finding:
   - **a.** A researcher finds a relationship between use of alcohol and unemployment and concludes that using alcohol causes workers to lose their jobs.
   - **b.** A researcher finds that the proportion of the population confined to mental hospitals for manic-depression is greater in the United States than it is in England. He concludes that the American environment is more conducive to the development of depression.
   - **c.** A researcher finds that the increase in the use of cell phones in recent years has been accompanied by an increase in the incidence of brain tumors. He concludes that using a cell phone can result in a brain tumor.
9. Find an example of an ex post facto study in the current literature. Identify (a) the problem, (b) the operational definitions of the variables, (c) the design used, (d) the steps taken to control extraneous variables, (e) the findings, and (f) the conclusions. Did the author include a caveat about the conclusion?
10. A recent study reported that of 4500 women who underwent dental X-rays during pregnancy, 1117 had low-birth-weight babies. The researchers concluded that women who have dental X-rays during pregnancy face an increased risk of having underweight babies. What questions might you want to

ask about this research before accepting their conclusion?

11. The following are examples of published studies. Which do you think involve post facto research methodology?
    a. Do After-School Programs Help Students Succeed?
    b. The Impact of Hope and Social Activity on Academic Performance of Midwestern College Students
    c. The Importance of Middle School Math on High School Mathematics Achievement
    d. The Effect of Using Classroom Lap Top Computers on the Writing of Middle School Students
12. Answer the following questions about the Gardner, Ritblatt, and Beatty (2000) study.
    a. What is the independent variable?
    b. What are the dependent variables?
    c. How are the dependent variables operationally defined?
    d. What is the population of interest?
    e. How was the sample selected?
    f. Is reverse causality credible?
    g. Is common cause credible?

## ANSWERS

1. The two major limitations are lack of randomization and lack of manipulation of the independent variable.
2. a. Attribute
   b. Active
   c. Attribute
   d. Attribute
   e. Active
   f. Attribute
3. a. Ex post facto
   b. Experimental
   c. Experimental
   d. Ex post facto (it is generally not feasible to manipulate class size, although it has been done)
4. To report that one variable ($X$) is the cause of another variable ($Y$), the researcher must first show that there is a statistical relationship between the variables. However, this finding by itself is not sufficient to establish a causal relationship. The researcher must examine the time ($X$ must precede $Y$) and must show that no other variables could cause $Y$.
5. Use the proactive design. That is, you would select two groups of high school students, one of whom works a certain number of hours each week ($X$), and would compare their grades and scores on standardized tests ($Y$) with those of a comparable group of students who do not work. Or, instead of studying the effect of working or not working, you could study the effect of the extent of work as measured by the number of hours worked each week. One group could work less than 10 hours a week and the other group could work 20 or more hours. You would have to control a number of variables, including aptitude scores, previous grades, socioeconomic status, and gender. If you used only one class level (only juniors or only seniors), you could control for the variable of class standing (if relevant).
6. A more positive self-concept could lead to higher achievement in school. Or the reverse causality hypothesis could account for the finding. That is, higher achievement in school could lead to a more positive self-concept. You would also want to explore a common-cause hypothesis. Perhaps another variable, such as intelligence, causes both achievement and self-concept.
7. Try to control any relevant extraneous variables that could affect the measured variable. Methods of control include (a) matching, (b) using homogeneous groups, (c) building extraneous variables into the design, and (d) using ANCOVA or other statistical means to partially control unwanted variables.
8. a. Perhaps unemployment leads to the use of alcohol. Or perhaps profound depression causes people to lose their jobs and also take to drink.
   b. Perhaps depressed people in England receive a different type of treatment and are less likely to be hospitalized than people in the United States. Perhaps English psychologists are less inclined to make a diagnosis of depression.

**c.** Perhaps there is a common cause. The stress and rapid pace of contemporary life are conducive to the use of cell phones and may also influence the development of brain tumors.

**9.** Answers will vary.

**10.** One would first want to ask why the women received X-rays because, normally, dentists are cautious about using X-rays during pregnancy. Perhaps there was some disease or infection that necessitated the use of X-rays, and it was the disease or infection that caused the low-birth-weight babies. The study's lead author stated that more research is needed to determine if dental X-rays really were the culprit.

**11.** b and c are ex post facto research.

## REFERENCES

Cicirelli, V. (1969). *The impact of Head Start: An evaluation of the effects of Head Start on children's cognitive and affective development* (report presented to the Office of Economic Opportunity pursuant to contract B89-4536). Westinghouse Learning Corporation and Ohio University. (Distributed by the Clearinghouse for Federal Scientific and Technical Information, U.S. Department of Commerce, National Bureau of Standards, Institute for Applied Technology, PB 184 328.)

Donahoo, W. T., & Eckel, R. H. (2004). Physiological mechanisms for the seasonal changes in serum cholesterol level. *Archives of Internal Medicine*, *164*(22), 2506–2507.

Gardner, P. W., Ritblatt, S. N., & Beatty, J. R. (2000). Academic achievement and parental school involvement as a function of high school size. *High School Journal*, *83*(2), 21–27.

Hotz, R. (2002, April 15). Outgoing toddlers have higher IQs as preteens. *Indianapolis Star*, p. 1.

Lee, V. E., & Loeb, S. (2000). School size in Chicago elementary schools: Effects on teachers' attitudes and students' achievement. *American Educational Research Journal*, *37*, 1–31.

CHAPTER 13

# Correlational Research

Correlation reveals relationship but not necessarily causation.

## INSTRUCTIONAL OBJECTIVES

After studying this chapter, the student will be able to:

1. Describe the nature of correlational research.
2. Describe the ways correlational research is used.
3. Describe the design of correlational research.
4. Discuss the limitations of correlational research.
5. Distinguish between correlational and ex post facto research.
6. List different types of correlational coefficients and state the conditions for their appropriate use.
7. Interpret correlation coefficients in terms of sign, magnitude, statistical significance, and practical significance.
8. Test a hypothesis about a correlation coefficient, *r*.
9. Define how large a random sample is needed to reject a null hypothesis for a given population correlation.
10. Define predictor and criterion.
11. Define multiple regression and explain when it is used.
12. Define discriminant analysis and explain its purpose.
13. Define factor analysis and explain its purpose.
14. Distinguish exploratory factor analysis and confirmatory factor analysis.
15. Define partial correlation and explain its purpose.
16. Identify studies in which *canonical correlation*, *path analysis*, or *structural equation modeling* would be appropriate.

**Correlational research** is nonexperimental research that is similar to ex post facto research in that they both employ data derived from preexisting variables. There is no manipulation of the variables in either type of research. They differ in that in ex post facto research, selected variables are used to make comparisons between two or more existing groups, whereas correlational research assesses the relationships among two or more variables in a single group. Ex post facto research investigates possible cause-and-effect relationships;

correlational research typically does not. An advantage of correlational research is that it provides information about the strength of relationships between variables. An ex post facto researcher might define those who make more than $200,000 per year as high earners and those who make less than $40,000 per year as low earners and then compare the mean percent of income paid in taxes for each group. The researcher's data show that the average percent of income paid in taxes by the low earners, at 19 percent, is greater than the 10 percent paid by the high earners. The conclusion is that low-income earners pay a higher percent of their income in taxes than do high-income earners. A correlational researcher would record the income and the percent of income paid in taxes for all people in the study. This researcher might report a correlation coefficient of −.6, indicating a strong negative correlation between the two variables.

Correlational research produces indexes that show both the direction and the strength of relationships among variables, taking into account the entire range of these variables. This index is called a **correlation coefficient**. Recall from Chapter 6 that in interpreting a **coefficient of correlation**, one looks at both its sign and its size. The sign (+ or −) of the coefficient indicates the direction of the relationship. If the coefficient has a positive sign, this means that as one variable increases, the other also increases. For example, the correlation between height and weight is positive because tall people tend to be heavier and short people lighter. A negative coefficient indicates that as one variable increases, the other decreases. The correlation between outdoor air temperature during the winter months and heating bills is negative; as temperature decreases, heating bills rise.

The size of the correlation coefficient indicates the strength of the relationship between the variables. The coefficient can range in value from +1.00 (indicating a perfect positive relationship) through 0 (indicating no relationship) to −1.00 (indicating a perfect negative relationship). A perfect positive relationship means that for every *z*-score unit increase in one variable there is an identical *z*-score unit increase in the other. A perfect negative relationship indicates that for every unit increase in one variable there is an identical unit decrease in the other. Few variables ever show perfect correlation, especially in relating human characteristics.

## THINK ABOUT IT 13.1

Interpret each of the following:

1. The correlation between time spent watching television and time spent reading is −.44.
2. The correlation between socioeconomic status and number of museums visited is .21.
3. The correlation between days absent from school and kindergarten reading scores is −.58.

**Answers**

1. The more time spent watching TV, the less time spent reading; there is a negative relationship.
2. The higher the socioeconomic status, the more museums visited; there is a positive relationship.
3. The more days absent, the lower the reading scores; there is a negative relationship.

# USES OF CORRELATIONAL RESEARCH

Correlational research is useful in a wide variety of studies. The most useful applications of correlation are (1) assessing relationships, (2) assessing consistency, and (3) prediction.

## ASSESSING RELATIONSHIPS

Correlational research methods are used to assess relationships and patterns of relationship among variables in a single group of subjects. For instance, correlational research is used to answer questions such as the following: Is there a relationship between math aptitude and achievement in computer science? What is the direction and strength of this relationship, if any? You would most likely predict that a positive relationship would be found between scores on a math aptitude test and grades in computer science. A correlational study would determine the extent of any relationship between these variables.

The following are additional examples of questions that could be investigated in a correlational study: What is the relationship between self-esteem and academic achievement? Is there a relationship between musical aptitude and mathematics achievement among 6-year-olds? and What is the relationship between watching media violence and aggression in children?

In some correlational studies, the researcher may be able to state a hypothesis about the expected relationship. For example, from phenomenological theory you might hypothesize that there is a positive relationship between first-grade children's perceptions of themselves and their achievement in reading. In other instances, the researcher may lack the information necessary to state a hypothesis.

Recall from Chapter 9 that the correlation between test scores and selected external variables is a widely used source of evidence in validity studies.

## ASSESSING CONSISTENCY

In Chapter 9, we noted that the reliability (consistency) of a test can be assessed through correlating test–retest, equivalent-forms, or split-half scores. Correlation can be used to measure consistency (or lack thereof) in a wide variety of cases. For example, how consistent are the independently assigned merit ratings given by the principal and the assistant principal to teachers in a school? How much agreement is there among Olympic judges rating the performance of a group of gymnasts? When a researcher asks a group of teachers to rank the severity of disruption created by each item on a list of behavior disorders, to what extent do their rankings agree?

## PREDICTION

If you find that two variables are correlated, then you can use one variable to predict the other. The higher the correlation, the more accurate the prediction. Prediction studies are frequently used in education. For example, correlational research has shown that high school grades and scholastic aptitude measures are related to college grade point average (GPA). If a student scores high on aptitude tests and has high grades in high school, he or she is more likely to make high grades in college than is a student who scores low on the two predictor variables. Researchers can predict with a certain degree of accuracy a student's

probable freshman GPA based on high school grades and aptitude test scores. This prediction will not hold for every case because other factors, such as motivation, initiative, or study habits, are not considered. However, in general, the prediction is good enough to be useful to college admissions officers.

### THINK ABOUT IT 13.2

Criticize the conclusions reached in the following examples:

1. The correlation between two variables in an investigation turned out to be negative. The researcher reported that there was no relationship between the variables.
2. A scatterplot showed that the points were all close to a straight line. The researcher concluded that this indicated a positive correlation between the variables.

**Answers**

1. There is a relationship: The negative correlation means that high scores on one variable are associated with low scores on the other. As one variable increases, the other decreases.
2. It could also indicate a negative correlation. It depends on the direction of the straight line.

## DESIGN OF CORRELATIONAL STUDIES

The basic design for correlational research is straightforward. First, the researcher specifies the problem by asking a question about the relationship between the variables of interest. The variables selected for investigation are generally based on a theory, previous research, or the researcher's observations. Because of the potential for spurious results, we do not recommend the "shotgun" approach in which one correlates a number of variables just to see what might show up. The population of interest is also identified at this time. In simple correlational studies, the researcher focuses on gathering data on two (or more) measures from a single group of subjects. For example, you might correlate vocabulary and reading comprehension scores for a group of middle school students. Occasionally, correlational studies investigate relationships between scores on one measure for logically paired groups such as twins, siblings, or husbands and wives. For instance, a researcher might want to study the correlation between the SAT scores of identical twins.

The following is an example of a typical correlational research question: What is the relationship between quantitative ability and achievement in science among high school students? The researcher determines how the constructs, ability and achievement, will be quantified. He or she may already be aware of well-accepted operational definitions of the constructs, may seek definitions in sources such as those described in Chapter 4, or may develop his or her own operational definitions and then assess their reliability and validity. In the example, the researcher may decide that quantitative ability will be defined as scores on the School and College Ability Test, Series III (SCAT III), and science achievement will be defined as scores on the science sections of the Sequential Tests of Educational Progress (STEP III).

You learned in Chapters 8 and 9 that it is important to select or develop measures that are appropriate indicators of the constructs to be investigated, and that it is especially important that these instruments have satisfactory reliability

and are valid for measuring the constructs under consideration. In correlation research, the size of a coefficient of correlation is influenced by the adequacy of the measuring instruments for their intended purpose. Instruments that are too easy or too difficult for the participants in a study would not discriminate among them and would result in a smaller correlation coefficient than instruments with appropriate difficulty levels. Studies using instruments with low reliability and questionable validity are unlikely to produce useful results.

Following the selection or development of instruments, the researcher specifies his or her population of interest and draws a random sample from that population. Finally, the researcher collects the quantitative data on the two or more variables for each of the students in the sample and then calculates the coefficient(s) of correlation between the paired scores. Before calculating the coefficient, the researcher should look at a scatterplot or a graph of the relationship between the variables.

## CORRELATION COEFFICIENTS

There are many different kinds of correlation coefficients. The researcher chooses the appropriate statistical procedure primarily on the basis of (1) the scale of measurement of the measures used and (2) the number of variables.

### PEARSON PRODUCT MOMENT COEFFICIENT OF CORRELATION

In Chapter 6, we introduced you to the Pearson product moment correlation coefficient, symbolized *r*, which is the most widely used descriptive statistic of correlation. Recall that the Pearson coefficient is appropriate for use when the variables to be correlated are normally distributed and measured on an interval or ratio scale. We briefly mention some of the other indexes of correlation without going into their computation. Interested students should consult statistics books for the computational procedures.

### COEFFICIENT OF DETERMINATION

Unsophisticated consumers of research often assume that a correlation indicates percentage of relationship, for example, that an *r* of .60 means the two variables are 60 percent related. In fact, *r* is the mean *z*-score product for the two variables, not a percentage. The absolute size of the correlation coefficient (how far it is from zero) indicates the strength of the relationship. Thus, a correlation of −.4 indicates a stronger relationship than a +.2 because it is further from zero. The sign has nothing to do with the strength of the relationship. Another way to see how closely two variables are related is to square the correlation coefficient. When you square the Pearson *r*, you get an index called the **coefficient of determination**, $r^2$, which tells you how much of the variance of *Y* is in common with the variance of *X*. A correlation of +.60 or −.60 means that the two variables have ($.60^2$) or 36 percent of their variance in common with each other. If the two variables were caffeine and reaction time, then the amount of caffeine one has consumed would be associated with 36 percent of the variance in one's reaction time. That leaves 64 percent of the variance in reaction time associated with factors other than variation in caffeine intake. The notion of common variance is illustrated in

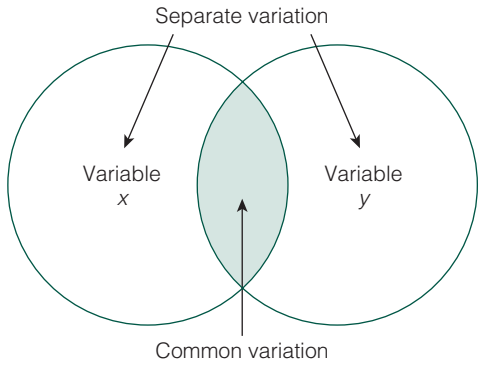



**Figure 13.1** The Common Variance between Two Variables

Figure 13.1, in which the total amount of variation in each variable is represented by a circle. The overlap of the circles represents the common variance.

An increase in the $r$ results in an accelerating increase in $r^2$. A correlation of .20 yields a coefficient of determination of .04. An $r$ of .4 yields an $r^2$ of .16. An $r$ of .8 yields an $r^2$ of .64, and so on. The coefficient of determination is a useful index for evaluating the meaning of size of a correlation. It also reminds one that positive and negative correlations of the same magnitude, for example, $r = .5$ and $r = -.5$, are equally useful for prediction and other uses because both have the same coefficient of determination, $r^2 = .25$. The coefficient of determination ranges from 0 to +1.00. If it is 1.00 ($r = +1.00$), you can predict individuals' scores on one variable perfectly from their scores on the other variable.

The Pearson $r$ and $r^2$ are only appropriate where the relationship between $X$ and $Y$ is *linear*. Linear means that a straight line is a good fit for showing the tilt of the cloud of data in a scattergram. In Chapter 6, several scattergrams are shown. All of them are linear except Figure 6.13. Look at those figures and note the difference between this figure and the rest.

Fortunately, most correlations found in the behavioral sciences are linear. However, before you proceed to calculate and interpret a Pearson $r$ for your data, have your computer print out a scatter diagram for your data. If the relationship is not linear, the Pearson $r$ is not appropriate for assessing the relationship between variables.

## SPEARMAN RHO COEFFICIENT OF CORRELATION

**Spearman rho (ρ),** an ordinal coefficient of correlation, is used when the data are ranks. For example, assume the principal and assistant principal have independently ranked the 15 teachers in their school from first, most effective, to fifteenth, least effective, and you want to assess how much their ranks agree. You can calculate the Spearman's rho by putting the paired ranks into the Pearson $r$ formula or by using a formula developed specifically for rho that is simpler that the Pearson $r$ if you are calculating "by hand."

Spearman rho is interpreted in the same way as the Pearson $r$. Like the Pearson product moment coefficient of correlation, it ranges from −1.00 to +1.00. When each individual has the same rank on both variables, the rho correlation will be +1.00, and when their ranks on one variable are exactly the opposite of

their ranks on the other variable, rho will be −1.00. If there is no relationship between the rankings, the rank correlation coefficient will be 0.

### THE PHI COEFFICIENT

The **phi (φ) coefficient** is used when both variables are genuine dichotomies scored 1 or 0. For example, phi would be used to describe the relationship between the gender of high school students and whether they are counseled to take college preparatory courses or not. Gender is dichotomized as male = 0, female = 1. Being counseled to take college preparatory courses is scored 1, and not being so counseled is scored 0. It is possible to enter the pairs of dichotomous scores (1's and 0's) into a program that computes Pearson *r*'s and arrive at the phi coefficient.

If you find the phi coefficient in school A is −.15, it indicates that there is a slight tendency to counsel more boys than girls to take college preparatory courses. If in school B the phi coefficient is −.51, it indicates a strong tendency in the same direction. As with the other correlations, the phi coefficient indicates both direction and strength of relationships.

A variety of correlation coefficients are available for use with ordinal and nominal data. These include coefficients for data that are more than just pairs; for example, assessing the agreement of three or more judges ranking the performance of the same subjects. We highly recommend Siegel and Castellan's *Nonparametric Statistics* (1988). We consider it a remarkably well organized and easy to understand text.

## CONSIDERATIONS FOR INTERPRETING A CORRELATION COEFFICIENT

The coefficient of correlation may be simple to calculate, but it can be tricky to interpret. It is probably one of the most misinterpreted and/or overinterpreted statistics available to researchers. Various considerations need to be taken into account when evaluating the practical utility of a correlation. The importance of the numerical value of a particular correlation may be evaluated in four ways: (1) considering the nature of its population and the shape of its distribution, (2) its relation to other correlations of the same or similar variables, (3) according to its absolute size and its predictive validity, or (4) in terms of its statistical significance.

### THE NATURE OF THE POPULATION AND THE SHAPE OF ITS DISTRIBUTION

The value of an observed correlation is influenced by the characteristics of the population in which it is observed. For example, a mathematics aptitude test that has a high correlation with subsequent math grades in a regular class where students range widely on both variables would have a low correlation in a gifted class. This is because the math aptitude scores in the gifted class are range restricted (less spread out) compared to those in a regular class. Range restrictions of either the predictor or the criterion scores reduce the strength of the observed correlation. Before proceeding to interpret your correlation results, produce a scattergram to determine if you have range restriction problem.

Also, if your population differs from the population in which a correlation was reported, that correlation only provides an estimate of correlation in your population of interest. The more your population differs from the original population, the less useful the estimate becomes.

In planning a correlational study, if you think variables such as home language or gender will influence your correlation of interest, you can draw random samples of equal numbers from each subgroup to assess the influence of these variables.

## COMPARISON TO OTHER CORRELATIONS

A useful correlation is one that is higher (in either direction) than other correlations of the same or similar variables. For example, an $r$ of .75 would be considered low for the relationship between the results of two equivalent forms of an achievement test because equivalent forms of most achievement tests correlate with each other by more than .90. A correlation of .80 between a measure of academic aptitude and GPA of middle school students would be considered high because the correlation for other measures of academic aptitude and GPA for this population is typically approximately .70.

As we have previously stated, a measure that can be used with high school seniors that correlates .60 with their subsequent college freshman GPA would be excellent because currently available measures correlate between .40 and .45 with college GPAs.

## PRACTICAL UTILITY

Always consider the practical significance of the correlation coefficient. Although a correlation coefficient may be statistically significant, it may have little practical utility. With a sample of 1000, a very small coefficient such as .08 would be statistically significant at the .01 level. But of what practical importance would this correlation be? Information on $X$ only accounts for less than 1 percent ($.08^2 = .0064$ or 00.64 percent) of the variance in $Y$ ($r^2$). In this case, it would hardly be worth the bother of collecting scores on a predictor variable, $X$, to predict another variable, $Y$. You want to avoid the **significance fallacy**—the assumption that a statistically significant correlation also has practical significance. Statistical significance alone is not sufficient. How worthwhile a correlation may be is partly a function of its predictive utility in relation to the cost of obtaining predictor data. A **predictor** with a high correlation that is difficult and expensive to obtain may be of less practical value than a cheap and easy predictor with a lower correlation. Also, note that a correlation coefficient only describes the degree of relationship between given operational definitions of predictor and predicted variables in a particular research situation for a given sample of subjects. It can easily change in value if the same variables are measured and correlated using different operational definitions and/or a different sample.

Failure to find a statistically significant relationship between two variables in one study does not necessarily mean there is *no* relationship between the variables. It only means that in that particular study, sufficient evidence for a relationship was not found. Recall from Chapter 6 that other factors, such as reliability of the measures used and range of possible values on the measures, influence the size of a correlation coefficient.

## STATISTICAL SIGNIFICANCE

In evaluating the size of a correlation, it is important to consider the size of the sample on which the correlation is based. Without knowing the sample size, you do not know if the correlation could easily have occurred merely as a result of chance or is likely to be an indication of a genuine relationship. If there were fewer than 20 cases in the sample (which we would not recommend), then a "high" $r$ of .50 could easily occur by chance. You should be very careful in attaching too much importance to large correlations when small sample sizes are involved; an $r$ found in a small sample does not necessarily mean that a correlation exists in the population.

To avoid the error of inferring a relationship in the population that does not really exist, the researcher should state the null hypothesis that the population correlation equals 0 ($H_0$: $\rho_{xy} = 0$) and then determine whether the obtained sample correlation departs sufficiently from 0 to justify the rejection of the null hypothesis. In Chapter 7, we showed you how to use Table A.3 in the Appendix, which lists critical values of $r$ for different numbers of degrees of freedom ($df$). By comparing the obtained $r$ with the critical values of $r$ listed in the table, you can determine the statistical significance of a product moment correlation. For example, assume a correlational study involving the paired math and spelling test scores of 92 students yields a correlation of .45.

Recall that for the Pearson $r$ the degrees of freedom are the number of paired scores minus 2 ($n - 2$). In Table A.3, we find that for 90 degrees of freedom, $r = .2050$ or greater is statistically significant at the .05 level of significance, .2673 or greater is statistically significant at the .01 level, and .3375 or greater is statistically significant at the .001 level (all two tailed). Therefore, the hypothesis that the population correlation is zero can be rejected at the .01 level and even at the .001 level; therefore, you conclude that there is a positive relationship between math and spelling scores.

## DETERMINING SAMPLE SIZE

The Pearson product moment correlation is a form of effect size. Therefore, Table A.3 in the Appendix can be used to determine the needed sample size for a predetermined level of significance and predetermined tolerable probability of Type I error.

For example, a researcher developed a measure of how much a person is willing to sacrifice to achieve success and found it had very satisfactory reliability when administered to high school seniors. He thinks it may be a useful predictor of success in college. Since previous research has shown that the predictor variables high school GPA, ACT test scores, and CEEB test scores of high school seniors all correlate around .40 with the criterion variable college freshmen GPAs, the researcher decides that if his scores correlate .40 or higher with college GPAs it is worth further investigation. If it is less than .40, it is not worth pursuing. He sets his desired level of significance at the two-tailed .01 level. You see in Table A.3 that if the true population correlation is .3932 or greater with 40 degrees of freedom, then $40 + 2$ subjects randomly selected from that population are needed to reject the null hypotheses that the population correlation is zero.

Recall from Chapter 7 that the larger the sample, the more likely the sample statistics are to approximate the population parameters. Note that this is true only when generalizing results from a random sample back to the population from which it was drawn. If the researcher drew the sample from high school seniors in Peoria, Illinois, he could only directly apply results to Peoria, Illinois, seniors. The usefulness of the result for predicting scores for a different population depends on how similar that population is to the Peoria senior population.

Before disseminating the results of this study, the researcher should calculate the scores on his sacrifice-for-success test with high school GPA and ACT and CEEB scores. If any or all of these scores correlate highly with the sacrifice-for-success measure, it is largely repeating information already known. Therefore, it is not adding enough to the prediction of college GPA to be worthwhile. If the correlations are low, the sacrifice-for-success scores would be useful for increasing the predictive validity of the combined weighted scores currently in use.

### THINK ABOUT IT 13.3

If this population Pearson product moment correlation is .40 or greater, how large should a randomly selected sample be to reject the null hypothesis that the population correlation is zero with a two-tailed test at the .001 level?

**Answer**

$70 + 2 = 72$

## CORRELATION AND CAUSATION

In evaluating a correlational study, one of the most frequent errors is to interpret a correlation as indicating a **cause-and-effect relationship**. Correlation is a necessary but never a sufficient condition for causation. For example, if a significant positive correlation is found between the number of hours of television watched per week and above average body weight among middle school pupils, that does not prove that excessive television watching causes obesity. Recall from Chapter 12 that when the independent variable is not under the investigator's control, alternate explanations must be considered. In this example, reverse causality is plausible. Perhaps the more overweight a child is, the more he or she is inclined to choose television watching instead of physical activities, games, and interacting with peers. The common-cause explanation is also plausible. Perhaps differences in family recreational patterns and lifestyle account for both differences in weight and time spent watching television.

Consider another example. Assume a researcher finds a relationship between measures of self-esteem and academic achievement (grades) for a sample of students. Table 13.1 summarizes the possibilities for interpreting this observed relationship. Any number of factors could act together to lead to both self-esteem and academic achievement: previous academic experiences, parents' education, peer relationships, motivation, and so on.

**Table 13.1** Possible Interpretations of a Relationship between Self-esteem and Academic Achievement

| | |
|---|---|
| Self-esteem → | Achievement (self-esteem causes achievement) |
| Achievement → | Self-esteem (achievement causes self-esteem) |
| Intelligence → | Self-esteem (a third factor causes both) |
| | Achievement |
| Home environment → | Self-esteem (a third factor causes both) |
| | Achievement |

**Table 13.2** Possible Interpretations of a Relationship between Children's Watching Television Violence and Aggression

| | |
|---|---|
| Watching TV violence → | Aggression in children |
| Aggressive children → | Choose to watch violent TV shows |
| Other factors such as home environment → | Both children's watching TV violence and aggression |

Let us consider the example of the relationship between the amount of violence children watch on television and their aggression. Most research has shown a relationship between these two variables, which many people assume is causal. However, Table 13.2 shows other explanations for this relationship.

We must stress, however, that correlation can bring evidence to bear for cause and effect. The Surgeon General's warning about the dangers of cigarette smoking is, in part, based on studies that found positive correlations between the number of cigarettes smoked per day and incidence of lung cancer and other maladies. Here, reverse causality (cancer leads to cigarette smoking) is not a credible explanation. Various common-cause hypotheses (e.g., people who live in areas with high air pollution smoke more and have higher cancer rates) have been shown not to be the case.

Although correlational research does not permit one to infer causality, it may generate causal hypotheses that can be investigated through experimental research methods. For example, finding the correlation between smoking and lung cancer led to animal experiments that allowed scientists to infer a causal link between smoking and lung cancer. Because the results of correlational studies on humans agree with the results of experimental studies on animals, the Surgeon General's warning is considered well founded.

## PARTIAL CORRELATION

The correlation techniques discussed so far are appropriate for examining the relationship between two variables. In most situations, however, a researcher must deal with more than two variables, and we need procedures that examine the relationship among several variables. **Partial correlation** is a technique used to determine what correlation remains between two variables when the effect of another variable is eliminated. We know that correlation between two variables may occur because both of them are correlated with a third variable. Partial correlation controls for this third variable. For example, assume you are

interested in the correlation between vocabulary and problem-solving skills. Both these variables are related to a third variable, chronological age. For example, 12-year-old children have more developed vocabularies than 8-year-old children, and they also have more highly developed problem-solving skills. Scores on vocabulary and problem solving will correlate with each other because both are correlated with chronological age. Partial correlation would be used with such data to obtain a measure of correlation with the effect of age removed. The remaining correlation between two variables when their correlation with a third variable is removed is called a *first-order partial correlation*. Partial correlation may be used to remove the effect of more than one variable. However, because of the difficulty of interpretation, partial correlation involving the elimination of more than one variable is not often used.

## MULTIPLE REGRESSION

**Multiple regression** is a correlational procedure that examines the relationships among several variables. Specifically, this technique enables researchers to find the best possible weighting of two or more independent variables to yield a maximum correlation with a single dependent variable. For example, colleges use data submitted by prospective freshmen to predict first-semester GPA. The predictor values may be scores on the SAT subtests [SAT verbal (SATV) and SAT math (SATM)], along with students' relative high school rank (RHSR). Relative high school rank, found by dividing a student's rank in the high school graduating class by the size of the class, adjusts for the variation in size of graduating classes. Table 13.3 shows the simple correlations between each of the predictors and the criterion.

You can see in Table 13.3 that none of the variables has a very high correlation with freshman GPA; the best single predictor is relative high school rank. However, we can use all three variables in a multiple regression analysis to determine the correlation of the best possible weighted combination of the three predictor variables with GPA. Computer programs produce a **prediction equation** with the **predictor variables** weighted in the appropriate way to yield the highest correlation with GPA and hence the best prediction. The university can use the equation with similar groups of prospective students whose SAT scores and relative high school ranks are known to predict their as yet unknown GPA at the university.

The regression equation would look as follows: $Y' = a + b_1X_1 + b_2X_2 + b_3X_3$, where $Y$ is the predicted score (GPA); $a$, $b_1$, $b_2$, and $b_3$ are constants provided by the regression analysis; and $X_1$, $X_{12}$, and $X_3$ are the independent variables (RHSR, SATV, and SATM, respectively). Assume a student has the following scores: SATV = 510, SATM = 540, and RHSR = 21. The student's predicted GPA at the university would be 2.97:

$$Y' = 2.0813 + (-.0131)21 + .0014(540) + .0008(510) = 2.97$$

The regression analysis also yields $R$, the **coefficient of multiple correlation**, which indicates the relationship between the predictor variables in combination and the criterion. If we square $R$ to get the coefficient of determination, we know the amount of variability in the criterion that is due to differences in scores on the predictor variables. For example, if $R$ is .50, then 25 percent of the variability in GPA can be accounted for by the combined SATV, SATM, and RHSR scores.

**Table 13.3** Correlations of Each Predictor with the Criterion

| | SATV | SATM | RHSR[a] |
|---|---|---|---|
| GPA | 0.31 | 0.39 | −0.42 |

[a]The negative correlation between relative high school rank and GPA is due to the way that rank in class is measured. The highest achiever in the class has a rank of 1 (the lowest number); the lowest achiever in the class has a rank equal to the size of the class (a high number). The students with the lowest *size* rank are predicted to have the highest GPA; hence, the correlation coefficient is negative.

In the development of a multiple regression equation, the variables should be measured on an interval scale. It is possible, however, to put categorical variables such as gender, social class, marital status, political preference, and the like into a prediction equation if they are recoded as binary variables. For instance, if the variable is gender, 1's can be assigned to females and 0's to males. Such recoded variables are referred to in multiple regression as **dummy variables**.

Because the computations are very complex, multiple regression is done by computer. Computer programs are available that provide not only the multiple correlation coefficient ($R$) and the prediction equation but also the proportion of variance in the criterion accounted for by the combination of predictors ($R^2$) and tests of statistical significance for the $R$ and for the contribution of each predictor. For further discussion of multiple regression, see Cohen, Cohen, West, and Aiken (2003).

### THINK ABOUT IT 13.4

Match the procedure listed in the left column with the definition in the right column:

| | |
|---|---|
| 1. Spearman rho | a. Shows sign and magnitude of correlation between two nominal variables |
| 2. Pearson *r* | b. Shows sign and magnitude of correlation between two ordinal variables |
| 3. Multiple regression | c. Shows sign and magnitude of correlation between two interval variables |
| 4. Phi coefficient | d. Uses a number of independent variables to predict a single dependent variable |
| 5. Eta correlation | e. Used when the relationship between two variables is curvilinear |

**Answers**

1. b; 2. c; 3. d; 4. a; 5. e

## FACTOR ANALYSIS

**Factor analysis**, or **exploratory factor analysis**, is a family of techniques used to detect patterns in a set of interval-level variables (Spicer, 2005). Factor analysis begins with a table of pairwise correlations (Pearson *r*'s) among all the variables of interest; this table is called a **correlation matrix**. The purpose of

the analysis is to try to reduce the set of measured variables to a smaller set of underlying factors that account for the pattern of relationships. The search follows the law of parsimony, which means that the data should be accounted for with the smallest number of factors. This reduction of the number of variables serves to make the data more manageable and interpretable.

There are two types of situations in which factor analysis is typically used. In the first, a researcher is interested in reducing a set of variables to a smaller set. For example, assume a technology company uses 10 different tests to select computer programmers. Factor analysis could be used to identify perhaps four underlying dimensions measured by those 10 tests so that tests of the four dimensions could be used just as effectively in the selection process as the 10 original tests.

The second type of situation is when researchers use factor analysis to determine the characteristics or underlying structure of a measuring instrument such as a measure of intelligence, personality, or attitudes. Assume a researcher has developed a new scale to measure self-esteem and thinks it is unidimensional (measuring one single dimension). If this is true, factor analysis should yield only one factor. In other cases, a researcher may be interested in investigating the nature of the underlying factors in an existing scale. In Chapter 9, we discussed the use of factor analysis in establishing the construct validity of tests.

Let us illustrate factor analysis with a simple example. Imagine that you have scores on six different subscales of an aptitude measure for 300 subjects. The correlations among all the pairs of scores are shown in Table 13.4. Each subject is shown both on the horizontal rows and vertical columns. The Pearson *r* for each pair of variables is shown where the columns and rows intersect.

The question is: Is there a simpler structure underlying these 15 correlations? Table 13.4 shows that all of the subscales are positively correlated, and we assume all of the correlations are statistically significant. The first two subscales (vocabulary and analogies) form a separate subgroup because they correlate .50 with each other but do not correlate with the other subscales. The next two subscales (arithmetic and numerical reasoning) correlate .55 and thus form another subgroup, and likewise the last two subscales (picture completion and block design) have a correlation of .52 with each other but negligible correlations with other subscales. In other words, the pattern of correlations among these variables seems to reflect three underlying factors, which we label verbal, numerical, and spatial.

**Table 13.4** Correlations among Subscale Scores

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1. Vocabulary | — | .50 | .15 | .12 | .12 | .15 |
| 2. Analogies | — | — | .12 | .15 | .10 | .18 |
| 3. Arithmetic | — | — | — | .55 | .15 | .12 |
| 4. Numerical reasoning | — | — | — | — | .20 | .22 |
| 5. Picture completion | — | — | — | — | — | .52 |
| 6. Block design | — | — | — | — | — | — |

In this simple example, we are dealing with only six variables. In most cases, there would be a greater number, and it would not be so easy to discover the factors by inspection. Thus, researchers turn to factor analysis. The mathematics of factor analysis is beyond the scope of this book. However, basically it involves searching for the clusters of variables that are all correlated with each other. The first cluster identified is called the first factor, and it represents the variables that are most intercorrelated with each other. Then, other factors are identified that account for decreasing amounts of the variance. The factor is represented as a score, which is generated for each subject in the sample. Next, a correlation coefficient is computed between subjects' factor score and their score on the particular variable entered into the factor analysis. This correlation between a variable and a factor is called the **factor loading**. The higher its loading, the more a variable contributes to and defines a particular factor. A factor loading is interpreted like a correlation coefficient: The larger it is (either positive or negative), the stronger the relationship of the variable to the factor. The result of the factor analysis is a factor matrix, which shows the number of important underlying factors and the weight (loading) of each original variable on the resulting factors. The square of the factor loading is the proportion of common variance between the test and the factor. Table 13.5 shows what the hypothetical factor matrix resulting from a factor analysis of the intelligence test in the previous example might look like.

The first two tests load strongly on factor 1; we might call this underlying factor "verbal ability." The next two tests load strongly on factor 2, which we might label "numerical ability." The last two tests load strongly on factor 3, which we might label "spatial ability." Our simple example thus suggests that there were three factors underlying performance on the intelligence test. This procedure did not involve a hypothesis to be tested but, rather, something to be explored. How does one decide on the "correct" number of factors? The first criterion is that all the factors should be interpretable; an uninterpretable factor serves no practical or theoretical function (Spicer, 2005). Second, the factors should account for a satisfactory amount of shared variance in the data. What is "satisfactory" is defined by the researcher. Some writers suggest that the analysis keep extracting factors as long as a factor accounts for at least another 10 percent of the variance. "There is general agreement that overfactoring is preferable to underfactoring" (Spicer, 2005, p. 195).

**Table 13.5** Hypothetical Factor Matrix from an Analysis of Scores on Six Subscales[a]

| Test Subscale | Factor 1 | Factor 2 | Factor 3 |
|---|---|---|---|
| Vocabulary | **.91** | .40 | .30 |
| Analogies | **.87** | .30 | .20 |
| Arithmetic | .25 | **.90** | .15 |
| Numerical reasoning | .22 | **.80** | .10 |
| Picture completion | .15 | .10 | **.85** |
| Block design | .09 | .05 | **.75** |

[a]The variables loading most strongly on each factor are set in boldface.

This preceding discussion illustrates exploratory factor analysis (EFA) because a researcher does not test any formal hypotheses about the number of underlying factors. The number of factors is determined empirically rather than being specified a priori. It is distinguished from the more advanced technique called confirmatory factor analysis (CFA), which we describe briefly next.

### CONFIRMATORY FACTOR ANALYSIS

**Confirmatory factor analysis**, like exploratory factor analysis, "is used to examine the relationships between a set of measured variables and a smaller set of factors that might account for them" (Spicer, 2005, p. 199). Confirmatory factor analysis, however, assumes relatively precise advance knowledge and allows a researcher to specify a priori what these relationships might look like and then to test the accuracy of these formal hypotheses.

The first step in CFA is to specify a model made up of a number of hypotheses about the number of underlying factors, whether or not they are correlated, and which variables are expected to load on which factors. The output of CFA allows the researcher to evaluate the factor model overall and at the level of individual variable–factor relationships. The researcher can use CFA and compare different models or factor solutions that might be proposed.

Researchers often use both EFA and CFA in the construction and evaluation of measuring instruments. They begin with EFA and then move to CFA at later stages in the research. CFA is beyond the scope of an introductory text. Interested readers may refer to Pedhazur (2006), Loehlin (2004), or Thompson (2004).

## OTHER COMPLEX CORRELATIONAL PROCEDURES

Several more complex techniques are available to investigate correlation of more than two variables. These analyses require more sophistication with statistics than is usually needed in a beginning research course. We briefly describe these techniques and refer interested students to other texts.

**Canonical correlation** is a generalization of multiple regression that adds more than one dependent variable (criterion) to the prediction equation. For more information on canonical correlation, see Thompson (1984).

**Discriminant analysis** is a statistical procedure related to multiple regression, but it differs in that the criterion is a categorical variable rather than a continuous one. A good source for this procedure is Huberty (1994).

**Structural equation modeling (SEM)** is a popular technique used in the analysis of causality. SEM combines confirmatory factor analysis and path analysis to test both a measurement model and a structural model. We refer the reader to Bentler and Chou (1988), Pedhazur (2006), and Loehlin (2004) for further discussion of SEM.

Pedhazur (2006) defines **path analysis** as "a method for studying direct and indirect effects of variables hypothesized as causes of variables treated as effects."

Pedhazur further states, "Path analysis is intended *not* to discover causes but to shed light on the tenability of the causal models a researcher formulates based on knowledge and theoretical considerations" (p. 769).

### THINK ABOUT IT 13.5

Match the procedure in the left column with the definition in the right column:

| | |
|---|---|
| 1. Factor analysis | a. Uses multiple independent variables to predict more than one dependent variable |
| 2. Discriminant analysis | b. Reduces a matrix of correlations among variables to a few underlying constructs |
| 3. Canonical correlation | c. Uses a number of variables to predict membership in categories |
| 4. Partial correlation | d. Uses theory to specify and test models of causation among variables |
| 5. Path analysis | e. Determines the relationship between two variables when the effect of another variable is eliminated |

**Answers**

1. b; 2. c; 3. a; 4. e; 5. d

## SUMMARY

Correlational research is nonexperimental research that studies the direction and strength of relationships among variables. It gathers data on two or more quantitative variables from the same group of subjects (or from two logically related groups) and then determines the correlation among the variables.

Correlational procedures are widely used in educational and psychological research. They enable researchers to better understand certain phenomena and to make predictions. Correlational designs are often valuable for generating hypotheses that can be further investigated in experimental or ex post facto research. Correlations must be interpreted appropriately. When assessing a correlation coefficient, one must take into account the population from which the sample was drawn, the shape of the distribution, the sample size, and its statistical and practical significance. The most serious error is to interpret correlation per se as an indicator of causation.

A number of different types of correlation coefficients are used with variables that are measured on different types of scales. Multiple regression is used to find the relationship between two or more independent variables and a dependent variable. It yields a prediction equation that the researcher can use later to predict the dependent variable for a new group of subjects, when the researcher has information only on the independent variables.

Some more sophisticated correlational procedures include partial correlation, discriminant analysis, factor analysis, canonical correlation, path analysis, and structural equation modeling. It is important to know the type of research situation in which each of these techniques would be useful.

## KEY CONCEPTS

assessing consistency
assessing relationships
canonical correlation
causation
cause-and-effect relationship
coefficient of correlation
coefficient of determination
coefficient of multiple correlation
confirmatory factor analysis
correlation coefficient
correlation matrix
correlation ratio
correlational research
discriminant analysis
dummy variable
exploratory factor analysis
factor analysis
factor loading
linear correlation
multiple regression
nonparametric indexes of relationships
partial correlation
path analysis
phi coefficient
prediction
prediction equation
predictor
predictor variables
significance fallacy
Spearman rho
structural equation modeling

## EXERCISES

**1.** Identify the following as
   1. Characteristic of correlational research
   2. Characteristic of ex post facto research
   3. Characteristic of both
   4. Characteristic of neither

   **a.** Actively manipulates variables
   **b.** Looks at differences between two or more groups
   **c.** Assesses the direction and strength of association among variables
   **d.** Produces indexes that range from −1.00 to +1.00
   **e.** Studies preexisting variables

**2.** Which correlation in each pair indicates the greatest extent of relationship between two variables?
   **a.** +.80 or −.80
   **b.** +.60 or+.85
   **c.** +.60 or −.85
   **d.** +.04 or −.15

**3.** A state official found a correlation of .65 between the number of bars in cities in the state and the number of crimes reported in those cities. He concluded that legislation to reduce the number of bars would reduce crime. What would you say about the official's conclusion?

**4.** Why are magnitude interpretations of correlation coefficients alone sometimes misleading?

**5.** Describe the relationship between the following pairs of scores ($X$ and $Y$): Constructing a scattergram may be useful.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $X$ | 2 | 3 | 10 | 6 | 4 | 9 | 7 | 3 | 6 |
| $Y$ | 10 | 9 | 1 | 3 | 7 | 1 | 2 | 5 | 4 |

**6.** A study finds a moderate positive relationship between measures of self-esteem and achievement scores for a group of middle school students. Which of the following correlation coefficients would best represent this relationship?
   **a.** .94
   **b.** .35
   **c.** −.62
   **d.** .20

**7.** A study finds almost no relationship between two physical fitness scores for a group of elementary school students. Which of the following coefficients would best represent this relationship?
   **a.** .25
   **b.** .75
   **c.** −.10
   **d.** .60
   **e.** −.80

**8.** A researcher administered a series of tests to a group of beginning first-year law school students at State University. At the end of the first year of law school, the researcher obtained the students' GPAs.
   **a.** What correlational technique should be used to determine the maximum relationship of these tests to first-year GPA?
   **b.** What assumptions would have to be made in order to generalize the findings to future applications to the same law school? To applications to other law schools?

**9.** How would you compare the correlations .70 and .35 in terms of magnitude?

**10.** A researcher is interested in anxiety and how it might affect performance on scholastic aptitude tests. He has a clinical psychologist assess the anxiety of subjects by ranking them from 1 through 20. Then he administers a standardized aptitude test to each of the 20 subjects and converts their IQ scores to ranks. Which correlation coefficient should the researcher calculate for the data? Explain your answer.

**11.** A researcher demonstrated a correlation of +.35 between principals' ratings of teacher attire and student academic performance across 150 grade schools in his state. He concluded that encouraging teachers to be properly attired will increase academic performance. Comment on his conclusion.

**12.** Determine the kind of index of correlation appropriate for use in solving each of the following problems:
- **a.** What is the coefficient of correlation between self-esteem and aptitude scores for a group of subjects?
- **b.** What is the correlation between gender and being hired or not hired for a job when you have 60 men and 60 women with identical scores on a qualifying test for that job?
- **c.** What is the correlation between the responses of 180 students to two test items when you know the right and wrong responses of these students on both test items?
- **d.** What is the relationship between sociometric scores of a group of students and their academic rank in a high school class?
- **e.** How is the value of an automobile from the day you drive it from the dealer's showroom until it becomes an antique related to its age?

**13.** To compute a correlation coefficient between traits A and B, which of the following must you have?
- **a.** One group of subjects, some of whom possess characteristics of trait A and the remainder of whom possess those of trait B
- **b.** Measures of trait A on one group of subjects and measures of trait B on another group
- **c.** One group of subjects, some who have both traits A and B, some who have neither trait, and some who have one trait but not the other
- **d.** Two groups of subjects, one of which could be classified as A or not A, and the other as B or not B
- **e.** Measures of traits A and B on each subject in one group of students

**14.** An investigation finds a positive correlation between IQ scores and length of attention span among 10-year-olds. From this finding, which of the following would be a reasonable conclusion?
- **a.** A long attention span is a cause of intelligence.
- **b.** A high IQ is a cause of long attention span.
- **c.** There is a high probability that a large sample of high-IQ 10-year-olds will have a shorter mean attention span than a large sample of low-IQ 10-year-olds.
- **d.** One would predict longer attention spans for high-IQ 10-year-olds than for low-IQ 10-year-olds.

**15.** Examine the following research topics and decide whether experimental research, ex post facto research, or correlational research is the appropriate research design for each one:
- **a.** The effect of parents' divorce on the achievement motivation of the children
- **b.** The effect of a specific program of vocabulary instruction on social studies achievement
- **c.** The relationship between self-esteem and grade point average
- **d.** The effect of phonics instruction on the reading grade level of fourth-grade students
- **e.** The relationship between the verbal aptitude scores of identical twins
- **f.** The relationship between preschool attendance or nonattendance and academic achievement in first-graders

**16.** A researcher found a correlation of −.42 between rank in high school class and achievement in college for a sample of 1500 college freshmen. How would you interpret this coefficient in terms of direction, magnitude, and percentage of shared variance?

**17.** A school administrator wants to predict achievement in foreign languages. She has available scores on an intelligence test, a

language aptitude test, and a reading test; she also knows the sex of the students. How would you recommend that she proceed to predict foreign language achievement?

18. Explain how it is possible for the measures of two variables to be associated in a fairly systematic way without the existence of any causal relationship between the variables?
19. A researcher is investigating the validity of an inventory designed to measure teacher stress. He wishes to know if the inventory is measuring the single construct, teacher stress, or more than one construct. What procedure would you recommend that the researcher use to answer this question?
20. A researcher wanted to predict achievement in the first year of law school. He conducted a multiple-regression analysis with a sample size of 1000 and used six predictor variables: undergraduate GPA, undergraduate major, LSAT score, score on a writing test, sex, and time lapse between undergraduate degree and application to law school. The resulting $R = .20$ was significant at the .01 level of significance. How would you interpret the researcher's findings?
21. A researcher wants to investigate the relationship between class size and student satisfaction in sections of beginning psychology at a university. Describe two research methodologies that might be appropriate to answer this question.
22. A researcher found a moderate negative correlation between two variables of interest. She concluded that the negative correlation meant that it would be useless to try to predict one variable from the other. Do you agree? Why or why not?
23. How are ex post facto and correlational research alike?
24. How are ex post facto and correlational research different?
25. If the population correlation is .32 or greater to reject the hypothesis that the population correlation is zero with a one-tailed .01 level, how many subjects do you need?
26. A football coach finds that his best men are either small, agile players or hulking, large ones.
    **a.** Describe the relationship between weight and success in his team.
    **b.** What index could he use to describe this relationship?
27. Classify the following as (1) attenuated independent, (2) attenuated dependent, or (3) neither a nor b:
    **a.** Vocabulary test scores and reading grades in class for gifted fifth-graders
    **b.** Vocabulary test scores and reading grades among third-graders in Concord, New Hampshire.
    **c.** Height and basketball scores among professional players

## ANSWERS

1. **a.** 4
   **b.** 2
   **c.** 1
   **d.** 1
   **e.** 3
2. **a.** The extent of relationship is the same.
   **b.** +.85
   **c.** −.85
   **d.** +.15
3. Correlation between the two variables does not indicate causation. In this case, common cause is a likely explanation. Large cities have more bars and more crime; smaller cities have fewer bars and fewer crimes.
4. To interpret the correlation, you need to know the sample size on which it is based and the magnitude of other correlations of the same variables in other studies.
5. The relationship is
   **a.** Linear
   **b.** High
   **c.** Negative
6. .35
7. −.10
8. **a.** Multiple regression
   **b.** The researcher would have to assume that the relationship among the variables in the population of applicants to the law school remains similar over time.

To apply the results to other law schools, the researcher would have to assume that the relationships among variables among applicants to other law schools are similar to these relationships among applicants to her law school.

**9.** An $r$ of .70 might be described as being a very high correlation and the $r$ of .35 as being moderate. One must *not* say that the $r$ of .70 is twice as high as the $r$ of .35. A good way to compare them is in terms of the coefficient of determination ($r^2$). With an $r$ of.70, 49 percent of the variance is shared variance; with an $r$ of .35, only 12 percent of the variance in $Y$ is accounted for by variance in $X$. Thus, the difference in magnitude of the relationship is more like 4 to 1 instead of 2 to 1, as might have been incorrectly assumed from the absolute size of the coefficients.

**10.** Because the researcher has ordinal or rank-order data, he should calculate the Spearman rho correlation coefficient.

**11.** The researcher has no justification for inferring a causal relationship merely on the basis of correlational evidence. Principals' ratings of teacher attire and student academic performance could very well be functions of some other variable.

**12.** **a.** Pearson product moment
**b.** Phi coefficient
**c.** Phi coefficient
**d.** Spearman rho
**e.** Eta squared

**13.** e

**14.** d

**15.** **a.** Ex post facto research
**b.** Experimental research
**c.** Correlational research
**d.** Experimental research
**e.** Correlational research
**f.** Ex post facto research

**16.** The $r$ of $-.42$ indicates a moderate negative relationship between rank in high school class and achievement in college; that is, students with a low numeric ranking would tend to have high achievement. The negative correlation is due to the way that rank in class is measured. The highest achiever in the class has a rank of 1 (the lowest number), and the lowest achiever in the class has a rank equal to the size of the class (the highest number). The student with a rank of 1 in a class of 400 would be expected to have a higher GPA in college than the student with a rank of 400. An $r = -.42$ indicates that approximately 18 percent of the variance in college grades can be accounted for by achievement in high school as indicated by rank.

**17.** The school administrator would select a sample of students and obtain their scores on the three tests and their grades in foreign languages. Sex would be coded 0 or 1. The data would be entered into a multiple regression analysis, with the intelligence test, language aptitude test, reading test, and gender as independent variables and grades in foreign language as the dependent variable. Once the multiple-regression equation is developed, it can be used to predict foreign language achievement of similar groups of students when only the independent variables are known.

**18.** Even when two variables are correlated, you cannot infer that one causes the other. Correlation does not indicate causation. A third variable, unmeasured by the researcher, may account for the observed relationship.

**19.** The researcher should subject the inventory to factor analysis.

**20.** With the large sample size, it is not surprising that a coefficient of only .20 would be statistically significant. It is only moderately useful in prediction. $R^2 = .04$; only 4 percent of the variance in law school achievement would be predictable from this combination of predictor variables.

**21.** You could use ex post facto methodology by selecting already existing sections with different numbers of students and comparing the mean satisfaction scores in the various sections. Or you could use correlational methodology by correlating class size with student satisfaction scores.

**22.** A negative correlation is just as good for predictive purposes as a positive correlation. It is the magnitude of the correlation that is important in this case.

**23.** Neither has active independent variables but depend on measures of preexisting

status. Therefore, you must be very cautious before interpreting the results as indicating cause and effect.

**24.** Ex post facto research examines differences between natural or arbitrary groups. Correlational studies assess both the direction and the strength of relationships between variables within a single group or set of matched pairs. Correlation indexes are easy to picture. Most range from −1.00 (perfect negative) thru 0 (no relationship) to +1.00 (perfect positive). Some range from 0 to +1.00.

**25.** 90 + 2 = 92

**26. a.** Curvilinear
**b.** Eta squared

**27. a.** 1
**b.** 3
**c.** 2

## REFERENCES


Bentler, P., & Chou, C. (1988). Practical issues in structural modeling. In J. S. Long (Ed.), *Common problems/proper solutions: Avoiding error in quantitative research*. Newbury Park, CA: Sage.

Cohen, J., Cohen, P., West, S., & Aiken, L. (2003). *Applied multiple regression/correlational analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Huberty, C. (1994). *Applied discriminant analysis*. New York: Wiley.

Loehlin, J. (2004). *Latent variable models: An introduction to factor, path, and structural equation analysis*. Mahwah, NJ: Erlbaum.

Pedhazur, E. (2006). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). Belmont, CA: Wadsworth.

Siegel, S., & Castellan, N. J., Jr. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York: McGraw-Hill.

Spicer, J. (2005). *Making sense of multivariate data analysis*. Thousand Oaks, CA: Sage.

Thompson, B. (1984). *Canonical correlation analysis: Uses and interpretation*. Beverly Hills, CA: Sage.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.

CHAPTER 14

# Survey Research

Ask and you shall know.

## INSTRUCTIONAL OBJECTIVES

After studying this chapter, the student will be able to:

1. State the purpose of survey research.
2. Describe the four categories of surveys classified according to their scope and focus.
3. Distinguish between longitudinal and cross-sectional surveys.
4. Describe the different types of longitudinal surveys.
5. List the steps involved in carrying out a survey.
6. Explain the importance of probability sampling in survey research.
7. Define margin of error and use sample data to calculate the margin of error and the confidence interval around the population parameter.
8. Calculate the sample size needed to achieve a desired margin of error in a survey.
9. List some of the factors that influence a researcher's decision about sample size.
10. State the merits and disadvantages of the interview as a data-gathering technique.
11. Describe data-gathering techniques used in survey research.
12. Define focus group.
13. Write both open-ended and closed-ended survey questions.
14. List guidelines for conduction an interview.
15. State the advantages and disadvantages of the questionnaire as a data-gathering technique.
16. State some rules for writing items for a questionnaire.
17. List guidelines to follow for developing the format of a questionnaire.
18. Explain the advantages of field testing a questionnaire.
19. Explain the follow-up procedures a researcher should employ with a mailed survey.
20. Write a cover letter for a questionnaire.
21. Outline procedures for dealing with nonrespondents after follow-up procedures have been used.

**22** Discuss the procedures for assessing the validity and reliability of questionnaires and/or interviews.

**23** Explain the data analyses that are appropriate for survey data.

**24** Explain the statistics used with cross tabulations.

**25** Describe issues with electronic surveys.

In **survey research,** investigators ask questions about peoples' beliefs, opinions, characteristics, and behavior. The survey questionnaire is widely used as a source of data in studies in sociology, business, psychology, political science, and education. It also provides information used for nonresearch planning and policy setting in the areas of government, business, health, and education. For example, the U.S. Census Bureau conducts a monthly survey for the Bureau of Labor Statistics that focuses on measuring labor force participation and unemployment. These data are used to produce the monthly unemployment figures for the United States as well as for the individual states. Surveys are taken of consumer choices, use of health services, numbers of women and minority faculty in universities, and so on. A survey researcher may want to investigate associations between respondents' characteristics such as age, education, social class, race, and their current attitudes toward some issue. Survey research typically does not make causal inferences but, rather, describes the distributions of variables in a specified group.

Although researchers want to gather information about the characteristics of populations, they usually study a smaller group (a sample) carefully drawn from the population and then use the findings from the sample to make inferences about the population.

The range of topics covered by surveys and the techniques used has increased significantly in the past several years. Hardly a week goes by that you are not exposed through the news media to the results of some survey. The well-known Gallup poll, for instance, surveys public opinion on a variety of issues. Market researchers ask what products you purchase or might purchase, political pollsters ask whom you are likely to vote for, and television networks want to know what shows you watch.

Many doctoral dissertations and much of the research published in educational journals involve survey methods. Public school districts also conduct surveys to gather data such as average teacher load, number of preschool children in the community, number of students who participate in extracurricular activities, opinions of parents and students, and data on a myriad of other topics. The 39th Annual Phi Delta Kappa/Gallup poll for 2007 (Rose & Gallup, 2007) surveyed 1005 adults to get their views on the current state of public schools. The data showed that the public's generally high regard for the public schools has dropped marginally from 2006, but, as always, parents gave a higher grade to the school their oldest child attended. The parts of the survey focused on the No Child Left Behind (NCLB) Act of 2002 indicated that a larger number of those polled felt knowledgeable about NCLB and that more people have an unfavorable view of NCLB than in previous polls. Criticism of standardized testing has increased, but

other options for improving schools are receiving more favorable consideration. Lack of adequate funding continues to be the major problem facing schools.

### THINK ABOUT IT 14.1

Which of the following questions would best be answered by survey methods?

1. Do voters in our school district think we should raise taxes in order to build new classrooms?
2. Do people who have taken driver education have fewer accidents than people who have not?
3. What do voters consider the most important issues in the upcoming election?
4. What do school principals consider to be the major problems in their schools?
5. Does dividing second-grade math students into ability groups produce greater math achievement than doing math instruction in a single group?

**Answers**

Questions 1, 3, and 4 would be suitable for survey research. Ex post facto would be appropriate for question 2, and experimental research would be suitable for question 5.

## TYPES OF SURVEYS

Before initiating survey research, the investigator must determine the format that is most appropriate for the proposed investigation. Surveys are classified according to their focus and scope (census and sample surveys) or according to the time frame for data collection (longitudinal and cross-sectional surveys). Becoming familiar with the options enables the researcher to select the method that will provide the most useful data.

### SURVEYS CLASSIFIED ACCORDING TO FOCUS AND SCOPE

A survey that covers the entire population of interest is referred to as a **census,** an example of which is the U.S. Census, undertaken by the government every 10 years. In research, however, "population" does not refer to all the people of a country. The term **population** is used to refer to the entire group of individuals to whom the findings of a study apply. The researcher defines the specific population of interest. It is often difficult or even impossible for researchers to study very large populations. Hence, they select a smaller portion, a **sample,** of the population for study. A survey that studies only a portion of the population is known as a **sample survey.**

Surveys may be confined to simple tabulations of **tangibles,** such as what proportion of children ride school buses and the average class enrollment. The most challenging type of survey is one that seeks to measure **intangibles,** such as attitudes, opinions, values, or other psychological and sociological constructs. In such a study, you must bring to bear not only the skills involved in proper sampling but also the skills involved in identifying or constructing

appropriate measures and employing the scores on these measures to make meaningful statements about the constructs involved. If you classify surveys on the basis of their scope (census versus sample) and their focus (tangibles versus intangibles), four categories emerge: (1) a **census of tangibles,** (2) a **census of intangibles,** (3) a **sample survey of tangibles,** and (4) a **sample survey of intangibles.** Each type has its own contributions to make and its own inherent problems.

### A Census of Tangibles

When you seek information about a small population, such as a single school, and when the variables involved are concrete, there is little challenge in finding the required answers. If a school principal wants to know how many desks are in the school, how many children ride the school bus, or how many teachers have master's degrees, a simple count will provide the information. Because the study covers the entire population, the principal can have all the confidence characteristic of perfect induction. Well-defined and unambiguous variables are being measured, and as long as the enumeration is accurate and honest, the principal can say, without much fear of contradiction, "On the first of September there were 647 children's desks in our school" or "Sixty-five percent of the present faculty has master's degrees." The strength of a census of this type lies in its irrefutability. Its weakness lies in its confinement to a single limited population at a single point in time. The information provided by such a census may be of immediate importance to a limited group, but typically such surveys add little to the general body of knowledge in education.

### A Census of Intangibles

Suppose the school principal now seeks information about pupil achievement or aspirations, teacher morale, or parents' attitudes toward school. The task will be more difficult because this census deals with constructs that are not directly observable but must be inferred from indirect measures. Test scores and responses to questionnaires serve to approximate constructs such as knowledge and attitudes. The National Study of School Evaluation publishes an opinion inventory designed to measure student, teacher, and parent attitudes and opinions about schools. Administering the inventory to all the students, teachers, or parents in the school system would represent a census of intangibles.

Another example of this type of census is the achievement-testing program carried out by most schools. All children are tested, and the test scores are used to compare their performance with national norms, their own previous performance, and so on.

The value of a census of intangibles is largely a question of the extent to which the instruments used actually measure the constructs of interest. Reasonably good instruments are available for measuring aptitude and achievement in a variety of academic areas. Many other variables remain very difficult to measure. Because researchers lack instruments that can meaningfully measure the constructs involved, many important questions in education have not been answered. Such variables as teacher success, student motivation, psychological adjustment, and leadership have been difficult to define and measure operationally.

### A Sample Survey of Tangibles

When investigators seek information about large groups, the expense involved in carrying out a census is often prohibitive. Therefore, researchers use sampling techniques and use the information they collect from the sample to make inferences about the population as a whole. When sampling is done well, the inferences made concerning the population can be quite reliable.

A classic example of a sample survey of tangibles is the report on equality of educational opportunities commissioned by the U.S. Department of Health, Education, and Welfare. This study, sometimes called the Coleman Report after James Coleman who developed the survey, was conducted in response to Section 402 of the Civil Rights Act of 1964, which directed the Commissioner of Education to conduct a survey of inequalities in educational opportunities among various groups in the United States. The sample survey included more than 600,000 children in grades 1, 3, 6, 9, and 12 of approximately 4000 schools. The schools were considered generally representative of all U.S. public schools, although there was some intentional overrepresentation of schools with minority group populations. From the data generated by the survey, the researchers concluded that 65 percent of blacks attended schools in which more than 90 percent of students were black and 80 percent of whites attended schools enrolling more than 90 percent white. When comparisons were made concerning class size, physical facilities, and teacher qualifications, relatively little difference was found among schools serving different racial and ethnic groups. However, these variables did differ between metropolitan and rural areas and between geographic regions. Those disadvantaged in regard to these variables appeared to be rural children and those in the South, regardless of race.

### A Sample Survey of Intangibles

The public opinion polls are examples of studies measuring intangible constructs. Opinion is not directly observable but must be inferred from responses made by the subjects to questionnaires or interviews. Opinion polling began in the 1930s and has grown tremendously. Where respondents have been willing to reveal their preferences freely before elections, for instance, pollsters have been quite accurate in inferring public opinion from which they have predicted subsequent election results. These polls have provided excellent examples of the usefulness of sample statistics in estimating population parameters. However, if people who support one candidate are reluctant to reveal their preference, whereas people who support the other candidate feel free to say so, considerable error is introduced into the results of the poll. For example, people are more willing to say they will vote against an incumbent than for him or her. A classic example occurred before the 1948 presidential election when several polls showed Dewey leading the incumbent, Truman, but with many people indicating they were undecided. Newspapers had already prepared headlines proclaiming Dewey the winner, but Truman won the election. Apparently, most of those who indicated they were undecided actually voted for Truman. Respondents are also reluctant to reveal a choice that may appear to be based on self-interest, prejudice, or lack of knowledge about the issues.

How someone is going to vote is an intangible, but what is marked on a ballot is tangible. The television network news services have done very well in predicting how states will vote when only a few precincts have reported because they

can use tangible measures of a sample (i.e., how some ballots have been marked) to predict the vote of a population. Therefore, the risks are only those involved in estimating population parameters from sample statistics. However, pollsters who estimate how a population *will* vote on the basis of how people *say* they will vote have the additional handicap of measuring what is intangible at the time the measurements are made. Surveys of intangibles are limited by the fact that the data researchers collect are only indirectly measuring the variables they are concerned about. The seriousness of this limitation depends on how well the observations measure the intangible variable.

The same survey may study tangibles and intangibles at the same time. The survey on equality of educational opportunities asked the students to answer questionnaires and administered intelligence and achievement tests in order to make inferences about social class, ability, and achievement, as well as the relationship of these variables to each other and to tangible variables in the study.

### THINK ABOUT IT 14.2

What type of survey is illustrated in the following examples?

1. Randomly selected teachers are asked how many years experience they have teaching?
2. The superintendent of School District 214 has Iowa Test of Basic Skills scores for all second-graders in his district.
3. The state superintendent of schools' list of the enrollment in each of the state's 1143 schools.
4. Some students in Shaw School are given a physical fitness test to get an estimate of the fitness of all the students in the school.

**Answers**

1. Sample survey of tangibles
2. Census of intangibles
3. Census of tangibles
4. Sample survey of intangibles

## SURVEYS CLASSIFIED ACCORDING TO THE TIME DIMENSION

Surveys are also classified according to the time of data collection: longitudinal surveys, which study changes across time, and cross-sectional surveys, which focus on a single point in time.

### Longitudinal Surveys

**Longitudinal surveys** gather information at different points in time in order to study the changes over extended periods of time. Three different designs are used in longitudinal survey research: panel studies, trend studies, and cohort research.

*Panel Studies* In a **panel study,** the *same* subjects are surveyed several times over an extended period of time. For example, a researcher studying the development of quantitative reasoning in elementary school children would select a sample of

first-graders and administer a measure of quantitative reasoning. This same group would be followed through successive grade levels and tested each year to assess how quantitative reasoning skills develop over time. Researchers have studied how age affects IQ by measuring the same individuals as adolescents and when they were college-aged, middle-aged, and older. Because the same subjects are studied over time, researchers can see the changes in the individuals' behavior and investigate the reasons for the changes. An example of a panel study is Terman's (1926) classic study of intelligence in which he followed exceptionally bright children to maturity.

*Trend Studies* A **trend study** differs from a panel study in that *different* individuals randomly drawn from the same general population are surveyed at intervals over a period of time. For example, researchers who have studied national trends in mathematics achievement sample middle school students at various intervals and measure their math performance. Although the same individuals are not tested each time, if the samples from the population of middle school students are selected randomly, the results each time can be considered representative of the middle school population from which the student samples were drawn. Test scores from year to year are compared to determine if any trends are evident. Another example of a trend study is the survey on alcohol, tobacco, and other drug use among Indiana youth conducted annually since 1991 by the Indiana Prevention Resource Center. The 2007 survey collected data from 158,632 pupils in public and private schools in Indiana. Results include a continuing long-term decline in cigarette smoking, an increase in binge drinking, a 3-year decline in marijuana use, and no change in the use of injection drugs. Among the relationships found was that adolescents involved in after-school activities were less likely to use alcohol, tobacco, and drugs. This continuing trend study permits researchers to evaluate the prevention and enforcement efforts directed at the teenage population and to plan future programs.

*Cohort Studies* In a **cohort study,** a *specific* population is followed over a length of time with different random samples studied at various points. Whereas trend studies sample a general population that changes in membership over time, a cohort study samples a specific population whose members do not change over the duration of the survey. Typically, a cohort group has age in common. For example, a school system might follow the high school graduating class(es) of 2004 over time and ask them questions about higher education, work experiences, attitudes, and so on. From a list of all the graduates, a random sample is drawn at different points in time, and data are collected from that sample. Thus, the population remains the same during the study, but the individuals surveyed are different each time.

### Cross-Sectional Surveys

**Cross-sectional surveys** study a cross section (sample) of a population at a single point in time. In a longitudinal study of vocabulary development, for example, a researcher would compare a measure of first-grade students' vocabulary skills in 2000 with one when they were fourth-grade students in 2003 and seventh-grade students in 2006. A cross-sectional study would compare the vocabulary skills of a sample of children from grades 1, 4, and 7 in 2006. The cross-sectional survey is the method of choice if you want to gather the data at one point in time.

**THINK ABOUT IT 14.3**

How would you administer a questionnaire to assess changes in students' political attitudes during college with a (a) cross-sectional approach, (b) panel study, (c) trend study, and (d) cohort study?

**Answers**

a. In the cross-sectional study, you would draw a random sample from each of the four levels and administer the questionnaire to them at the same time.
b. Panel, trend, and cohort studies are all longitudinal. In all three, you first randomly draw a sample of freshmen from your population of interest. In a panel study, you assess your original sample and study the same individuals again when they are sophomores, juniors, and seniors.
c. In the trend study, you draw a random sample of sophomores from the population. A year later, you draw a random sample of juniors, and then in the final year you draw a random sample of seniors.
d. The cohort study would differ from the trend study in that the subsequent samples are drawn only from the population who were enrolled as freshmen when the study began and does not include students who transferred in later.

Longitudinal surveys are more time-consuming and expensive to conduct because the researcher must keep up with the subjects and maintain their cooperation over a long period of time. Cross-sectional surveys, in contrast, do not require years to complete. Hence, they are less expensive. A major disadvantage of the cross-sectional method is that chance differences between samples may seriously bias the results. You may by chance draw a sample of first-graders who are more mature than average and a sample of fourth-graders who are less mature than average, with the result that the difference between the groups appears much smaller than it really is. However, researchers can usually obtain larger samples for cross-sectional studies than for longitudinal studies, and the larger samples mitigate the problem of chance differences.

## SURVEY TECHNIQUE

The survey permits you to gather information from a large sample of people relatively quickly and inexpensively. Conducting a good survey, however, is not as easy as it might initially appear. It requires careful planning, implementation, and analysis if it is to yield reliable and valid information.

### SIX BASIC STEPS INVOLVED IN SURVEY RESEARCH

1. *Planning.* Survey research begins with a question that the researcher believes can be answered most appropriately by means of the survey method. For example, "How do elementary teachers feel about retaining students?" and "What is the extent of tobacco use among the high school students in this district?" are questions that a survey could answer. The research question in

survey research typically concerns the beliefs, preferences, attitudes, or other self-reported behaviors of the people (respondents) in the study. A literature review reveals what other researchers have learned about the question.

2. *Defining the population.* One of the first important steps is to define the population under study. To whom will you distribute the survey? The population may be quite large, or it may be rather limited. For instance, the population might be all elementary teachers in the United States or all elementary teachers in the state of Indiana. Or you might further restrict the population to "all first-year male elementary teachers in the state of Indiana." Defining the population is essential for identifying the appropriate subjects to select and for knowing to whom the results can be generalized.

   Once the population has been defined, the researcher must obtain or construct a complete list of all individuals in the population. This list, called the **sampling frame,** can be very difficult and time-consuming to construct if such a list is not already available.

3. *Sampling.* Because researchers generally cannot survey an entire population, they select a *sample* from that population. It is very important to select a sample that will provide results similar to those that would have been obtained if the entire population had been surveyed. In other words, the sample must be representative of the population. The extent to which this happens depends on the way subjects are selected. The sampling procedure that is most likely to yield a representative sample is some form of probability sampling (see Chapter 7). Probability sampling permits you to estimate how far sample results are likely to deviate from the population values.

4. *Constructing the instrument.* A major task in survey research is constructing the instrument that will be used to gather the data from the sample. The two basic types of data-gathering instruments are interviews and questionnaires.

5. *Conducting the survey.* Once the data-gathering instrument is prepared, it must be field tested to determine if it will provide the desired data. Also included in this step are training the users of the instrument, interviewing subjects or distributing questionnaires to them, and verifying the accuracy of the data gathered.

6. *Processing the data.* The last step includes coding the data, statistical analysis, interpreting the results, and reporting the findings.

Many considerations are involved in implementing the foregoing steps. The balance of this chapter discusses these considerations in detail.

## DATA-GATHERING TECHNIQUES

There are two basic data-gathering techniques in survey research: interviews and questionnaires. Interviews involve some form of direct contact between the people in the sample group and the interviewer (the researcher or someone trained by the researcher), who presents the questions to each person in the sample group and records their responses. When a questionnaire is used, the questions are sent to all the members of the sample group, who record and return their responses to the questions.

### Personal Interviews

In a **personal interview,** the interviewer reads the questions to the respondent in a face-to-face setting and records the answers. One of the most important aspects of the interview is its flexibility. The interviewer has the opportunity to observe the subject and the total situation in which he or she is responding. Questions can be repeated or their meanings explained in case they are not understood by the respondents. The interviewer can also press for additional information when a response seems incomplete or not entirely relevant.

A greater response rate is another obvious advantage of the personal interview. The term **response rate** refers to the proportion of the selected sample that agrees to be interviewed or returns a completed questionnaire. With interviews, response rates are very high—perhaps 90 percent or better. Personal contact increases the likelihood that the individual will participate and will provide the desired information. With mailed questionnaires, the personal contact is missing, and people are more likely to refuse to cooperate. This results in many *nonreturns* (people who do not complete and return the questionnaire). The low response rate typical for a mailed questionnaire (less than 30 percent is common) not only reduces the sample size but also may bias the results (Fowler, 2002). However, an interviewer can get an answer to all or most of the questions. Missing data represent a serious problem for the mailed questionnaire.

Another advantage is the control that the interviewer has over the order with which questions are considered. In some cases, it is very important that respondents not know the nature of later questions because their responses to these questions may influence earlier responses. This problem is eliminated in an interview, in which the subject does not know what questions are coming up and cannot go back and change answers previously given. For individuals who cannot read and understand a written questionnaire, interviews provide the only possible information-gathering technique.

The main disadvantage of the personal interview is that it is more expensive than other survey methods. The selection and training of the interviewers, their salary, and their travel to the interview site make this procedure costly. It takes a great deal of time to contact potential respondents, set up appointments, and actually conduct the interview. Another disadvantage is the possibility of **interviewer bias,** which occurs when the interviewer's own feelings and attitudes or the interviewer's gender, race, age, and other characteristics influence the way questions are asked or interpreted. As a general rule, interviewers of the same ethnic/racial group get the most accurate answers to race-related questions. On other issues, however, two studies found that blacks reported income from welfare and voting more accurately to white interviewers than to black interviewers (Fowler, 2002). The gender of the interviewer may be a factor in surveys of opinions on abortion and gender equality issues. Women talking to women interviewers may express different opinions than they would if the interviewer were male. Researchers should consider the interaction between the subject matter of a survey and the demographic characteristics of the interviewers and respondents. If race, ethnicity, or some other characteristic is very relevant to the answers to be given, then the researcher should consider controlling the relationship of interviewer and respondent characteristics.

Another problem is **social desirability bias,** in which respondents want to please the interviewer by giving socially acceptable responses that they would not necessarily give on an anonymous questionnaire. They may say what they think the interviewer wants to hear. For example, in preference polls in elections involving minority candidates, the proportion of respondents who said they would vote for the minority candidates was often higher than the proportion of votes these candidates actually received in the election. To account for this error, researchers speculate that white voters may have feared they would appear racist if they admitted to interviewers that they preferred a white candidate. Without realizing it, the interviewer also may verbally or nonverbally encourage or reward "correct" responses that fit his or her expectations.

### Focus Groups

A specific category of interviews is the **focus group.** Several subjects are interviewed at the same time. An advantage of a focus group is that participants respond not only to the researcher but also to other participants and their responses. The interaction between participants usually reveals more about the subjects' point of view than would be the case with a researcher-dominated interview. The focus group's interaction enables the researcher to see how subjects incorporate the viewpoints of the others in structuring their own understandings.

This method can provide the researcher with insight into how disagreements are or are not resolved. Sometimes the researcher can report a final consensus.

Focus groups are often used in qualitative research (see Chapter 13). The researcher invites people who are interested in the same general topic to assemble to discuss it. They are assured that they will be free to express themselves in their own words and to respond not only to the researcher but also to other participants and their responses.

In quantitative research, it is very difficult and often very expensive to assemble individuals who will gather in the same place at the same time to respond to the quantitative researchers' predetermined questions. Among those willing to bear the expense of assembling people to focus on a predetermined topic are manufacturers of consumer products who want to learn what product characteristics lead to high sales. Also, politicians often use focus groups to determine what stances might help get them elected or re-elected to public office. You may remember the 2008 U.S. presidential campaigns when some candidates asserted their independence by declaring that their platforms were not based on polls and focus groups.

### Telephone Interviews

The telephone interview is popular, and studies show that it compares quite favorably with face-to-face interviewing. In fact, the past 50 years have seen a gradual replacement of face-to-face interviewing with telephone interviewing as the dominant mode of survey data collection in the United States (Holbrook, Green, & Krosnick, 2003). Its major advantages are lower cost and faster completion, with relatively high response rates. The average response rate may reach 80 percent or higher (Neuman & Kreuger, 2003). Telephone interviews can be conducted over a relatively short time span with people scattered over a large geographic area. For example, national polling organizations often use the telephone to obtain nationwide opinions among voters near election time.

Large-scale surveys in major cities often use the telephone instead of sending interviewers into unsafe areas. The phone permits the survey to reach people who would not open their doors to an interviewer but who might be willing to talk on the telephone. Another advantage is that respondents have a greater feeling of anonymity—and hence there may be less interviewer bias and less social desirability bias than with personal interviews.

The main disadvantage of the telephone interview is that there is less opportunity for establishing rapport with the respondent than in a face-to-face situation. It takes a great deal of skill to carry out a telephone interview so that valid results are obtained. The interviewer often finds it difficult to overcome the suspicions of the surprised respondents. We recommend that you identify yourself right away and explain that you are doing a survey and are not asking for money. An advance letter that informs the potential respondents of the approaching call is sometimes used to deal with this problem, but the letter can induce another problem: The recipient has time to think about responses or to prepare a refusal to participate when the call comes.

Another limitation of telephone interviews is that complex questions are sometimes difficult for respondents to follow. If they misunderstand the questions, the interviewer may not know. It is best that interview questions be short with a limited number of options. The phenomenon of multitasking may affect the quality of telephone interviews. Without the interviewer's knowing, the respondent may be watching television, stirring soup, or writing checks while answering the survey questions. Telephone interviews can be very time-consuming. If the sample is very large, a researcher will need a number of people to help with the interviews. We recommend that telephone interviews be relatively brief.

Another disadvantage is that households without telephones, those with unlisted numbers, and those with cell phones only are automatically excluded from the survey, which may bias results. Almost all U.S. homes now have telephones, so this is not the problem it was years ago. Neuman and Kreuger (2003) state that approximately 95 percent of the population can be reached by telephone. A technique known as "random-digit dialing" solves the problem of unlisted numbers (although it does not reach households without a telephone). In random digit dialing, a computer randomly generates a list of telephone numbers based on all possible numbers thought to be in use in an area. Because of the random determination, this technique ensures that every household with telephone service has an equal chance of being included in the sample. Random-digit dialing has greatly improved the sampling in telephone surveys. See Fowler (2002) for a thorough discussion of random-digit dialing.

Other limitations of telephone surveys arise from new technology that may make it increasingly difficult to reach potential respondents by phone. Services such as caller identification, phone number blocking, and similar procedures enable residential phone customers to have much greater control over incoming calls. People may simply ignore calls from the unfamiliar number of the surveyor and telephone response rates may continue to drop.

### Computer-Assisted Telephone Interviewing (CATI)

Computer and telecommunications technology has been applied to telephone surveys. Wearing earphones, the interviewer sits at a computer while it randomly

selects a telephone number (through random-digit dialing or from a database) and dials. When the respondent answers, the interviewer reads the first question that appears on the computer screen and types the answer directly into the computer. The computer program displays the next screen containing the next question, and so on through the entire survey. Using CATI saves a great deal of time. The surveyor can fill in forms on a computer screen or type answers to open-ended questions very quickly. The major advantage is that CATI software immediately formats responses into a data file as they are keyed in, which saves the researcher time usually spent in coding and manually transferring responses from paper into the computer for analysis.

### Conducting the Interview

Whether the interview is conducted in person or by telephone, the interviewer's main job is to ask the questions in such a way as to obtain valid responses and to record the responses accurately and completely. The initial task for the interviewer is to create an atmosphere that will put the respondent at ease. After introducing yourself in a friendly way, briefly state the purpose of the interview but avoid giving too much information about the study, which could bias the respondent. It is well to begin the interview with fairly simple, nonthreatening questions.

The interviewer also has the responsibility of keeping the respondent's attention focused on the task and for keeping the interview moving along smoothly. This can best be done if you are thoroughly familiar with the questions and their sequence so that you can ask the questions in a conversational tone and without constantly pausing to find what question is coming next. Of course, you must refrain from expressing approval, surprise, or shock at any of the respondent's answers.

Interviews can be more or less structured. In a less structured interview, the same questions are asked of all respondents, but the interview is more conversational and the interviewer has more freedom to arrange the order of the questions or to rephrase the questions. If comparable data are to be obtained, however, the interviewer must standardize the procedure by using a structured interview schedule. A structured interview schedule contains specific questions in a fixed order, to be asked of all respondents, along with transition phrases and **probes** (questions used to clarify a response or that push a little further into a topic). For example, if the respondent starts to hedge, digress, or give irrelevant responses, or if he or she has obviously misinterpreted the question, then the interviewer may use a fixed probe such as "Explain your answer a little further" or "Can you tell me a little more about that?" Another important technique besides the probe is the **pause.** A good interviewer needs skill in listening and is quiet at times until the respondent answers. In less structured interviews, any marked deviations from the protocol should be documented so that the information can be taken into account when analyzing the interviewee's response. In using probes, take care not to suggest or give hints about possible responses. It takes less training time to teach interviewers to administer a structured interview than it does an unstructured one because everything they need to say or do is contained in the interview schedule. For this reason, the structured interview is the most widely used format for large studies with numerous interviewers.

### Training the Interviewer

It is essential that potential interviewers receive training before being sent out to conduct interviews. Quality of interviewers is probably one of the least appreciated aspects of survey research. "Interviewers have a great deal of potential for influencing the quality of the data they collect" (Fowler, 2002, p.117).

There are certain aspects of interviews that need to be standardized and, therefore, they should always be included in interviewer training: (1) procedures for contacting respondents and introducing the study, (2) instructions on asking questions so that interviewers ask all questions in a consistent and standardized way, (3) procedures for probing inadequate answers in a nondirective way, (4) procedures for recording answers to open-ended and closed-ended questions, and (5) rules for handling the interpersonal aspects of the interview in a nonbiasing way (Fowler, 2002). To be able to answer respondents' questions, interviewers should also know the purpose of the project, the sponsorship, the sampling approach used, and the steps that will be taken with respect to confidentiality.

Interviewer trainees should be provided with written manuals on interviewing procedures. They should observe interviews being conducted by trained individuals and should be supervised in conducting practice interviews. In the practice interviews, the interviewees should be individuals drawn from the same population that will be used in the research project.

### Mailed Questionnaires

The direct one-on-one contact with subjects in a personal interview is time-consuming and expensive. Often, much of the same information can be obtained by means of a **questionnaire** mailed to each individual in the sample, with a request that he or she complete and return it by a given date. Because the questionnaire is mailed, it is possible to include a larger number of subjects as well as subjects in more diverse locations than is practical with the interview.

A **mailed questionnaire** has the advantage of guaranteeing confidentiality or anonymity, thus perhaps eliciting more truthful responses than would be obtained with a personal interview. In an interview, subjects may be reluctant to express unpopular or politically incorrect points of view or to give information they think might be used against them at a later time. The mailed questionnaire also eliminates the problem of interviewer bias.

A disadvantage of the mailed questionnaire is the possibility of respondents misinterpreting the questions. It is extremely difficult to formulate a series of questions whose meanings are crystal clear to every reader. The investigator may know exactly what is meant by a question, but because of poor wording or different meanings of terms, the respondent makes a significantly different interpretation. Furthermore, large segments of the population may not be able to read or may read only in another language and may not be able to respond to a mailed questionnaire. Only people with considerable education may be able to complete a very complex questionnaire.

Another important limitation of mailed questionnaires is the low return rate. It is easy for the individual who receives a questionnaire to lay it aside and simply forget to complete and return it. A low response rate limits the generalizability of the results of a questionnaire study. It cannot be assumed that nonresponse is randomly distributed throughout a group. Studies have shown that there are

usually systematic differences in the characteristics of respondents and nonrespondents to questionnaire studies. Response rate is often higher among the more intelligent, better educated, more conscientious, and those more interested or generally more favorable to the issue involved in the questionnaires. The goal in a questionnaire study is 100 percent returns, although a more reasonable expectation may be 40 to 75 percent returns.

A number of factors have been found to influence the rate of returns for a mailed questionnaire, including (1) length of the questionnaire, (2) cover letter, (3) sponsorship of the questionnaire, (4) attractiveness of the questionnaire, (5) ease of completing it and mailing it back, (6) interest aroused by the content, (7) use of a monetary incentive, and (8) follow-up procedures used. We discuss these factors in more detail later.

### Electronic Mail Surveys

As computers have become common, researchers have used electronic mail (e-mail) to deliver questionnaires. Dillman (2000) found that e-mail surveys have the advantage of prompter returns, lower item nonresponse, and more complete answers to open-ended questions. **Electronic mail surveys** can be completed at a pace the respondents choose, and they cannot be mislaid like a mail survey. The main disadvantage is that they are appropriate only when the researcher has e-mail addresses for all members of a finite population, such as all elementary teachers in a given school district or all members of a local union or fraternal lodge. Research shows that some of the factors found to be important for regular mail surveys are also important for e-mail surveys. For example, people who received a prior e-mail notification about the survey were more likely to respond; also, surveys addressed individually to a person (rather than being part of a mailing list) had higher response rates.

### Internet Surveys

The Internet has become a popular methodology for survey research. The questionnaire is placed on a website constructed by the investigator. Respondents are able to answer the questions and submit the questionnaire online. **Web-based surveys** have a number of advantages. They have the potential of reaching large populations and permit the collection of larger amounts of data than would be possible with traditional survey methods. They can be conducted quickly and easily and are less expensive than mailed surveys. The cost advantage increases as the size of the sample increases. Dillman (2000) states, "Once the electronic data collection has been developed, the cost of surveying each additional person is much less, compared with both telephone interview and postal procedures. In some instances these technologies may result in decisions to survey entire populations rather than only a sample" (p. 353). Another important benefit is in the processing of survey data. Web-based surveys can significantly reduce the amount of time and effort and the costs associated with getting the data into a system for analysis. Furthermore, because they are available 24 hours a day, respondents can reply when and where they choose. In a University of Colorado survey, 55 percent of respondents cited ease of use as one of the things they liked most about answering a web survey (Cook, Heath, & Thompson, 2000).

The major limitation of Internet surveys is that samples are restricted to those with access to the technology and who choose to respond. Samples are dominated by relatively affluent, well-educated, urban, white-collar, technically sophisticated young males (Flatley, 2001). The large number of potential responses in a web survey does not overcome the problem of sampling error due to lack of representativeness. We need to be sure the survey is reaching the desired respondents.

There is somewhat conflicting evidence about response rates in web-based surveys. The consensus, however, is that response rates for web surveys are lower than rates obtained by other methods. A meta-analysis of 68 web surveys found a mean response rate of 39.6 percent (Cook et al., 2000). Dillman and Bowker (2001) compared response rates to a questionnaire administered by telephone, mail, or Internet. The response rate for the Internet was 50 percent compared with 80 percent for telephone and mail surveys. To generate enthusiasm and maximize response, one should send an introductory letter separate from the instrument that explains what the survey is about, that requests their cooperation, and that provides an incentive for completing the survey. The problem, however, is that the introductory letter is usually sent by e-mail and may be deleted before it is ever read by the potential respondent. Because of the relative anonymity provided by the Internet, it is more difficult to determine if respondents are who they say they are. It would also be possible for people to use different identities and respond more than once to a survey.

Meister and Melnick (2003) conducted an Internet survey of first- and second-year teachers to identify their concerns in the areas of classroom management, time management, communication with parents, and academic preparation. Ten flyers, each containing a request for new teachers to visit the website and respond to the survey, were sent to 1000 principals in all 50 states. They were directed to distribute the flyers to first- and second-year teachers. A second mailing via e-mail was sent to 500 principals. They were asked to forward the e-mail to any first- or second-year teachers in their building. A total of 273 teachers from 41 states responded. Sixty percent of the respondents were elementary teachers, 27 percent high school, and 13 percent middle school. Rural, suburban, and urban school districts were all represented. The responses were immediately recorded in a database as teachers answered items on the website.

The findings showed that new teachers were less confident of their knowledge and skills in the areas of discipline, time management, and communication skills. They reported needing assistance in handling disruptive students and those with special needs. Eighty-four percent reported that they felt sometimes "overwhelmed" by the paperwork and other noninstructional demands on their time. One in four teachers did not feel well prepared by the student teaching experience, especially in the areas of reading and language arts. Forty percent of beginning elementary teachers responded that they were not prepared to teach reading.

This research example illustrates the advantages and disadvantages of Internet surveys. The survey had the potential of reaching teachers over a wide geographic area and permitted them to respond at their convenience. The fact

that the data were put into the system ready for analysis saved time and effort for the researchers. The disadvantage is that the response rate appears to be low, but actually we cannot calculate the response in the usual way. There is no way of knowing how many teachers actually received a flyer or the e-mail message about the survey.

The reader is referred to DeVaus (2002), who discusses a variety of applications for web-based surveys, as well as tips on designing the questionnaire.

### Directly Administered Questionnaires

A **directly administered questionnaire** is given to a group of people assembled at a certain place for a specific purpose. Examples include surveying the freshmen or their parents attending summer orientation at a university. Surveys at universities are often administered in classrooms or in residence halls. For example, the annual survey of freshmen coordinated by UCLA that was referred to previously is administered during orientation at the various colleges and universities.

The main advantage of directly administering questionnaires is the high response rate, which typically is close to 100 percent. Other advantages are the low cost and the fact that the researcher is present to provide assistance or answer questions. The disadvantage is that the researcher is usually restricted in terms of where and when the questionnaire can be administered. Also, when a population is limited (e.g., parents of freshmen in a specific university), the results of the survey will be equally limited in terms of generalizability.

As we have seen, researchers have a choice among several data collection methods for survey research. They choose the one best suited for their particular study. However, a multimode approach in which researchers use combinations of these methods in the same study is quite common. In fact, Fowler (2002) states that mixing modes is one of the best ways to minimize survey nonresponse because it enables researchers to reach people who are inaccessible via a single mode. Table 14.1 provides a summary of the advantages and disadvantages of the various data collection methods.

## STANDARD ERROR OF THE SAMPLING PROPORTION

We have seen in Chapter 7 that even with random sampling there will always be some error in estimating a population parameter from sample statistics. The statistic most commonly reported in a sample survey is a proportion or a percentage of the sample that gives a particular response. The discrepancy between the known sample proportion and the unknown population value is referred to as **sampling error.** The first step in assessing how much sample results are likely to deviate from the population values is to calculate the standard error of the sampling proportion.

### CALCULATING THE STANDARD ERROR

Using the obtained sample proportions, you can calculate the variance and then the standard error of the proportion. The variance of a proportion is $pq$, where $p$ is the proportion agreeing or having a certain characteristic, and $q(1 - p)$ is the

**Table 14.1** Comparison of Data Collection Methods

| Method | Advantages | Disadvantages |
|---|---|---|
| Personal interview | Researcher is present<br>Can establish rapport<br>Flexibility<br>High response rate<br>Fewer incomplete answers<br>Good for surveys on complicated issues<br>Time for thoughtful answers | Time-consuming<br>More expensive<br>Interviewer bias<br>Social desirability bias<br>Need trained interviewers<br>No anonymity<br>Less safe for researchers |
| Telephone interview | Low cost<br>Convenient<br>High response rate<br>Quick<br>Greater safety for interviewers and respondents | Less rapport<br>Individuals may refuse to talk<br>Questions need to be short and simple<br>Limited to those with telephones |
| Mailed questionnaire | Low cost<br>Allows anonymity<br>No interviewer bias<br>Convenient<br>Nonthreatening | Slow<br>Low response rate and nonresponse bias<br>Cannot clarify question<br>Missing data<br>Literacy required |
| Electronic or Internet questionnaires | Low cost<br>Easy for researcher<br>Convenient for respondent<br>Allows anonymity<br>Potential of quick response<br>Time for thoughtful answers<br>Low social desirability bias<br>Greater safety for researcher and respondent | Limited to computer savvy<br>Difficulty getting cooperation<br>Potentially lower response rate<br>Literacy requirement<br>May not be able to identify respondents |
| Directly administered questionnaire | Low cost<br>High response rate<br>Researcher present | Not as flexible with respect to time and place |

proportion not agreeing or not having the characteristic. If the variance for the proportion is *pq,* the formula for the standard error of the proportion is

$$\mathrm{SE}(\sigma_p) = \sqrt{\frac{\mathrm{Var}}{n}} = \sqrt{\frac{pq}{n}} \tag{14.1}$$

where

$\mathrm{SE}$ or $\sigma_p$ = standard error of the proportion
$p$ = proportion agreeing
$q$ = proportion not agreeing $(1 - p)$
$n$ = size of the sample

For example, assume you survey a random sample of 100 parents of prekindergarten children in a school district and ask if they are in favor of full-day kindergarten in the district. You find that 80 say "Yes" to this question and 20 say "No." The standard error of the proportion is

$$\mathrm{SE} = \sigma_p = \sqrt{\frac{(.80)(.20)}{\sqrt{100}}} = \sqrt{\frac{.1600}{100}} = \sqrt{.0016} = .04$$

From probability theory, we say that we are 95 percent confident that the sample statistic is within 1.96 standard errors of the parameter.

In our example, standard error of .04 multiplied by 1.96 gives us .0784. Rounding that to .08 and expressing it as a percent, we have 8 percent. Thus, we would report that 80 percent of parents of prekindergarten children are in favor of a full day with a **margin of error** of 8 percent. This means that we have 95 percent confidence that the population proportion is between 72 and 88 percent.

### THINK ABOUT IT 14.4

A community survey asked the question, "Are you in favor of having a Wal-Mart in our community?" From the 899 completed surveys, it was found that 63 percent of the residents answered "Yes" to this question. Calculate the margin of error and set up the 95 percent confidence interval for these data:

**Answer**

$$\sigma_p = \sqrt{\frac{(.63)(.37)}{899}} = \sqrt{\frac{.2331}{899}} = \sqrt{.000259} = .0161$$

$1.96 \times .0161 = .0316$ or a 3 percent margin of error. The confidence interval would be $63 \pm 3$. One would be 95 percent confident that the percent in the population favoring a Wal-Mart in the community is between 60 and 66 percent. In surveys conducted by the major polling organizations, a margin of error of 3 percent is generally considered acceptable.

## CALCULATING SAMPLE SIZE

How large a sample to draw is one of the early questions researchers must answer. We know that sampling error decreases as the size of the sample increases. Once you decide on an acceptable margin of error, you can determine the sample size needed to achieve that margin of error by applying Formula 14.2.

$$n = \left( \frac{\frac{1}{E}}{\sqrt{pq}} \right)^2 (z)^2 \tag{14.2}$$

where

$n$ = sample size needed
$E$ = desired margin of error
$pq$ = variance of hypothesized proportion
$z$ = $z$ score of confidence level

For example, if we want a $\pm 3$ percent margin of error at the 95 percent confidence level ($z = 1.96$) for a hypothesized $p = .5$, $q = .5$, the required $n$ would be

$$n = \left( \frac{\frac{1}{.03}}{\sqrt{(.5)(.5)}} \right)^2 (1.96)^2 = \left( \frac{1}{.06} \right)^2 (3.8416) = (277.7783\ 3)(.8416)$$

$$n = 1067.1131$$

If the hypothesized population proportions are .50/.50, then 1068 subjects would be needed to yield a margin of error of 3 percent. Remember that this is 1068 usable responses. If you are using a mailed questionnaire and anticipate a 50 percent return, you would need to mail out twice that number of questionnaires. If a 10 percent margin of error would be acceptable, then only 97 respondents would be needed.

$$n = \left( \frac{\frac{1}{10}}{\sqrt{(.5)(.5)}} \right)^2 (1.96)^2 = \left( \frac{1}{.2} \right)^2 (3.8416) = (25)(3.8416) = 96.04$$

Using $p = .50$ is always a safe way to calculate the needed sample size because $pq$ is at its maximum possible value when $p$ and $q$ both equal .50 and $pq = .25$. Using population estimates of $p = q = .50$ in Formula 14.2, the researcher will be confident that the margin of error will be as small as or smaller than the value specified. If there are compelling reasons to hypothesize unequal population proportions for $p$ and $q$, these may be substituted for .50 and .50, respectively, and the result will be a lower required number of subjects. For instance, if 20 percent of the high school boys in a school district tried out for intramural sports last year, you might use $p = .20$ and $q = .80$ in calculating the number of subjects needed for a survey of this year's boys. Note that all these statistics are based on the assumption that the sample is a random sample of the population of interest. The credibility of the results is a function of the validity of that assumption. Table 14.2 shows the sample size needed to have a certain margin of error for a given $p$. Note that the largest sample is required when the desired margin of error is lowest and $p = .50$. As the acceptable margin of error increases and $p$ varies from .50, the required sample size decreases.

As with anything else that relies on sampling, the representativeness of the sample is more important than the size of the sample. An unrepresentative sample leads to inappropriate conclusions regardless of its size. For example, a university wanted to determine the need for evening cafeteria hours. It conducted a random survey of all full-time students enrolled at the university. However, the majority of the students taking evening classes and thus potentially more likely to use the cafeteria in the evening were part-time, not full-time, students. A survey of full-time students could lead to inaccurate conclusions about the demand for evening cafeteria hours.

**Table 14.2** Minimum Sample Sizes Required for Various Margins of Error around the Parameter Estimation at the .95 Confidence Level

| Maximum Margin of Error (%) | Value of p | | |
|---|---|---|---|
| | .10 or .90[a] | .25 or .75 | .50 |
| 1 | 3462 | 7212 | 9616 |
| 2 | 866 | 1803 | 2404 |
| 3 | 385 | 802 | 1069 |
| 5 | 139 | 289 | 385 |
| 10 | 35 | 73 | 97 |

[a]The sample variance, $pq$, is the same when $pq = (.10)(.90)$ as when $pq = (.90)(.10)$.

Another factor influencing the decision about sample size is the variability of the population to be sampled. If the population is fairly homogeneous, then you can use a smaller sample than if the population is more variable. A population of college students, for example, would be expected to show less variability than a population of adults in general, and thus a smaller sample could be used. Check the published research studies to learn about the variability of population of interest and the sample size used in survey research on the population.

Sample size does not need to be a certain percentage of the population. When random sampling is used, a sample size that is only a small percentage of the population can represent the population well. The major public opinion polls, for example, do not use really large samples. The Nielsen rating service uses a sample of only approximately 5000 television-watching households in the United States to estimate the popularity of various programs. The Nielsen ratings are a major variable in determining which programs continue and which do not.

## CONSTRUCTING THE INSTRUMENT

### FORMAT OF QUESTIONS

Once the overall research question has been determined, the next task is to construct an instrument that will provide the desired information. Because survey data consist of peoples' responses to individual questions, it is essential to start with good questions. Two basic types of questions are used in survey instruments: closed-ended or fixed alternative and open-ended or free-response questions. Use **closed-ended questions** when all the possible, relevant responses to a question can be specified, and the number of possible responses is limited. For example, in a survey of undergraduates, a question about class level would be closed-ended. The possible answers are known and are few in number: *freshman, sophomore, junior,* or *senior*. Other closed-ended questions might ask about residency status (in-state or out-of-state) or sex. **Open-ended questions** are used when there are a great number of possible answers or when the researcher cannot predict all the possible answers. For example, a question about the students' reasons for selecting a particular university would probably be open-ended. A question about the college major would be open-ended because the researcher would probably not want to include the long list of possible majors. Both formats can be used in the same question—that is, a number of closed-ended responses can be followed by "Other" as the last possible response. For example, you might ask a professor a question such as follows:

1. What type of writing assignments do you typically require in your course? (circle as many as apply)
   a. Reports
   b. Themes or essays
   c. Research papers
   d. Take-home essay examinations
   e. Minute papers
   f. Other (please specify)

There are advantages and disadvantages to both these question formats. The open-ended question permits a free response rather than restricting the respondent

to a choice from among stated alternatives. Individuals are free to respond from their own frame of reference. Open-ended questions are easier to construct, but analyzing them is very tedious and time-consuming. The researcher must read and interpret each response and then develop a coding system that will permit a quantitative analysis of the responses. Some responses may be unclear, and the researcher may be unsure how to classify or code the response. The responses to open-ended questions typically differ in length, and some respondents may give more than one response to a particular question. If asked the most important reason for choosing a certain university, a respondent might answer, "I chose University X because of its academic reputation and because it is in-state and less expensive for me." The researcher must then decide whether to use either both answers or only the first one, assuming the first answer is the more important one. An option to consider with such a question is rank ordering, using a combination of closed- and open-ended responses or using new technology to directly enter responses electronically.

Closed-ended questions take more time to construct, but the responses are easier to tabulate. Responses to closed-ended questions can be coded directly on scannable sheets that can be "read" and the data entered into a computer database for analysis. Closed-ended questions can be answered more easily and quickly by respondents. A closed format also ensures that all subjects have the same frame of reference in responding and may also make it easier for subjects to respond to questions on sensitive or private topics.

A limitation of the closed-ended question is that it does not provide much insight into whether respondents really have any information or any clearly formulated opinions about an issue. It is easier for the uninformed respondent to choose one of the suggested answers than to admit to lack of knowledge on an issue. For example, in response to the question, "What is the effect of outsourcing on the U.S. economy?" the respondent who has little knowledge of outsourcing and the reasons for it could easily select a reasonable answer from among the alternatives provided. In contrast, respondents who have the knowledge or who have well-informed opinions on the issue may dislike being restricted to simple response categories that do not permit them to qualify their answers. It is possible to get the benefits of both open-ended and closed-ended questions. A researcher can first use the open-ended format with a small sample to identify possible alternative responses to the questions. The researcher can then list the most frequent responses, followed by "OTHER (please specify)".

## STRUCTURE OF QUESTIONS

We have discussed the advantages and disadvantages of the closed-ended and open-ended questions in survey research. Now let us look at the ways these two formats can be used to structure questions for interviews and questionnaires.

1. **Completion,** or **fill-in, items** are open-ended questions to which respondents must supply their own answers in their own words. For example, "What is the major weakness you have observed in your students' preparation for college?"

2. **Checklists** are questions that present a number of possible answers, and the respondents are asked to check those that apply. For example,

   What type of teaching aids do you use in your classes? (check as many as apply)

   _____ 1) CHALKBOARD
   _____ 2) OVERHEAD PROJECTOR
   _____ 3) COMPUTER PROJECTOR
   _____ 4) VIDEOTAPES
   _____ 5) OTHER (please specify)

3. **Scaled items** ask respondents to rate a concept, event, or situation on such dimensions as quantity or intensity, indicating "how much"; on quality, indicating "how well"; or on frequency, indicating "how often." For example,

   How would you rate the writing skills of students you are teaching this semester? (check one)

   _____ 1) VERY POOR
   _____ 2) LESS THAN ADEQUATE
   _____ 3) ADEQUATE
   _____ 4) MORE THAN ADEQUATE
   _____ 5) EXCELLENT
   _____ 6) INSUFFICIENT INFORMATION

   or

   How well prepared in basic math skills are the students who typically enroll in your course? (check one)

   _____ 1) NOT AT ALL PREPARED
   _____ 2) SOMEWHAT PREPARED
   _____ 3) WELL PREPARED
   _____ 4) EXTREMELY WELL PREPARED

4. **Ranking items** ask respondents to indicate the order of their preference among a number of options. Rankings should not involve more than six options because otherwise it becomes too difficult for respondents to make the comparisons. An example of a ranking item follows:

   Please rank the order of difficulty your students have in reading each of the following materials, with 1 being the most difficult and 4 the least difficult.

   _____ 1) TEXTBOOKS
   _____ 2) OTHER REFERENCE BOOKS
   _____ 3) JOURNAL ARTICLES
   _____ 4) OTHER (please specify)

5. **Likert-type items** let subjects indicate their responses to selected statements on a continuum from *strongly agree* to *strongly disagree*. Likert-type scales were discussed in Chapter 8. An advantage of this type of item is that points can be assigned to the various responses, and thus measures of

central tendency, variability, correlation, and the like can be calculated. For example,

> The students who typically enroll in my course are underprepared in basic math skills. (circle one)
>
> strongly agree agree undecided disagree strongly disagree

## THINK ABOUT IT 14.5

Assume you are conducting a survey of students in the local high school on the extent of alcohol and drug use among students in the school. Construct one item of each of the following types that you might include in this survey: (1) completion, (2) checklist, (3) scaled item, (4) ranking, and (5) Likert type.

**Suggested Answer**

**Completion**

How would you describe the extent of alcohol use in your school?

**Checklist**

Are you (check one)

_____ 1) Male

_____ 2) Female

**Scaled item**

On average, how often do the students that you know use alcohol? (check one)

_____ 1) Twice a week

_____ 2) Once a week

_____ 3) Twice a month

_____ 4) Once a month

**Ranking**

How would you rank the following in terms of extent of use among students you know? (1 most common and 5 least common)

_____ 1) Beer

_____ 2) Wine

_____ 3) Other alcoholic beverages

_____ 4) Marijuana

_____ 5) Hard drugs

**Liken-type item**

Alcohol consumption is a serious problem among students in this school. (circle one)

SA A U D SD

## WRITING SURVEY QUESTIONS

Now let us examine the task of formulating good questions. Deciding how to word the questions that are asked in a survey is a challenge. Before beginning to write a structured set of survey questions, it can be helpful to have a focus group discuss the questions in a nonstructured form. You do this by bringing together a few groups of 5 to 10 people representative of the study population to discuss the topics covered in the survey. A moderator keeps the discussion focused on a preset agenda and asks questions to clarify comments.

A focus group can be very valuable in questionnaire development. Focus group discussions help the researcher understand how people talk about the survey issues, which is helpful in choosing vocabulary and in phrasing questions. A focus group can often suggest issues, concerns, or points of view about the topic that the researcher had not considered.

A significant body of research has shown that changes in such things as phrasing, the amount of information offered, and the choice of answers available to respondents can influence the outcome of a survey to a greater or lesser degree. The way you ask a question may prescribe the answer. It is possible for surveyors with competing agendas to come up with entirely different responses. Consider the following three ways to ask a question:

1. Would you support increased taxes to pay for full-day kindergarten?
2. Would you support an increase in your property taxes to pay for full-day kindergarten?
3. Would you support a 5 percent increase in your property taxes to pay for full-day kindergarten?

Individuals who might agree with the first general question would not necessarily agree when they consider that their property taxes would be increased.

Or consider the following ways to ask opinions on abortion:

1. Do you believe in women's right to choose?
2. Do you believe women should be allowed to decide whether to continue a pregnancy?
3. Do you believe women should be allowed to kill their unborn child?

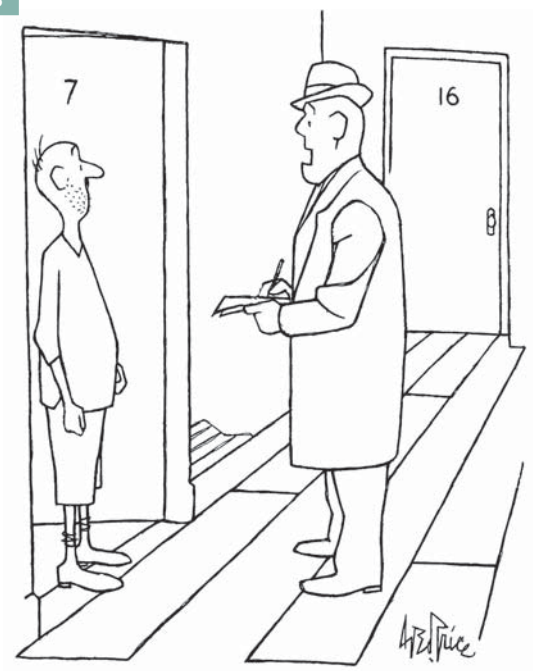
The following are basic guidelines for writing good questions:

1. *Questions should be short, simple, and direct.* Eliminate any words and phrases not essential to the clear meaning of the question. Short questions are easier to understand. Most of the questions should have fewer than 10 words (one line); none should have more than 20.
2. *Phrase questions so that they can be understood by every respondent.* The vocabulary used should be nontechnical and should be geared to the least educated respondent. At the same time, avoid talking down to respondents or choosing words that sound patronizing. It is a good idea to have some other people, preferably ones whose background is similar to those who will be included in the study, read and give their interpretation of the content of each question. For example, questions using terms such as "authentic assessment," "distance learning," and "total quality management" may not be appropriate in a survey designed for the general public. Also, be careful not to use slang, abbreviations, or acronyms that may not be familiar to all.
3. *Phrase questions so as to elicit unambiguous answers.* The question, "Did you vote in the last election?" is ambiguous because it does not specify which election. Quantify responses whenever possible. Words such as *often* and *sometimes* have different meanings for different people. For example, in a survey on how often undergraduates use the main library for studying, the responses should be quantified (*daily, five times per week, twice per week,* and so on) instead of using responses such as *usually, sometimes,* and *often.*
4. *Phrase questions so as to avoid bias that may predetermine a respondent's answer.* The wording of a question should not influence the respondent in a certain direction. For this reason, avoid stereotyped, prestige-carrying,

emotionally loaded, or superlative words. Some words have such an emotional appeal in U.S. culture that they tend to bias questions regardless of how they are used. For example, the wording "Have you exercised your American right and registered to vote?" would undoubtedly bias the question. Simply asking, "Are you registered to vote?" would be preferable. Dillman (2000) states that words such as *freedom, equality, private enterprise, justice,* and *honesty* have a strong positive appeal in our culture. Words such as *bureaucratic, socialist, boss,* and *government planning* have a strong negative appeal. Avoid such words if possible.

5. *Avoid questions that might mislead because of unstated assumptions.* The frame of reference for answering the questions should be clear and consistent for all respondents. If any assumptions must be made before respondents give an answer, then also include questions designed to inquire into these assumptions. For example, in a survey designed for high school seniors, the question, "Do you think your high school has adequately prepared you for college?" assumes the student is going to college and knows what is required in the way of preparation. The question, "Have you registered to vote for the next presidential election?" assumes the high school student is 18 years of age, which may not be true.
6. *Avoid leading questions, which imply a desired response.* For example, "What do you think of the biased coverage of the Iraq War by the major TV networks?" is a leading question.
7. *Avoid questions that may elicit embarrassment, suspicion, or hostility in the respondent.* Questions should not put the respondent on the defensive. For example, people often resent questions about their age, income, religion, or educational status. Instead of asking a subject's age, the researcher can ask for his or her year of birth. People seem less concerned about giving their year of birth than about giving their age. The question, "Do you have a high school diploma?" may embarrass someone who did not graduate from high school. The question might instead ask, "What grade had you completed when you left school?" In fact, it is best to avoid personal questions entirely unless the information is essential to the research.
8. *Avoid* ***"double-barreled" questions****, which attempt to ask two questions in one.* For example, "Do you feel that the university should provide basic skills courses for students and give credit for those courses?" is a double-barreled question. When a respondent answers such a question, the researcher does not know whether the answer applies to both parts of the question or just to one. A yes answer to the preceding question may mean either that the respondent believes the university should offer basic skills courses and give credit for them or that it should offer the courses but not give credit for them. You can identify a double-barreled question from noting the *and, or,* or some other conjunction in the wording.
9. *Make sure the alternatives to each questionnaire item are exhaustive—that they express all the possible alternatives on the issue.* For example, "What is your marital status?" should include not only the alternatives "married" and "single" but also "unmarried cohabiting," "widowed," "divorced," and "separated." In developing the alternatives for questionnaire items designed

**PICTURE THIS**




*"Next question: I believe that life is a constant striving for balance, requiring frequent tradeoffs between morality and necessity, within a cyclic pattern of joy and sadness, forging a trail of bittersweet memories until one slips, inevitably, into the jaws of death. Agree or disagree?"*



to identify attitudes or opinions on issues, it is a good idea first to present the questions in an open-ended form to a small sample of respondents. Their answers can then be used as alternatives in the final product. On questions with a wide variety of possible responses, always include the alternative "other," along with a request that the respondent explain that choice. The question, "What is your position in the school system?" might be followed by the alternatives "administrator," "teacher," "librarian," and "other (please specify)."

10. *Keep the questionnaire as brief as possible so that it requires a minimum of the respondents' time.* Respondents are much more likely to complete and return a short questionnaire. The researcher must make an effort to eliminate all unnecessary items, especially those whose answers are available from other sources. All the items of a questionnaire should serve a research problem function; that is, they should elicit data needed to test the hypotheses or answer the questions of the research study. For example, you can eliminate a question that asks the respondent's age in a study where this information is not needed in the data analysis.
11. M*ake sure the respondents have the information necessary to answer the questions.* Avoid questions dealing with experiences or topics you know are unfamiliar to your sample.

See Fowler (2002) for additional suggestions on writing survey questions.

**THINK ABOUT IT 14.6**

Which of the basic guidelines for writing good questions are violated in the previous cartoon?

**Answer**
1, 2, 3, 8, 10, 11

## USING A MAILED QUESTIONNAIRE

As we discussed previously, it is not always practical to obtain survey data using an interview format—in these cases questionnaires are used, and quite often they are mailed to respondents. A well-constructed questionnaire is an important factor influencing response rate. We have already offered guidelines for writing the questions. In this section, we discuss the overall arrangement of the questions, the cover letter, the follow-ups, and other factors that contribute to the success of a mailed questionnaire.

### DIRECTIONS

It is very important to begin with precise directions that tell the respondents exactly what to do. Indicate how and where they are to mark their responses; for example, "Please indicate your response to the following questions by placing an X in the box next to the answer of your choice" or "Please use a No. 2 pencil and indicate your answers by blackening in the appropriate bubbles on the separate scannable answer sheet." If the format changes within the questionnaire, include new directions for that section.

### ORDER OF QUESTIONS

Once the questions are written, they must be arranged in an appropriate order. The order of questions is important because it can influence the respondent's interest in completing the questionnaire. The very first question should be especially interesting and easy enough for all respondents to interpret and answer. If respondents are motivated to answer the first question, they are more likely to continue with the questionnaire. The first question should seek worthwhile information that is clearly related to the topic under consideration. For this reason, never begin a questionnaire with questions relating to age, gender, education, occupation, ethnic origin, marital status, and the like. The respondents may regard these questions as irrelevant or as an invasion of privacy and may therefore decide not to continue with the questionnaire. It could end up in the nearest wastebasket. Researchers also recommend that the first few questions be of the closed-ended type, which the respondent can complete quickly, instead of open-ended ones that may require a long, written response.

Group together questions that are similar in content. For example, in a questionnaire asking university faculty about the basic academic skills of their students, all the questions on reading would be placed together. Then questions on writing would appear together, followed later by questions related to mathematics

skills. Within the content areas, group items according to the type of question. For example, the questions requiring a simple yes or no answer would be placed together, as would items requiring respondents to rank or to indicate the extent of agreement or disagreement.

Within each of the topic areas, arrange the questions in good psychological order. A logical or psychological arrangement contributes to better thought-out answers on the part of the respondents. For example, first ascertain whether respondents were satisfied with working conditions before asking them to recommend changes. If both general and specific questions are asked on a topic, place the general questions first. Objective items on an issue or situation should precede the more subjective questions. Questions that are less likely to be objectionable should precede items that are more objectionable. People are sometimes reluctant to answer questions about attitudes, preferences, motives, behavior, personal feelings, and the like, but if objective questions can be used first to clarify and specify the situation, it may be easier for individuals to respond. For example, a researcher who wanted to survey students on the extent of marijuana usage might begin by asking more objective questions first, such as "How would you describe marijuana usage in your school: serious problem, moderate problem, slight problem, or no problem?" This could be followed by the questions, "Do you think the frequency of marijuana smoking has increased, stayed about the same, or decreased this year?" and "Do you know students who use marijuana?" Then perhaps "Do you ever smoke marijuana?" could be asked. This principle of placing less objectionable questions before more objectionable ones implies that items dealing with demographic data such as age, gender, and occupation should be placed at the end of the questionnaire rather than at the beginning. The respondent will have fewer objections to giving this personal information after completing the questionnaire and understanding why such data would be relevant.

## MATRIX SAMPLING

A procedure called **matrix sampling** is sometimes used when the survey is long and the accessible population is large. This technique involves randomly selecting respondents, each of whom is administered a subset of questions, randomly selected from the total set of items. The practical advantage of using matrix sampling is the decrease in the time required for each individual to respond. This is an important advantage because one obstacle to obtaining a high response rate is the unwillingness of some individuals to take the time to answer a long questionnaire.

## FORMAT OF THE QUESTIONNAIRE

The questionnaire should be formatted so that it is attractive, easy for the respondent to read and answer, and convenient for the researcher to code and score. To achieve these ends, incorporate the following suggestions into the design of the survey document:

1. *Number questions consecutively* throughout the questionnaire without any repetitions or omissions. Having a unique number for each question avoids confusion in coding responses.

2. *Differentiate questions from answer categories* by using regular type for the questions and uppercase letters for the answers. Put any specific directions for responding inside parentheses and in lowercase. For example,

   Do you favor setting standards in basic skills as a requirement for high school graduation? (check your response)

   _____ 1) NO
   _____ 2) YES

3. *Use numbers to identify the various response categories.* The numbers assigned to the options represent a form of pre-coding that will facilitate processing the data. You can place a blank or box in front of the response options and ask the respondent to place an *X* or check mark in the space. For example:

   What is the highest level of education that you have completed? (check one)

   _____ 1) GRADE SCHOOL
   _____ 2) SOME HIGH SCHOOL
   _____ 3) COMPLETED HIGH SCHOOL
   _____ 4) SOME COLLEGE
   _____ 5) COMPLETED COLLEGE
   _____ 6) SOME GRADUATE WORK
   _____ 7) GRADUATE DEGREE

   A check or *X* in front of item 5 indicates that the individual has completed college, and a count can easily be made for category. Because of the varying lengths of the options, survey designers recommend placing the numbers for the response options at the beginning, to the left of the response categories rather than at the right, for ease of scoring.

4. *Be consistent in assigning numbers to the various answer categories.* Always use the same number for the same answer throughout the questionnaire. It is conventional to assign low numbers to the negative responses and higher numbers to the positive responses. For example, 1 is assigned to "no" and 2 is assigned to "yes"; 1 is assigned to "unfavorable" and 2 to "favorable." Whatever number scheme is chosen, use it throughout because it is confusing for a respondent to associate 1 with "no" in the first part of the questionnaire and then find 1 associated with "yes" in another part.

5. *Response categories should be arranged vertically rather than horizontally.* The vertical arrangement makes the questionnaire appear less crowded and eliminates the common error of checking the space on the wrong side of the answer, as might occur for the following:

   What is your present marital status?

   _____ 1) NEVER MARRIED _____ 2) MARRIED _____ 3) DIVORCED
   _____ 4) SEPARATED _____ 5) WIDOWED _____ 6) UNMARRIED, COHABITING

The previous confusing arrangement could be improved in the following way:

What is your present marital status? (check one)

_____ 1) NEVER MARRIED
_____ 2) MARRIED
_____ 3) DIVORCED
_____ 4) SEPARATED
_____ 5) WIDOWED
_____ 6) UNMARRIED, COHABITING

6. *Use contingency questions when not every question will be relevant to all respondents.* A **contingency question** is one whose relevancy depends on the answer to a prior question. For example, in a survey designed to assess faculty interest in using computer-assisted testing, the following question might be used:

   1. Are you interested in using computer-assisted testing in your classes?
      1) NO (if NO, please go to question 6)
      2) YES
      3) UNDECIDED
   2. If YES, would you use the computer-assisted testing for
      1) QUIZZES
      2) TESTS
      3) BOTH

   In the preceding example, the second question is a contingency question because its relevance is contingent on the answer given in the first question. Faculty members who indicated in question 1 that they were not interested in computer-assisted testing would not even have to read the next four questions but could proceed to question 6, the next relevant question. Thus, contingency questions save time for the respondent and provide more accurate data for the researcher. Contingency questions can be set off by arrows, indenting, or enclosed boxes.

7. *Have the questionnaire reproduced by a high-quality printing method.* Quality printing gives the questionnaire a more professional appearance and makes a more favorable impression on respondents.
8. *Keep the questionnaire as short as possible.*

## RECORDING ANSWERS

The previous discussion has assumed that respondents will mark their answers directly on the questionnaire form. Whenever possible, we recommend that researchers provide scannable sheets and direct the respondents to mark all responses directly on the sheets. General-purpose scannable sheets with either 5 or 10 options per question are available at a nominal cost. The sheets can be read by an optical scanner, and the data can be put on a disk or sent directly

to a computer for analysis. Using scannable answer sheets saves time for the researcher and reduces human error in coding responses. The survey questions would be printed on a separate form, and the scannable sheet would be included for the responses. For example,

1. How many semesters of the *same* foreign language did you have in high school?

   A. NONE (0)
   B. ONE
   C. TWO
   D. THREE
   E. MORE THAN THREE

The response to the preceding question could be coded directly on the answer sheet. In some cases, it is possible to print survey questions directly on a scannable sheet. Figure 14.1 shows a sample of some survey questions used in a study of freshman students' perceptions of the adequacy of their high school preparation for college.

## FIELD TESTING

Before the final printing, the researcher must **field test** the instrument to identify ambiguities, misunderstandings, or other inadequacies. First, it is a good idea to ask colleagues who are familiar with the study to examine a draft of the questionnaire and give their opinions on whether the instrument will obtain the desired data and whether they see any problems that may have been overlooked.

Next, administer the questionnaire personally and one at a time to a small group drawn from the population to be considered in the study. Respondents answer the questions and provide feedback to the researcher on any difficulties they have with the items. Pay attention to such comments as "I don't know what you mean here" and "More than one of these answers apply to me." Try to ascertain whether all respondents interpret the questions in the same way. You might even state some of the questions in different ways to see if different responses are given to different versions of the same question. Observations made of the respondents as they fill out the questionnaire can also be enlightening. Spending an undue amount of time on a question or leaving a question blank and returning to it later can be clues that there are problems with some items.

The results of field tests can be used to clarify the items or perhaps to eliminate some. It is especially important to determine whether the questions will operate equally well for the different social classes and ethnic groups of the population to be studied. The following are some specific issues that should be addressed by field testing:

1. Do the respondents seem comfortable with the questionnaire and motivated to complete it?
2. Are certain items confusing?
3. Could some items result in hostility or embarrassment to respondents?
4. Are the instructions clear?

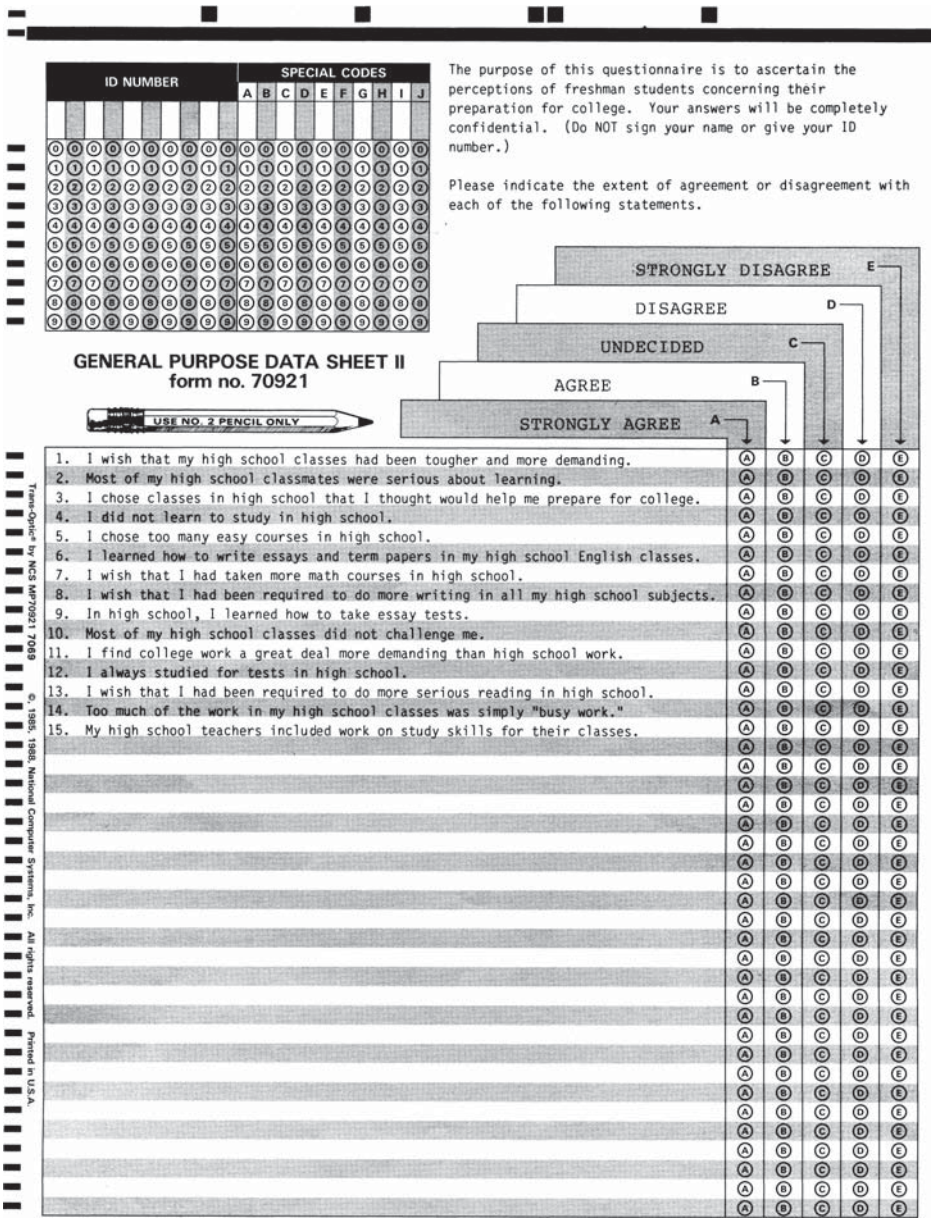



**Figure 14.1** Example of a Scannable Sheet with Survey Statements

*Source*: Reproduced with permission from National Computer Systems Inc.

5. How long will it take a respondent to complete the questionnaire?
6. Do all respondents interpret the items in the same way?

## PREPARING THE COVER LETTER

Researchers often find it useful to mail an introductory letter to potential respondents in advance of the questionnaire. This procedure alerts the subject to the study so that he or she is not overwhelmed by the questionnaire package. In any

case, a **cover letter** addressed to the respondent by name and title must accompany the questionnaire. One can use the "mail merge" feature of word processing packages to personalize the letter.

Figure 14.2 shows a cover letter with the important parts identified. The cover letter introduces the potential respondents to the questionnaire and "sells" them on responding. It should include the following elements:

1. *The purpose of the study.* The first paragraph of the letter should explain the purpose of the study and its potential usefulness. It will be helpful to relate the importance of the study to a reference group with which the individuals may identify. For example, a cover letter with a questionnaire for graduate students should stress the importance of the data for improving graduate education at the university.

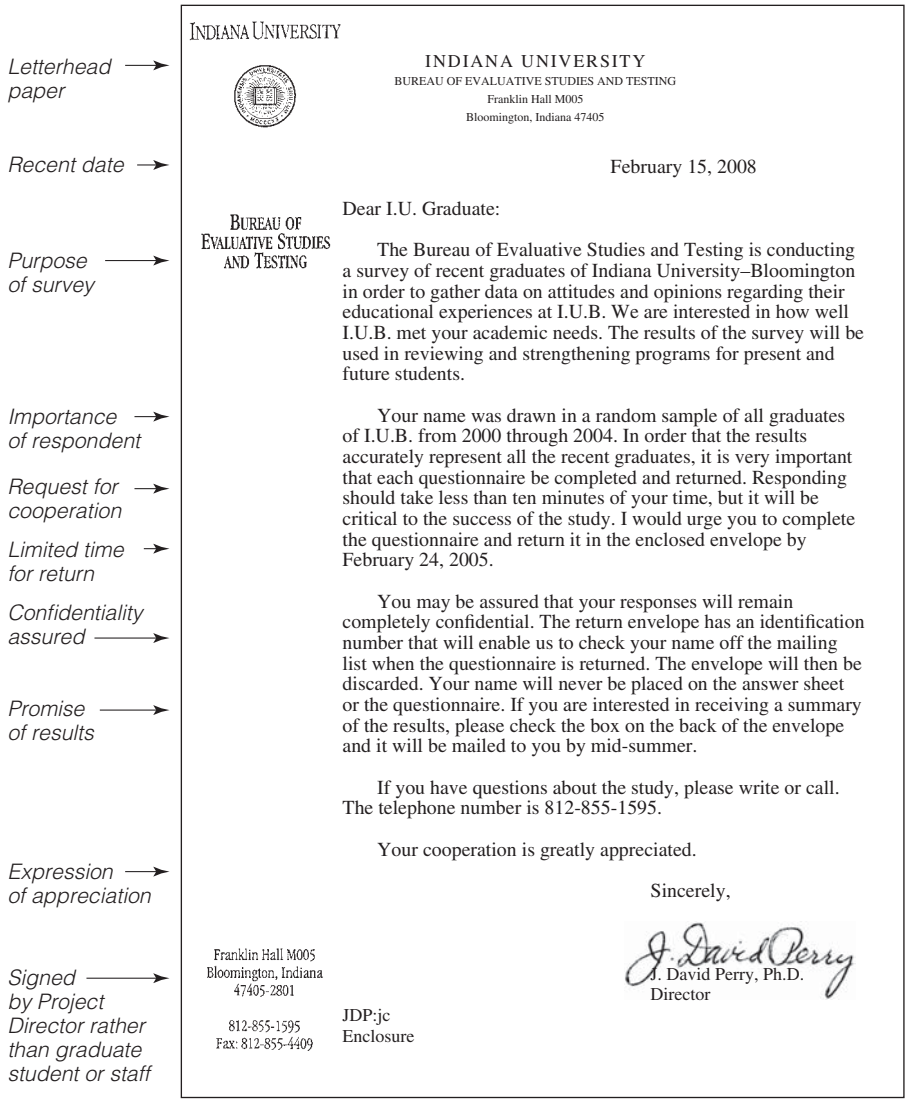



**Figure 14.2** Example of a Cover Letter for a Survey

2. *A request for cooperation.* The letter should explain why the potential respondent was included in the sample and should make an appeal for the respondent's cooperation. Respondents should be made to feel they could make an important contribution to the study.
3. *The protection provided the respondent.* The letter must not only assure the respondents that their responses will be confidential but also explain how that confidentiality will be maintained. To facilitate the follow-up procedure necessary for a high return rate, use identification numbers on the questionnaires. If there is no identification, the problem of nonresponse bias is compounded because there is no way to know who has responded and who has not, and follow-up procedures become very confused. If identification numbers are used, respondents must be told that the numbers are there simply to let the researcher check the respondents' names off the mailing list when the questionnaires are returned. The respondents must be assured that their names will never be placed on the questionnaires themselves; thus, there will be no way to associate particular responses with any individuals. If the researcher intends to destroy the questionnaires immediately after the responses have been rostered, this information should be conveyed in the letter to reassure the respondents of their anonymity.

   Some researchers prefer not to use any identification system at all, especially when the topic is sensitive. In this case, it is recommended that one include in the mail-out package a postcard that the respondent can mail separately to indicate that the questionnaire has also been mailed. This postcard contains a pretyped message that the questionnaire has been returned and a place for the respondent to write his or her name or a coded identifier. In this way, a record can be kept of the returned questionnaires.
4. *Sponsorship of the study.* The signature on the letter is important in influencing the return of the questionnaire. If the study is part of a doctoral dissertation, it is helpful if a person well-known to the respondents, such as the head of a university department or the dean of the college, signs or countersigns the letter. Such a signature is likely to be more effective than that of an unknown graduate student. If there is a sponsor for the study, such as a foundation or some agency, mention this. Use a university or agency letterhead.
5. *Promise of results.* An offer may be made to share the findings of the study with the respondents if they are interested. Tell them how to make their request for the results. One method is to provide a place for checking on the back of the return envelope, as well as a place for the respondent's name and address.
6. *Appreciation.* Include an expression of appreciation for their assistance and cooperation with the study.
7. *Recent date on the letter.* Date the cover letter near the day of mailing. A potential respondent will not be impressed by a letter dated several weeks before receipt.
8. *Request for immediate return.* It is also important to urge immediate return of the questionnaire. If a time period such as 2 weeks or 1 month is suggested, the respondent may lay the questionnaire aside and, despite good intentions, forget about it. A questionnaire that fails to receive attention within a week is not likely ever to be returned.

All of the preceding elements should be included, but at the same time the cover letter should be as brief as possible. One page is the maximum recommended length. Enclose the letter in an envelope along with the questionnaire. Always include a self-addressed, stamped return envelope for the respondent's use. This is indispensable for a good return rate.

Research shows that the type of postage used on the mailing can also influence returns. Stamps increase response rate over bulk-printed postage. Evidently, the stamp makes the questionnaire appear more personal and important and less like junk mail.

## MAXIMIZING RESPONSE RATES

Research shows that response rates for most U.S. national surveys of all types have been declining during the past four decades (Cook et al., 2000). Helgeson, Voss, and Terpening (2002) conceptualize the process of responding to a request to complete a survey as involving four phases through which the researcher must move the potential respondent (Figure 14.3).

1. *Attention,* the entry point to the response process, is critical for maximizing the probability of a response. If we can gain the respondent's attention, then we have moved the respondents into the process and increased the likelihood that there will be a follow-through to the end.
2. *Intention,* the next phase, involves the potential respondent's estimating the time and effort that must be expended to comply with the researcher's request for information and making a decision to continue with the survey.
3. *Completion* is the phase in which the potential respondent moves from the consideration of the survey completion process to the physical and mental activity necessary for actual survey completion.
4. *Return* involves putting the completed survey into a return envelope and mailing it, a relatively low-cost action compared to earlier phases of the decision process.

We will summarize some of the aspects of survey design that may affect the potential respondent at each phase.

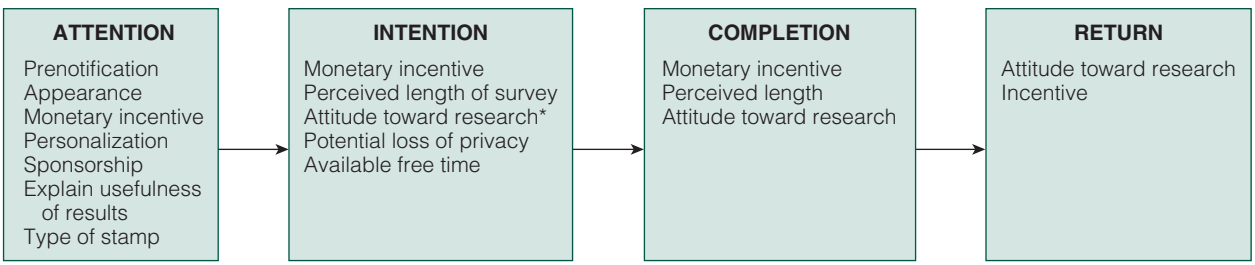



**Figure 14.3** Maximizing Response Rates

*Helgeson et al. found that an individual's attitude toward research is an important factor in responding to a survey.

*Source*: James G. Helgeson, Kevin E. Voss, and Willbann D. Terpening, "Determinants of Mail-Survey Response: Survey Design Factors and Respondent Factors," *Psychology and Marketing, 19*(3), March 2002, pp. 303–328. Reprinted by permission of John Wiley & Sons, Inc.

## MONETARY INCENTIVE

A great deal of research has shown that a token monetary incentive consistently increases response rate. The gratuity offered can be a nickel, a dime, a quarter, or even a dollar or more. Researchers believe that the payment creates a feeling of obligation on the part of the recipient and a need to reciprocate that is satisfied by completing and returning the questionnaire. Monetary incentives are most effective when prepaid rather than promised and when included with the first mailing of the questionnaire rather than with a follow-up. Among several survey design factors examined, Helgeson et al. (2002) found that the one-dollar incentive included with the mailed survey had the most effect on returns. Jobber, Saunders, and Mitchell's (2004) research not only supports the use of a prepaid monetary incentive but also shows that the response rate increases as the value of the incentive increases. They concluded from their studies that the inclusion of any incentive, regardless of amount, raises the response rate by an average of 15 percent. The size of the incentive had an additional effect at the rate of 2 percent per dollar within the observed range of incentives.

Newby, Watson, and Woodliff (2003) also found a monetary incentive to be an effective way to increase rates. Of course, offering a payment is not always possible because even a token amount can greatly increase the cost of the survey. If the sample is not too large, however, it is an option worth considering.

## FOLLOW-UPS

To reach the maximum percentage of returns in a mailed questionnaire survey, planned **follow-up** mailings are essential. Several steps are typically taken in the follow-ups and are explained here.

### First Reminder

If the questionnaire has not been returned in a week or 10 days after the initial mailing, send a postcard to the respondent. This card serves as a polite reminder that a questionnaire was sent earlier and that the response is very important to the study. Urge respondents to complete and return the questionnaire immediately ("today"). Of course, express thanks to those who may have already mailed the questionnaire. An offer can be made to send another questionnaire to people who may have misplaced or never received it. Usually, the postcard reminder brings in a relatively large number of responses.

### Second Follow-Up

This follow-up, which should be sent approximately 3 weeks after the original mailing, involves a letter, another copy of the questionnaire, and an addressed return envelope. The letter should first tell the nonrespondents that their questionnaires have not been received and should reiterate the usefulness of the study. Emphasize that a replacement questionnaire is enclosed and make a strong appeal to complete and return it. Tell the respondents not to respond a second time if they have already mailed the questionnaire.

### Third Follow-Up

The third and final follow-up is sent out 6 or 7 weeks after the initial mailing. It is similar to the second, having both a letter and a replacement questionnaire.

Many researchers send this follow-up by certified mail. If a researcher has 75 to 90 percent returns after three follow-ups, he or she may be ready to terminate the survey and to declare the remaining subjects nonrespondents. The researcher must decide whether the responses obtained through further follow-up efforts would be worth the cost and time involved. Also, you do not want potential respondents to feel harassed. It is sometimes suggested that the researcher include in the third follow-up a postcard on which subjects could indicate that they do not wish to participate in the survey and will not be returning the questionnaire. Such a procedure permits definite identification of nonrespondents.

## DEALING WITH NONRESPONSE

**Nonresponse** can be a serious problem in survey research. Nonresponse can bias survey data especially when it is nonrandom and if it is in some way correlated with the variables measured in the study. Nonrespondents may differ systematically from respondents. Research shows that respondents tend to differ from nonrespondents in characteristics such as education, intelligence, motivation, and interest in the topic of the survey. A survey with low response rate can thus be biased even though the researcher started out by mailing questionnaires to a representative sample. Recent studies, however, suggest that the effect of nonresponse may not be as pronounced s was once thought and that low response rates may not necessarily indicate bias (McCarty, 2003). Teitler, Reichman, and Sprachman (2003) investigated the costs and benefits of improving response rates for a difficult to reach population. They concluded that efforts to improve response rate were beneficial in obtaining a representative sample, but there was a point of diminishing returns beyond which the benefits were marginal. So where does this leave us? The findings from these studies do not justify low response rates; there is no question that high response rates are preferable to lower ones. They simply indicate that lower response rates do not necessarily translate into biased data.

What does a researcher do about the nonrespondents? We must not ignore them. If, after all follow-up attempts, response rate remain below approximately 75 percent, try to learn something about the characteristics of the nonrespondents and the extent to which they might differ from the respondents. There are several ways to do this:

1. *Compare respondents to population.* If you have access to information on characteristics of the population—such as age, gender, education, ethnicity, socioeconomic status, and so on—you can compare these characteristics of the respondents with those of the population from which the sample was selected. If your data show that your respondents closely resemble the population of interest, it is reasonable to assume that the nonrespondents also resemble the parent population. Thus, further study of nonrespondents is not necessary.
2. *Compare early to late respondents.* Research has shown that nonrespondents are often similar to late respondents. Thus, a second way to estimate the possible responses of nonrespondents is to look at those of the late respondents. Prior to this step, however, categorize the respondents into

early and late groups and compare their responses to check for any important differences. If no important differences appear between early and late respondents, you can assume the respondents are an unbiased sample of the recipients and you can thus generalize to the total group.

3. *Interview a sample of nonrespondents.* A more systematic approach is to interview either personally or by telephone a small random sample (perhaps 10 percent) of the nonrespondents. This sample of nonrespondents drawn for comparison purposes is sometimes called a "double-dipped sample." Using the questionnaire as an interview schedule, the investigator then gathers responses from the random sample of nonrespondents. The mean responses or the proportion of responses of the nonrespondents can be compared to those of the respondents to determine whether the two groups differ significantly. If no important differences are found when the responses of the initial respondents are compared with those of the interview sample, then the researcher can reasonably assume that the respondents represent an unbiased sample of all who received the questionnaire. The data can be pooled and generalizations made to the total sample and to the population. Without such a check, you have no way of knowing if the respondents are different and therefore biased. Using a double-dipped sample is the preferred method of checking for bias because the direction and the extent of bias that is due to nonresponse can be directly assessed. However, it is more costly and time-consuming.

If you find that certain clearly identifiable subgroups did not return the questionnaire, you may need to change the original research question to exclude these subgroups. For example, if secondary teachers showed a much lower rate of return of a survey than did elementary teachers, a researcher might conclude that the questionnaire had little relevance or interest for secondary teachers and decide to restrict the study to elementary teachers. The research question would be restated to indicate the change.

## VALIDITY

Attention should be given to the validity of interviews and questionnaires—that is, whether they are really measuring what they are supposed to measure. *Face validity,* which was discussed in Chapter 9, can be important in survey research. Subjects should perceive questions to be relevant. Consider the following questions:

1. The recipe for angel food cake calls for half a cup of sugar. You are only making half a cake. How much sugar should you use?
2. A pilot began a flight with half a tank of fuel. During this flight, he used half of his fuel supply. How much is left?

Both ask basically the same question. Candidates for pilot training would consider question 2 appropriate and question 1 inappropriate. The reverse would be true for culinary arts students. Question 2 has face validity for pilot training candidates. Question 1 has face validity for culinary arts students. Subjects are less inclined to complete and return a questionnaire they perceive as being inappropriate.

*Construct validity* can be assessed by having some colleagues who are familiar with the purpose of the survey examine the items to judge whether they are appropriate for measuring what they are supposed to measure and whether they are a representative sample of the behavior domain under investigation.

*Criterion-related validity* can be based on the relationship of survey responses to other variables. Direct observation of behavior, for example, has been a criterion used to validate surveys. After responses were obtained, observations were made to determine whether the actual behavior of the subjects agreed with their expressed attitudes, opinions, or other answers. If you find agreement between survey responses and actual behavior, you have some evidence for the criterion-related validity of the survey. Other data sources, such as third parties, may also be used as criteria.

Five potential problems may influence the validity of a questionnaire:

1. Respondents often report what they think is true but is not, or what they wish was true whether or not that is the case. Direct observation of a random sample of respondents is a good way to learn whether what they report is or is not correct.
2. Respondents may give untrue answers that are more socially acceptable than what is actually the case. Take this into account when interpreting survey results. If you read that a survey found that 40 percent of adult Americans would give a kidney transplant to a total stranger, ask yourself if 40 percent of adult Americans would yield a prime parking spot to a total stranger.
3. Respondents may give a response that they think the researcher wants to hear.
4. In interviews, respondents with little or no interest in the topic of the research may give off-hand, thoughtless responses just to get the investigator "off their back."
5. Respondents who are not sure their anonymity is assured may give "safe" answers.

For problems 1 through 4, direct observation of the behavior of a random sample of respondents is an excellent way to assess validity. Problem 5 can be avoided by making sure the respondents know that their anonymity is safeguarded.

## RELIABILITY

Recall from Chapter 9 that in order to be useful, a measure must first have **reliability** (consistency). Reliability indexes range from 0 (absolutely no consistency) to 1.0 (perfectly consistent). The strength of the reliability index sets the upper limit of the validity of a measure.

A procedure for assessing the reliability of an interview procedure is to have two or more interviewers ask the same subjects identical questions and then assess the consistency of the responses that the interviewers report. With questionnaires, internal consistency may be checked by building some redundancy into the instrument—items on the same topic may be rephrased and repeated in the questionnaire or interview. The more consistent the responses, the higher the reliability.

Such procedures are often expensive and time-consuming, however, and somewhat impractical because it is not easy to find subjects willing to repeat the questionnaire or interview. Another problem with this approach is that some answers to questions dealing with less stable aspects of behavior may legitimately change over time. However, it is important to assess consistency because if you employ a procedure with poor consistency, your data are worthless.

## STATISTICAL ANALYSIS IN SURVEYS

Surveys do not require complex statistical analyses. Data analysis may simply consist of determining the frequencies and percentages of responses for the questions of the study. For example, a survey of library resources may report the number of fiction books, the number of nonfiction books, and so on. A survey of people's attitudes on an issue may report the number and percentage of the respondents who gave each response, such as "strongly agree," "agree," "disagree," and so on. It is useful to convert numbers to percentages in order to be able to talk about the proportion responding a certain way and to be able to make comparisons. Consider the hypothetical frequency data in Table 14.3, based on a survey of 440 students concerning their opinions of a proposed increase in student activity fees at the university they attend. Looking at the raw frequencies, you might say that upperclassmen (66 juniors and seniors) and graduate students (66) are equally opposed to the increase in fees. However, if you calculate *percentages* based on the total number of students *in each group,* you can see that more than twice the proportion of graduate students disapprove of the increase compared to juniors and seniors. This is because 66 of the 88 graduate students (75 percent) disapprove compared to 66 of 192 (34 percent) juniors and seniors.

Thus, the table shows that graduate students are more likely to disapprove of the activity fee increase than undergraduates. To avoid a misinterpretation, always make sure that the total numbers for different groups are presented in tables such as this one.

### CONTROLLING VARIABLES IN A SURVEY ANALYSIS

Consider the hypothetical results of a survey of attitudes toward a new library tax for improving and expanding the county public library. Table 14.4 shows that 63 percent (150/240) of city residents favor the library tax compared with 37 percent (96/260) of the county residents. It appears from these data that there is a relationship between place of residence and attitude toward the library tax. A chi-square test will show whether there is a statistically significant relationship between the two variables (see Chapter 7 for discussion of chi square). Table 14.5

**Table 14.3** Students' Attitudes toward Increase in Activity Fee

| | Approve | No Opinion | Disapprove | Total |
|---|---|---|---|---|
| Freshmen and sophomores | 60 | 68 | 32 | 160 |
| Juniors and seniors | 80 | 46 | 66 | 192 |
| Graduate students | 12 | 10 | 66 | 88 |
| Total | 152 | 124 | 164 | 440 |

**Table 14.4** Attitudes toward Library Tax by Residence

| | City | County | Total |
|---|---|---|---|
| Favor | 150 | 96 | 246 |
| Oppose | 90 | 164 | 254 |
| Total | 240 | 260 | 500 |

**Table 14.5** Observed and Expected Frequencies for Attitude Data

| | City | County | Total |
|---|---|---|---|
| Favor | 150 (118) | 96 (128) | 246 |
| Oppose | 90 (122) | 164 (132) | 254 |
| Total | 240 | 260 | 500 |

$$\chi^2 = \frac{(150-118)^2}{118} + \frac{(96-128)^2}{128} + \frac{(90-122)^2}{122} + \frac{(164-132)^2}{132}$$

$$\chi^2 = 32.83$$

shows the calculation of chi square for these data. The expected frequencies for each cell are shown in parentheses.

Referring to Table A.4 of the Appendix with 1 degree of freedom, you can see that the chi square of 32.83 is highly significant (.01). You might conclude that there is a statistically significant relationship between place of residence and attitude toward the library tax. A more cautious observer, however, might point out that the city (the location of a major state university) has a greater proportion of educated people and that it may be educational level rather than place of residence per se that accounts for the favorable attitude toward the library tax.

To explore this alternative explanation, you can control for the variable of educational level by holding it constant and then observe whether the relationship between the first two variables continues to exist. The simplest way to hold the variable constant is to divide the subjects into separate groups, each having a different value on that variable, and to look at the crosstabs for each of these groups separately. In this case, you could categorize the 500 respondents into college educated and non-college-educated (assuming this information is available) and look at the relationship within the two separate groups.

The data in Table 14.6 show that the alternative explanation is correct: If college education is controlled by holding it constant, there is no relationship between the variables "place of residence" and "attitude toward the tax." Among the college educated, 80 percent (144/180) of the city residents favor the tax, and so do 80 percent (80/100) of the county residents. Among the non-college educated, 10 percent (6/60) of the city residents favor the tax, as do 10 percent (16/160) of the rural residents.

Suppose the data had been as shown in Table 14.7. Although educational level has been held constant within each table, the relationship between the variables "residence" and "attitude" is evident. Among the college educated, 67 percent (80/120) of the city residents favor the tax, compared with 33 percent (20/60) of the rural residents. Among the non-college educated, 58 percent of the city

**Table 14.6** Attitudes toward Library Tax Related to Education, Not to Residence

| | College-Educated | | |
|---|---|---|---|
| | **City** | **County** | **Total** |
| Favor | 144 | 80 | 224 |
| Oppose | 36 | 20 | 56 |
| Total | 180 | 100 | 280 |
| | **Non-College-Educated** | | |
| | **City** | **County** | **Total** |
| Favor | 6 | 16 | 22 |
| Oppose | 54 | 144 | 198 |
| Total | 60 | 160 | 220 |

**Table 14.7** Attitudes toward Library Tax Related to Residence, Not to Education

| | College-Educated | | |
|---|---|---|---|
| | **City** | **County** | **Total** |
| Favor | 80 | 20 | 100 |
| Oppose | 40 | 40 | 80 |
| Total | 120 | 60 | 180 |
| | **Non-College-Educated** | | |
| | **City** | **County** | **Total** |
| Favor | 70 | 76 | 146 |
| Oppose | 50 | 124 | 174 |
| Total | 120 | 200 | 320 |

residents favor the tax, compared with 38 percent of the rural residents. In this case, there *is* something about the city and county respondents, other than their education, that leads them to respond differently about the library tax.

You might also want to investigate sex differences in responses to a survey. In this case, the crosstabs would show the frequency of responses to the questions for males and females separately. Social class differences could be examined by categorizing the subjects into separate groups on the basis of appropriate criteria and looking at the responses for each of the groups. Tables that show frequencies of different groups' responses are often the best way to illustrate the relationship among the variables of a survey. These tables are called cross tabulations, or **crosstabs,** because they permit comparisons across groups. The simplest crosstabs contain two variables with two categories for each variable. More complex forms are possible, however, such as $2 \times 3$, $2 \times 4$, $3 \times 3$, and so on. Cross tabulations are used most often with categorical or nominal data.

## STATISTICS FOR CROSSTABS

Crosstabs are a widely used way to graphically show the differences in responses among various groups and whether or not a relationship may exist between the variables. When both variables in the cross tabulation are measured at the

nominal level, the chi-square test may be used to determine whether a systematic relationship exists between the two variables. Chi square, however, indicates *only* whether the variables are related or are independent. It does not tell the extent to which they are related. For example, the value of $\chi^2$ for the "college-educated" versus the "non-college-educated" data in Table 14.6 is 17.09, which is statistically significant at the .01 level. Thus, you know that there is a significant relationship between place of residence and attitude toward the tax among the college educated, but you do not know how strongly the two variables are related.

To learn the extent of the relationship, you must calculate a coefficient of correlation. A coefficient that is frequently used with nominal data in $2 \times 2$ tables is the *phi coefficient* ($\varphi$). The phi coefficient is a mathematical simplification of the Pearson product moment coefficient for $2 \times 2$ tables. Thus, phi has a value of 0 when no relationship exists, +1.00 when a perfect positive relationship exists, and 1.00 when a perfect negative relationship exists. The phi coefficient for the college educated in Table 14.6 is .32. The phi coefficient is interpreted like any other Pearson coefficient. The .32 indicates a moderate positive relationship between place of residence and attitude toward the library tax among the college educated.

For tables larger than $2 \times 2$, an appropriate measure of relationship is the kappa statistic ($\kappa$) (Siegel & Castellan, 1988, p. 284). If there is a perfect relationship between the variables, $\kappa$ will equal 1.00. If agreement between variables is exactly what would be expected through chance, $\kappa$ equals 0. If agreement is less than what is expected by chance, $\kappa$ will be a negative number. If the two variables in the cross tabulation are ordinal, statistics such as the Kendall coefficient of concordance ($W$) or the gamma statistic can be used to indicate the strength of the relationship between variables. [You can find comprehensive discussion of correlation procedures for nominal and ordinal variables in Chapter 9 of Siegel and Castellan's (1988) text.]

## SUMMARY

The survey is a widely used research method for gathering data ranging from physical counts and frequencies to attitudes and opinions. Surveys are classified according to their focus, scope, and time of data collection. They should involve careful planning, unbiased sampling of a population, thoughtful development of data-gathering instruments, and careful analysis of the results.

If researchers choose some type of probability sampling, they can infer population values from the sample results. The usual procedure is to set up an interval or range that has a high probability of including the population value. The width of this interval is a function of the risk they are willing to take of being wrong and the sample size. The interval narrows as the acceptable probability of error increases and as sample size increases. You can also use this procedure to estimate the sample size needed for any desired margin of error.

Interviews and questionnaires are the major means of data collection for a survey. Both procedures involve asking questions of selected subjects, but each has unique advantages and disadvantages. It is important that the instruments used be as valid and reliable as possible. Various follow-up procedures have proved effective in increasing returns from mailed questionnaires.

Cross tabulations provide an excellent way to show the relationship existing among the variables in a survey.

## KEY CONCEPTS

census
census of intangibles
census of tangibles
checklists
closed-ended question
cohort study
completion (fill-in) items
contingency question
cover letter
cross-sectional survey
crosstabs
directly administered questionnaire
double-barreled question
electronic mail survey
field test
focus group
follow-up
intangibles
interviewer bias
Likert-type item
longitudinal survey
mailed questionnaire
margin of error
matrix sampling
nonresponse
open-ended question
panel study
pause
personal interview
population
probe
questionnaire
ranking items
reliability
response rate
sample
sample survey
sample survey of intangibles
sample survey of tangibles
sampling error
sampling frame
scaled items
social desirability bias
survey research
tangibles
trend study
web-based survey

## EXERCISES

1. Suggest a research question that can best be answered by means of a survey. Write two open-ended items, two closed-end items, and two Likert scale items that would provide data relevant to the question.
2. What data-gathering technique would you use for each of the following surveys?
   a. A survey of a sample of high school teachers throughout the state concerning the use of mandatory competency exams for high school graduation
   b. A survey of the opinions of people in a major metropolitan area on the way juveniles who commit violent crimes are currently handled in the state's court system
   c. A survey of the opinions of a sample of professors on the campus of State University about the use of pass/fail grades instead of letter grades (You want a very high response rate.)
   d. A freshman survey of certain noncognitive characteristics that might relate to academic achievement in the first year of college
   e. A survey of 500 people from throughout the United States about their attitudes toward the administration's immigration policy
3. How does the proportion of yes/no responses influence the sample size needed?
4. The evening news on television reported the following: In a recently conducted survey of the American public, 37 percent of the respondents said they approve of the president's performance. What else would you want to know before you interpreted this report?
5. How would you interpret the following report? "A poll of 1000 randomly selected registered voters in Indiana found that 45 percent favored using state lottery profits for education. Figures from this poll are subject to a sampling error of 3 percent. The confidence level is 95 percent."
6. For each of the following three sample sizes, calculate the 95 percent margin of error for the population proportion. Assume the sample proportion is .40 for each. What effect does an increase in sample size have on the width of the interval? Why?
   $n$ of sample A = 100
   $n$ of sample B = 1000
   $n$ of sample C = 10,000

7. A school superintendent wants to determine what proportion of the 5000 middle and high school students in the district use cigarettes. He will distribute a questionnaire to a random sample of the students. How many students will he need to sample in order to have a 95 percent margin of error of ±5 percent in his estimate?
8. A national polling organization wants to be able to predict the outcome of the presidential election to within 3 percent. How large a random sample will be needed to achieve this level of precision? Assume a 95 percent confidence level.
9. A survey had an initial response rate of 51 percent. What would you suggest to the researcher for dealing with this low response rate?
10. The administration at State University wants to estimate the incoming freshman class's interest in a computer science major. Because it does not have the financial resources to survey all 5000 freshmen, the administration surveys a random sample of 500. It finds that 100 students report they are interested in such a major.
    **a.** Calculate the margin of error in this survey.
    **b.** What is the best estimate of the number of freshmen who would be interested in majoring in computer science?
11. Which of the following would be biased samples of the population of college students at a large university?
    **a.** A random sample of students entering the library on Friday evening
    **b.** A random sample of students registered for classes
    **c.** A random sample of students buying season tickets for basketball
    **d.** A sample composed of students who volunteered for a project after seeing a notice in the school newspaper
12. Assume you are conducting a survey to determine how elementary school teachers in a district feel about the policy of retaining students.
    **a.** Write two closed-ended questions for this survey.
    **b.** Write two open-ended questions for this survey.
    **c.** Write a contingency question.
13. On the basis of the time of data collection, classify each of the following surveys:
    **a.** Terman's study of adults who were intellectually gifted as children
    **b.** A comparison of math achievement in public middle schools in the United States in 1990, 1995, and 2000
    **c.** A follow-up of the 1995 graduates of the Indiana University School of Business
    **d.** A survey of reading achievement at different grade levels in a school system in 2000
14. Which of the following is an advantage of the longitudinal type of survey?
    **a.** More intensive individual study
    **b.** Providing data for different age groups at the same time
    **c.** Prompt data gathering
    **d.** No sampling errors
15. A popular weekly news magazine included a detailed questionnaire on attitudes about crime. The editors invited readers to fill out the questionnaire and return it to the magazine. The magazine subsequently printed an article on the widespread fear of crime among U.S. citizens, especially older people. How would you evaluate this survey?
16. A graduate student is planning to use a survey to gather data for her dissertation but is unsure whether to use mailed questionnaires or telephone interviews. Examine the following list of considerations and indicate whether a questionnaire or an interview would be more appropriate.
    **a.** She is on a tight financial budget for her data collection.
    **b.** Her goal is a 90 percent response rate from the sample.
    **c.** Some of the questions may involve sensitive issues.
    **d.** She hopes to minimize the time needed for coding and organizing data and getting it ready for computer analysis.

## ANSWERS

1. Answers will vary. *Example:* How do adults in a community feel about the building of a new elementary school?
2. **a.** Mailed questionnaire
   **b.** Telephone interview
   **c.** Personal interview
   **d.** Directly administered questionnaire
   **e.** Telephone interview
3. The nearer the proportions are to .50/.50, the larger the sample size needed.
4. One would want to know the size of the sample, how it was drawn, the width of the interval around the estimated population parameter (the margin of error), and the confidence level.
5. The .95 confidence interval for the proportion of registered voters who favor using lottery funds for education is between 42 and 48 percent.
6. **Sample A**

$$\sigma_p = \sqrt{\frac{(.40)(.60)}{100}} = \sqrt{\frac{.24}{100}} = \sqrt{.0024} = .04899$$

$.04899 \times 1.96 = .0960$ or .10 margin of error

**Sample B**

$$\sigma_p = \sqrt{\frac{(.40)(.60)}{1000}} = \sqrt{\frac{.24}{1000}} = \sqrt{.0024} = .0155$$

$.0155 \times 1.96 = .03$ margin of error

**Sample C**

$$\sigma_p = \sqrt{\frac{(.40)(.60)}{10{,}000}} = \sqrt{\frac{.24}{10{,}000}} = \sqrt{.000024} = .004899$$

$.0049 \times 1.96 = .01$ margin of error

An increase in sample size decreases the width of the interval. As n increases, the sample error decreases and hence the size of the interval.

7. Use Formula 14.2. Assume $p = q = .50$:

$$n = \left(\frac{\frac{1}{E}}{\sqrt{pq}}\right)^2 (1.96)^2 = \left(\frac{\frac{1}{.05}}{\sqrt{(.5)(.5)}}\right)^2 (1.96)^2$$

$$= \left(\frac{1}{.10}\right)^2 = (1.96)^2 = 100(3.8416) = 384$$

8. Assume $p = q = .50$:

$$n = \left(\frac{\frac{1}{E}}{\sqrt{pq}}\right)^2 (1.96)^2 = \left(\frac{\frac{1}{.03}}{\sqrt{(.5)(.5)}}\right)^2 (1.96)^2$$

$$= \left(\frac{1}{.06}\right)^2 (1.96)^2 = (277.78)(3.8416) = 1067$$

9. The researcher should follow up with a postcard reminder and then another mailing or two of the questionnaire. After follow-up efforts have been completed, the researcher should try to interview some of the remaining nonrespondents to find out about their characteristics and to obtain their responses in order to determine if they differ significantly from the respondents.
10. **a.** $p = .20$
    $q = .80$

$$\sigma_p = \sqrt{\frac{pq}{n}}$$

$$\sigma_p = \sqrt{\frac{(.20)(.80)}{500}}$$

$$= \sqrt{.00032} = .0179$$

margin of error $= 1.96 \times .0179 = .035$ or $\pm 3.5\%$

   **b.** Between 16.5 and 23.5 percent or between 825 and 1175 students would be interested in a computer science major.
11. Samples a, c, and d would not be representative of the population of college students at a large university.
12. Answers will vary.
13. **a.** Longitudinal panel study
    **b.** Longitudinal trend study
    **c.** Longitudinal cohort study
    **d.** Cross-sectional survey
14. a
15. Those who completed and returned the questionnaire would not be a representative sample of all U.S. citizens. A number of factors would bias this sample, such as socioeconomic level, educational level, sufficient interest in the topic to complete the questionnaire, and payment of postage to return it.
16. **a.** Mailed questionnaire
    **b.** Telephone interview
    **c.** Mailed questionnaire
    **d.** Telephone interview

## REFERENCES


Cook, C., Heath, F., & Thompson, R. (2000). A meta-analysis of response rates in web or Internet-based surveys. *Educational and Psychological Measurement, 60*(6), 821–836.

DeVaus, D. (2002). *Conducting surveys using the Internet.* Thousand Oaks, CA: Sage.

Dillman, D. A. (2000). *Mail and Internet surveys: The tailored design method* (2nd ed.). New York: Wiley.

Dillman, D. A., & Bowker, D. K. (2001). The web questionnaire challenge to survey methodologists. In U. D. Reips & M. Bosnjak (Eds.), *Dimensions of Internet science.* Lengerich, Germany: Pabst Science.

Flatley, J. (2001, May 11–12). *The Internet as a mode of data collection in government social surveys: Issues and investigation.* Paper presented at the International Conference on Survey Research Methods, Latimer, UK.

Fowler, F. J. (2002). *Survey research methods* (3rd ed.). Thousand Oaks, CA: Sage.

Helgeson, J., Voss, K., & Terpening, W. (2002). Determinants of mail survey response: Survey design factors and respondent factors. *Psychology and Marketing, 19,* 303–328.

Holbrook, A., Green, M., & Krosnick, J. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires. *Public Opinion Quarterly, 67,* 79–125.

Jobber, D., Saunders, J., & Mitchell, V. (2004). Prepaid monetary incentive effects on mail survey research. *Journal of Business Research, 57*(4), 347–350.

McCarty, C. (2003). Differences in response rates using most recent versus final dispositions in telephone surveys. *Public Opinion Quarterly, 67,* 396–406.

Meister, D., & Melnick, S. (2003). National new teacher study: Beginning teachers' concerns. *Action in Teacher Education, 24*(4), 87–94.

Neuman, W. L., & Kreuger, L. (2003). *Social work research methods: Qualitative and quantitative approaches.* Boston: Allyn & Bacon.

Newby, R., Watson, J., & Woodliff, D. (2003). SME survey methodology: Response rates, data quality, and cost effectiveness. *Entrepreneurship Theory and Practice, 28,* 163–172.

Rose, L., & Gallup, A. (2007). The 39th annual Phi Delta Kappa/Gallup poll of the public's attitudes toward the public schools. *Phi Delta Kappan, 89*(1), 33.

Siegel, S., & Castellan, N. J., Jr. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York: McGraw-Hill.

Teitler, J., Reichman, N., & Sprachman, S. (2003). Costs and benefits of improving response rates for a hard-to-reach population. *Public Opinion Quarterly, 67,* 126–138.

Terman, L. M. (1926). The mental and physical traits of a thousand gifted children. In *Genetic studies of genius* (Vol. 1). Stanford, CA: Stanford University Press.

CHAPTER 15

# Defining and Designing Qualitative Research

The research methods we choose say something about our views on what qualifies as valuable knowledge and our perspective on the nature of reality.

Glesne and Peshkin (1992)

## INSTRUCTIONAL OBJECTIVES

After studying this chapter, the student will be able to:

1. Define qualitative research and identify the critical aspects of this methodology.
2. Distinguish between quantitative and qualitative research.
3. Describe ways qualitative inquirers gather data.
4. Distinguish between participant and nonparticipant observation.
5. Identify ethical issues associated with qualitative research.
6. Explain different approaches to selecting participants for a qualitative study.

The research methods described in Chapters 10 through 14 primarily use numeric data to answer questions. Such procedures are classified as quantitative research because they employ such quantitative measures as frequencies, means, correlations, and statistical tests. In contrast, qualitative research employs words and images to answer questions. The two approaches to research differ in the type of data collected and reported and in what the researcher believes warrants the claims that are made. One form of research is not superior to the other. They are designed to answer different questions, have different assumptions, and have different ways of inquiring into realities. Qualitative research investigates the quality of relationships, activities, situations, and materials. It focuses on understanding the context and attempts to explain the intentionality of behaviors.

## DISTINGUISHING QUALITATIVE INQUIRY FROM QUANTITATIVE INQUIRY

The phrase **qualitative inquiry** is a generic term for an array of educational research approaches, such as ethnography, naturalistic inquiry, narrative research, case studies, interpretive research, fieldwork, field studies, and participant observation. These approaches use different methodologies, but certain features are typical of qualitative research. These characteristics set qualitative research apart from the quantitative approach to educational research outlined in Chapter 5 and the methodology described in Chapters 10 through 14. Both qualitative and quantitative researchers approach their studies by stating a purpose, posing a problem or question, defining a research population, collecting and analyzing data, and presenting results. Both use theories and both are concerned with the rigor of their inquiry. They differ in their views about the nature of reality, their assumptions about the role of the researcher, and in how they define knowledge.

### APPROACH

The quantitative approach to the study of social and behavioral phenomena holds that the aim and methods of the social sciences are, at least in principle, the same as the aim and methods of the natural or physical sciences. **Quantitative research** strives for testable and confirmable theories that explain phenomena by showing how they are derived from theoretical assumptions (see the discussion of scientific theory in Chapter 1). It seeks scientific explanation that includes the discovery of laws governing not only the behavior of the physical world but also human behavior.

Qualitative inquiry begins from a different assumption, namely that the subject matter of the social or human sciences differs fundamentally from the subject matter of the physical or natural sciences and therefore requires a different goal for inquiry and a different set of methods for investigation. Qualitative inquirers argue that human behavior is always bound to the context in which it occurs, that social reality cannot be reduced to variables in the same manner as physical reality, and that what is most important in the social disciplines is understanding and portraying the meaning that is constructed by the participants involved in particular social settings or events.

Qualitative inquiry seeks to understand and interpret human and social behavior as it is lived by participants in a particular social setting. It is an intensely personal kind of research, one that freely acknowledges and admits the subjective perceptions and biases of both participants and researcher. Defenders of qualitative approaches argue that, in contrast, quantitative inquiry is principally concerned with the discovery of "social facts" devoid of subjective perceptions of intentions and divorced from particular social and historical contexts.

### EXPLANATION

Quantitative approaches in the human sciences rely on a hypothetico-deductive model of explanation. Inquiry begins with a theory of the phenomenon to be investigated. From that theory any number of hypotheses are deduced that, in turn, are tested using a predetermined procedure such as an experimental, ex post facto, or correlational design. The ultimate goal of researchers using this hypothetico-deductive model is to revise and support theories, or law-like statements, of social and behavioral phenomena based on the results of hypothesis testing. Theories are

refined and extended (and sometimes abandoned) to account for the results of testing their implications or instances (deductions). One goal of quantitative approaches is to generalize findings from a randomized sample to a larger population.

Qualitative inquiry relies on a different model of explanation and argues for a different goal of inquiry. In general, its practitioners hold that the search for generalizations is misguided. Human behavior is always bound to a particular historical, social, temporal, and cultural context. Qualitative inquirers seek to interpret human actions, institutions, events, customs, and the like, and in so doing they construct a "reading," or portrayal, of what is being studied. The ultimate goal of this kind of inquiry is to portray the complex pattern of what is being studied in sufficient depth and detail so that someone who has not experienced it can understand it. When qualitative inquirers interpret or explain the meaning of events, actions, and so forth, they generally use one of the following types of interpretation: (1) construction of patterns through analysis and resynthesis of constituent parts, (2) interpretation of the social meaning of events, or (3) analysis of relationships between events and external factors. These interpretations may lead to the generation of theories, be guided by existing theories or concept maps, or seek to further explicate or expound upon a theory.

## METHODS

Quantitative methods use empirical approaches, experimental designs, and often statistical testing compared to the more naturalistic, emergent, and field-based methods typical of qualitative research. The primary instrument used for data collection in qualitative research is the researcher him- or herself, often collecting data through direct observation or interviews. Quantitative research more typically relies on measurement tools such as scales, tests, observation checklists, and questionnaires. The selection of subjects for study also differs. The ideal selection in quantitative research is random sampling, which allows for control of variables that may influence findings. Qualitative studies more typically use nonrandom or purposive selection techniques based on particular criteria.

## VALUES

Quantitative inquirers admit that the inquirer's values may play a role in deciding what topic or problem to investigate but maintain that the actual investigation should aim to be as value free as possible; that is, the inquirer must follow procedures specifically designed to isolate and remove subjective elements to the extent possible. The goal is to control or remove personal value from the inquiry situation so that what remains are just the "objective facts." For example, imagine an experimental study involving two different classes of third-graders in which one third-grade class is the experimental group and the other is the control group. Imagine further that observers are placed in each classroom to record interactions between teachers and students. Quantitative inquirers prefer that the observers be unaware of whether they are observing the experimental or the control group, that they be unaware of subject characteristics (their social class, IQ, previous academic achievement, etc.), and that they use highly structured observational protocols that require only low-level inferences and as little as possible interpretation about what is happening in the interactions between teacher and students. These procedures are used in quantitative inquiry to ensure that the observers'

values and beliefs will minimally influence or contaminate the observations that they make. By following these procedures for making observations, the quantitative inquirer provides strong assurance that the inquiry is value free.

In contrast, the qualitative approach suggests that inquiry is always value bound; it can never be considered value free, and inquirers must be explicit about the roles that values play in any given study. Qualitative inquirers argue that inquiry is value bound in the choice of a problem to investigate, in the choice of whether to adopt a quantitative or qualitative approach to a problem, in the choice of methods used to investigate that problem, in the choice of a way to interpret results or findings, and by the values inherent in the context where the study takes place. Qualitative inquirers believe that it is impossible to develop a meaningful understanding of human experience without taking into account the interplay of both the inquirers' and the participants' values and beliefs. They believe that rather than try to eliminate bias, it is important to identify and monitor biases and how they may affect data collection and interpretation. Furthermore, qualitative inquirers argue that human inquiry requires frequent, continuing, and meaningful interaction between inquirers and their respondents (subjects) and that inquiry must maximize rather than minimize this kind of contact.

Corbin and Strauss (2008) make the argument that in contrast to the objectivity valued in quantitative research, *sensitivity* in qualitative research requires being tuned in to what is happening, being able to present the world as seen by the participants. The ability to be sensitive may come more naturally to some researchers, but it can be developed with training. The researcher must understand that findings and interpretations are both the result of the data and the experiences, beliefs, and values the researcher brings to the task and that his or her background and knowledge also contribute to the ability to see connections between concepts.

## THINK ABOUT IT 15.1

For each of the following statements, indicate whether it is more descriptive of qualitative (QL) or quantitative (QT) research. Which method:

1. Relies more on the inductive approach
2. Is more likely to use random sampling
3. Relies more on the deductive approach
4. Is more likely to use purposive sampling
5. Is more likely to include a statistical report
6. Is more likely to include extensive quotations
7. Is more interested in generalizing
8. Is more likely to include intense interactions between the researcher and the subject
9. Is more likely to allow the researcher to modify, delete, or add interview questions during the interview
10. Is more likely to investigate a topic about which little is known

**Answers**

1. QL; 2. QT; 3. QT; 4. QL; 5. QT; 6. QL; 7. QT; 8. QL; 9. QL; 10. QL

**Table 15.1** Comparison of Quantitative Inquiry and Qualitative Inquiry

| Quantitative Inquiry | Qualitative Inquiry |
|---|---|
| **Purpose** | |
| To generalize findings | To contextualize findings |
| To predict behavior | To interpret behavior and intention |
| To provide causal explanations | To understand perspectives |
| **Approach** | |
| Uses theory to ground the study | May create theory grounded in the findings |
| Uses manipulation and control of variables | Portrays the natural context |
| Deductive then inductive | Inductive then deductive |
| Seeks to analyze discrete components | Searches for larger patterns |
| Looks for the norm | Looks for complexity |
| Reduces data to numbers | Relies on words and only minor use of numbers |
| Reports written in precise, abstract language | Reports written in descriptive, holistic language |
| **Assumptions** | |
| There is an objective reality | Reality is socially constructed |
| The world is stable | The world is not stable |
| Variables can be identified and measured | Complex variables are difficult to measure |
| Is rooted in logical empiricism | Is rooted in symbolic interactionism |
| **Role of Researcher** | |
| Detached and impartial | Personally involved |
| Objective portrayal | Empathic understanding |
| Inquiry should be as value free as possible | Inquiry is always value bound |
| **Methods** | |
| Focused on quantity (how much, how many) | Focused on quality (nature, essence) |
| Experimental, empirical, statistical focus | Fieldwork, ethnographic, naturalistic focus |
| Predetermined, structured methods, precise | Flexible, evolving, emergent methods |
| Random sampling is the ideal | Typically uses purposive sampling |
| Uses inanimate instruments (scales, tests, questionnaires, observation checklists, etc.) | Researcher as the primary instrument (observations, interviews, document analysis) |

*Source*: Adapted from Merriam (1998). Also adapted from Glesne and Peshkin, *Becoming Qualitative Researchers: An Introdution*, p. 7. Published by Allyn & Bacon, Boston. Copyright © 1992 by Pearson Education. Adapted by permission of the publisher.

Table 15.1 summarizes key differences between quantitative and qualitative inquiry. The contrasting approach in the table is not always as clear in the actual conduct of research but is designed to give the reader a simple base from which to compare the two approaches.

## MAJOR CHARACTERISTICS OF QUALITATIVE RESEARCH

Although qualitative inquirers work in many different ways, their studies have certain characteristics in common that set this approach apart from quantitative research. Some of the more important aspects of qualitative research are discussed next.

## CONCERN FOR CONTEXT AND MEANING

Qualitative inquiry shows **concern for context and meaning.** It assumes that human behavior is context bound—that human experience takes its meaning from and, therefore, is inseparable from social, historical, political, and cultural influences. Thus, inquiry is always bounded by a particular context or setting. Qualitative researchers focus on how people make sense of or interpret their experience. Qualitative inquiry aims to understand intention. There is no attempt to predict what will happen in the future but, rather, to understand a unique and particular context. Proponents of qualitative inquiry argue that the quantitative approach to the study of human experience seeks to isolate human behavior from its context; it engages in **context stripping.**

## NATURALLY OCCURING SETTINGS

Qualitative research studies behavior as it occurs naturally in a classroom, an entire school, a playground, or in an organization or community. Qualitative inquiry takes place in the field, in settings as they are found. It is not a setting contrived specifically for research, and there is no attempt to manipulate behavior. The researcher acknowledges that where the research is conducted (the setting) influences the findings. The researcher goes physically to the people, the setting, or the institution to observe behavior. Virtual presence may be possible with newer technologies. In addition, qualitative inquiry places no prior constraints on what is to be studied. It does not identify, define, and investigate or test the relationship between independent and dependent variables in a particular setting; rather, it studies human experience holistically, taking into account a broad range of factors and influences in a given situation.

## HUMAN AS INSTRUMENT

One of the distinguishing characteristics of qualitative research is the methods used to collect and analyze data. In qualitative studies, the human investigator is the primary instrument for the gathering and analyzing of data. Lincoln and Guba (1985) introduced the concept of **human as instrument** to emphasize the unique role that qualitative researchers play in their inquiry. Because qualitative research studies human experiences and situations, researchers need an instrument flexible enough to capture the complexity of the human experience, an instrument capable of adapting and responding to the environment. It is believed that only a human instrument is capable of this task. He or she talks with people in the setting, observes their activities, reads their documents and written records, and records this information in field notes and journals. Qualitative inquiry relies on fieldwork methods (interviewing, observation, and document analysis) as the principal means of collecting data, avoiding the use of paper-and-pencil tests, checklists, mechanical instruments, and highly structured observational protocols.

## DESCRIPTIVE DATA

The qualitative inquirer deals with data that are in the form of words or pictures rather than numbers and statistics. Data in the form of quotes from documents, field notes, and interviews or excerpts from videotapes, audiotapes, or

electronic communications are used to present the findings of the study. The data collected are the participant experiences and perspectives; the qualitative researcher attempts to arrive at a rich description of the people, objects, events, places, conversations, and so on. Occasionally, some numeric data may be collected. Managing the large volume of descriptive data generated from interviews, observations, and the collection of documents is an important consideration in qualitative studies. Qualitative investigators also typically keep a personal or reflexive log or journal in which they record accounts of their thoughts, feelings, assumptions, motives, and rationale for decisions made. This is one way that the qualitative inquirer addresses the issue of the inquiry being value bound.

## EMERGENT DESIGN

In quantitative studies, researchers carefully design all aspects of a study *before* they actually collect any data; they specify variables, measures for those variables, statistics to be used to analyze data, and so forth. In contrast, while qualitative inquirers broadly specify aspects of a design before beginning a study, the design continues to *emerge* as the study unfolds, hence the term **emergent design.** They adjust their methods and way of proceeding (design) to the subject matter at hand. This is necessary because the qualitative inquirer is never quite sure just what will be learned in a particular setting because what can be learned in a particular setting depends on the nature and types of interactions between the inquirer and the people and setting, and those interactions are not fully predictable, and also because important features in need of investigation cannot always be known until they are actually witnessed by the investigator.

## INDUCTIVE ANALYSIS

In most qualitative studies, data collection and data analysis take place simultaneously. In other words, the inquirer does not wait until all the data are "in" before beginning to interpret them. From the outset of the first interview or observation, the qualitative inquirer is reflecting on the meaning of what he or she has heard and seen, developing hunches (working hypotheses) about what it means, and seeking to confirm or disconfirm those hunches in subsequent interviews or observations. It is a process of **inductive data analysis;** it proceeds from data to theory or interpretation. As the inquirer reduces and reconstructs the data through the processes of coding and categorization, he or she aims at interpreting the phenomena being observed.

### THINK ABOUT IT 15.2

For each research question listed below, indicate whether you would choose a qualitative (QL) or quantitative (QT) research approach. (Note: In some cases, the question could be answered using more than one approach. Select the one you believe would be most appropriate.)

1. How are the social relations of adolescents who use illicit drugs different from those who do not use them?
2. How do school attendance and grades earned in school differ between adolescents who use illicit drugs and those who do not?

3. To what extent does family income predict whether a student will choose to attend a commuter or residential campus?
4. How do Hispanic and Latino students experience their first year in an urban community college?
5. Do students who have high scores on reading tests also have high scores on writing tests?
6. How do middle school students of differing ability levels approach reading?
7. What are the characteristics of mathematics lessons in Japanese and U.S. middle school textbooks?
8. How do U.S. middle school students compare to Japanese middle school students in performance on standardized mathematics examinations?
9. Do mainstreamed students in science classes using cooperative grouping differ in their performance from those not exposed to cooperative groups?
10. What are the helping behaviors of students in cooperative learning groups?
11. What are the personal and educational interactions in a group of teachers developing a high school chemistry curriculum?
12. How do gangs recruit members in schools?

**Answers**

1. QL; 2. QT; 3. QT; 4. QL; 5. QT; 6. QL; 7. QL; 8. QT; 9. QT; 10. QL; 11. QL; 12. QL

Maxwell (2005) discusses researcher goals for which he believes qualitative studies are especially suited: understanding meaning for the participants, understanding a particular context, identifying unanticipated phenomena and influences through which new theories may be generated, understanding process, and developing causal explanations (though he recognizes the dispute inherent when considering the more traditionalist research views). He contends that the qualitative focus on understanding process rather than regularities is aligned with current science philosophies.

## DESIGNING QUALITATIVE RESEARCH

The research design is the researcher's plan of how to proceed to gain an understanding of some group or some phenomenon in its context. Individuals, families, and a variety of groups, organizations, industries, and more can be examined using qualitative research. The qualitative researcher begins from a conceptual framework—a "system of concepts, assumptions, expectations, beliefs, and theories" (Maxwell, 2005) that informs the design. The design begins with a general statement of a **research problem** or topic. This initial topic that a qualitative researcher chooses for investigation is referred to as the **focus of inquiry.** To develop the focus of inquiry, the beginning researcher needs to think about some topic in which he or she has an interest and wants to know more about. The research question may be one that comes from the investigator's observations and experiences with particular topics, settings, or groups. Qualitative problems examine the context of events, real-world setting, subjects' perspectives,

unfolding and uncontrolled events, reasons for the events, and phenomena needing exploration and explanation. For example, qualitative researchers may ask the following questions: How do social workers cope with the stress of their jobs? How do teachers in rural schools react to distance-learning technologies? How do chronically ill children deal with pain? What is the relationship of schools to indigenous communities? What is the mentoring relationship like for a beginning teacher? What is the first-year college experience like for students who have been home schooled? and How do students interact in virtual worlds?

## CHOOSING A PROBLEM

Topics that are often investigated through qualitative methods include those that defy quantification, those that are best understood within a setting, those that involve studies of group activities over time, those that involve the study of roles and behaviors, those that involve studying an organization in its entirety, those about which little is known, and those that involve closed cultures. Types of questions asked in qualitative studies include the following: What is happening? What does something mean? How are events organized or related? What are the perspectives of the participants? How do participants interact? and What are the relationships among structure, events, and participants? Answering qualitative questions can help to illuminate everyday life, provide specific details to help understand a particular setting, examine interpretations of local meanings, and illuminate differences across settings.

The following are some practical suggestions for choosing a qualitative research problem (Bogdan & Biklen, 2006). Choose a problem that is (1) interesting to you, (2) significant because it will contribute to the body of knowledge or to solving some educational problem, (3) reasonable in size and complexity so that you can complete a study of it within the time and with the resources to which you will have access, and, if possible, (4) one in which you are not directly involved. It is more difficult, though not impossible, to assume the role of researcher in contexts in which you are directly involved.

A review of the relevant literature at this step will help familiarize the researcher with theory in the area and with previous research designs and findings related to the problem. It will be most useful if you find related research that is qualitative. Although quantitative studies may provide background information on the topic, beginning qualitative researchers especially need to see how the qualitative paradigm has been used in research on relevant topics.

The choice of the research question is crucial because the question (what you really want to understand) determines the design. Maxwell (2005) describes types of research questions posed in qualitative research. **Particularizing questions** ask about a specific context—what is happening in this particular school?—and are less concerned about generalizing but, rather, focus on developing rich descriptions and interpretations. Case studies typically use particularizing questions, justifying the case selection in terms of goals and existing theory. **Generic questions** about a broader population are more typically used in quantitative research with samples selected as representative in an attempt to generalize. Generic questions can be used in qualitative research, such as with multisite studies, but must be used with caution. **Process questions** examine how things

happen—the process by which a phenomenon takes place. Questions asking about meaning, influences, and context are process oriented. **Variance questions** (questions that ask to what extent or about differences) are best answered by quantitative studies rather than by qualitative studies. **Instrumentalist questions** are formulated in terms of observable, measurable data and are the norm in quantitative studies. **Realist questions** treat unobserved phenomena (feelings, beliefs, intentions, etc.) as real and are common is qualitative studies. The difference (although perhaps subtle) between these might be seen in a study involving interviews in which the research question is posed as "What is the effect on other students when a classmate is a victim of gang violence?" (a realist question) or as "What effects do students self-report when a classmate is a victim of gang violence" (an instrumentalist approach).

Qualitative researchers do not begin a study with no questions; they begin with a base of experience, theoretical knowledge, and certain goals that drive provisional questions that may evolve with time. If the initial research questions are too diffuse, the researcher may have difficulty in the design phase or in connecting to research goals. If the research questions are too focused, it may create "tunnel vision." If unexamined assumptions are imposed through the research question, it may impose later interpretations that do not fit reality. An example is asking the question, "Why do students prefer face-to-face classes to distance education?" which assumes that they do indeed prefer one over the other. The research question must be answerable through the type of study the researcher has chosen to conduct.

After deciding on the problem and questions and determining that qualitative methodology is indeed appropriate, next you need to make decisions about the particular qualitative approach, the main data collection tools, the setting for the study, the participants, the size of sample, and the behaviors to study. A qualitative design, however, is flexible and may be changed as the researcher gets into the setting. Several criteria are available for evaluating the qualitative design to be used to answer the research question: One criterion is **informational adequacy.** That is, does the research plan maximize the possibility that the researcher will understand the setting thoroughly and accurately? A second criterion is **efficiency.** Does the plan allow adequate data to be collected in a cost- and time-effective manner? A third criterion to use is **ethical considerations.** Will the proposed method put participants at risk or violate their human rights? We discuss ethics further in a later section.

## SAMPLING

**Sampling,** or selection of participants or sites, is important in qualitative research just as it is in quantitative research. Qualitative researchers cannot observe everything about the group or site that might be relevant to the research problem. They may try to obtain a sample of observations believed to be *representative* of everything they could observe, but typically it is not a random sample.

Qualitative researchers are purposeful in selecting participants and settings. They select **purposive samples** believed to be sufficient to provide maximum insight and understanding of what they are studying. They use their experience

and knowledge to select a sample of participants that they believe can provide the relevant information about the topic or setting. In a classic work, Guba and Lincoln (1981) wrote, "Sampling is almost never representative or random but purposive, intended to exploit competing views and fresh perspectives as fully as possible" (p. 276). There is increasing recognition that when there are a small number of participants, even though selected to be representative, the process may not reflect substantial diversity that may occur in a setting, leading to concerns about whether the resulting data are valid.

The qualitative researcher should first decide what in the situation is essential to a study of the problem. Let us say a qualitative study focuses on discipline in a school system. In this case, the researcher decides what personnel to interview (principals, assistant principals, teachers, coaches, or students) and what settings (classroom, playground, or cafeteria) to observe. The researcher must develop a plan to sample personnel and settings that in his or her judgment will provide an accurate picture of attitudes and disciplinary techniques employed in the school system.

Because of the depth and extent of the information sought in qualitative studies, purposive samples are typically small. How large should the sample be? There is no general rule about the number of participants to include in a qualitative study. Of course, practical considerations such as time, money, and availability of participants influence the size of the sample. However, the primary criterion of sample size is redundancy of information. Sampling should be terminated when no new information is forthcoming from new units. A unit is an individual participant, group, organization, event, setting, document, or artifact selected as part of the qualitative study. This point is referred to as **data saturation.**

Several variations on purposive sampling are used in qualitative research, as described in Miles and Huberman (1994) and in Marshall and Rossman (2006). These strategies can be combined or mixed within studies as well to meet multiple needs or to provide triangulation.

1. **Comprehensive sampling.** In comprehensive sampling, every unit is included in the sample. For example, a study of physically disabled students in a high school would include all such students in the school. Comprehensive sampling is used when the number of units is small.
2. **Critical case sampling.** Critical case sampling involves the selection of a single unit that provides a crucial test of a theory or program. An example is selecting a single school that has decided to adopt a well-known character education program in order to change the culture of the school and observing the school during a year-long implementation to determine the impact on behaviors and interactions in the school. Examination of critical cases can enhance the ability to generalize or apply findings to other cases.
3. **Maximum variation sampling.** In maximum variation sampling, units are included that maximize differences on specified characteristics. For example, a study of U.S. high school students might include students from schools that differ in location, student characteristics, parental involvement, and other factors. This type of sampling reveals differences but may also identify commonalities across the units.

4. **Extreme, deviant, or unique case sampling.** Extreme case sampling selects units that are atypical, special, or unusual. For example, you might choose to study a high-poverty, inner-city elementary school that has achieved exemplary reading and mathematics test scores. Such a study might identify practices, teaching methods, and student characteristics that may be relevant to their superior performance.
5. **Typical case sampling.** Typical case sampling selects units that are considered typical of the phenomenon to be studied. In a study of an elementary school, you would select a school considered typical rather than a very high achieving school or a very low achieving school. This approach highlights what is normal or average.
6. **Negative or discrepant case sampling.** This method of sampling selects units that are examples of exceptions to expectations. The researcher would intentionally look for examples that appear not to confirm the theory being developed. This strategy is also called confirming and disconfirming sampling.
7. **Homogeneous sampling.** Homogeneous sampling selects a subgroup that is considered homogeneous in attitudes, experiences, and so on. For example, you might choose only a sample of special education teachers from a population of teachers. This approach may be used with focus group interviewing.
8. **Snowball, chain, or network sampling.** Snowball, chain, or network sampling occurs when the initially selected subjects suggest the names of others who would be appropriate for the sample. These next subjects might then suggest others and so on. Such sampling occurs when potential respondents are not centrally located but scattered in different sites. For example, a researcher might ask one teacher who has been nominated for an environmental science teaching award to name another teacher who he or she considers to be an exemplary environmental educator.
9. **Intensity sampling.** Intensity sampling involves selecting participants who exhibit different levels of the phenomenon of interest to the researcher. The researcher would select several cases at each of several levels of variation of the phenomenon. For example, the researcher may select some high-achieving, average-achieving, and low-achieving students or in a study of bullying, may select students who have different levels of aggressive tendencies.
10. **Stratified purposeful sampling.** Stratified purposeful sampling attempts to ensure that subgroups are represented so that comparisons can be facilitated. For example, in a study of teaching practices, experienced and inexperienced teachers would be included for observation.
11. **Random purposeful sampling.** When the potential purposeful sample is too large (e.g., when resources are limited), the credibility of the study can be enhanced by randomly selecting participants or sites from the larger group.

12. **Theoretical** or theory-based **sampling.** In theoretical sampling, the researcher begins by selecting a person or site that exemplifies the theoretical construct and continues to select new cases that reflect the developing theory to include as the research unfolds and the theory emerges.
13. **Criterion sampling.** In this type of sampling, the researcher sets the criterion and includes all cases that meet that criterion. For example, if a researcher wanted to examine practices of successful elementary reading teachers in a school district, the researcher could interview or observe all teachers whose classes had average gain scores on the state reading assessment the previous year of more than 10 points.
14. **Opportunistic sampling.** Opportunistic sampling takes advantage of new leads or unexpected opportunities.
15. **Convenience sampling.** Convenience sampling is choosing a sample based on availability, time, location, or ease of access. Convenience sampling is not recommended because it may produce evidence that is not credible. Studies of your children or your workplace are examples of convenience sampling.

## DATA COLLECTION

The next step is to choose the data collection method(s) that will be used. The most common data collection methods used in qualitative research are (1) observation, (2) interviewing, and (3) document or artifact analysis. Artifacts may include audio and video recordings, photographs, games, artwork, or other items that provide insight related to the context or participants. The researcher may use one or more of these methods in a study.

### Observation

**Observation** is a basic method for obtaining data in qualitative research and is more than just "hanging out." It is a more global type of observation than the systematic, structured observation used in quantitative research. The qualitative researcher's goal is a complete description of behavior in a specific setting rather than a numeric summary of occurrence or duration of observed behaviors. Qualitative observation usually takes place over a more extended period of time than quantitative observation. Also, qualitative observation is more likely to proceed without any prior hypotheses. Quantitative observations often use checklists and behavior observation tools developed prior to the observation to record or document observed behaviors. Qualitative observations rely on narrative or words to describe the setting, the behaviors, and the interactions. The goal is to understand complex interactions in natural settings. There are some specialized approaches to observation, such as **interaction analysis** (sometimes used in small group or classroom settings). Two types of interaction analysis are *kinesics* (the study of body movements and how those motions communicate messages) and *proxemics* (the study of how people use space). In both kinesics and proxemics, there are limitations related to cultural awareness because gestures

and nonverbal behaviors as well as use of space (e.g., interpersonal distance) can have different meanings across cultures.

There are several benefits as well as drawbacks to the use of observation for data collection. Observation may allow the researcher to determine whether what is said actually matches actions or may illuminate subtleties that may be outside the consciousness of the person or that the person cannot articulate. However, observation approaches can be very time-consuming or intrusive, and the researcher may infer meanings that are not those of the observed—a danger particularly present in cross-cultural studies.

*Choosing an Observation Site* Guided by the research question, the qualitative investigator must select a site in which to observe. You must seek to gain access to a particular site or group of people in which the topic of interest can be studied and then negotiate entry by adopting a role as a full participant observer, just an observer, or some combination of the two. Some sites are very specific to the research question. For example, if you want to investigate college majors in indigenous studies and their perspective about research, you must find universities that offer majors in indigenous studies.

In many cases, once the site is selected, the researcher has to get permission to conduct the study in that setting. It may take a long time from the first contact with a site until final approval is given. Most school systems, for instance, have specific procedures they follow before giving approval to researchers. There are concerns about disruptions in the school, loss of time for teaching, who will have access to the findings, and how the school might benefit from its involvement. Negotiating entry, adopting a role, and developing and maintaining trust with the participants in a study are fairly complex topics about which much has been written. Some more recent writings have begun addressing specific issues related to culture or to indigenous peoples in the selection of and entrance into settings for qualitative research.

*Determining Researcher Roles* Marshall and Rossman (2006) provide advice to consider related to the **researcher's role** in a study. First is the amount of "participantness" or level of involvement of the researcher in the setting, from complete participant to complete observer as described later. Second is the extent of "revealedness" or how well the participants are informed about the study, which may range from full disclosure to complete secrecy. Third is "intensiveness" or "extensiveness," indicating the amount of time and duration of time spent in the setting. Fourth is study focus, ranging from the specific with well-developed research questions and data identification needs to the diffuse with more exploratory questions and less well-defined data needs.

As noted previously, the qualitative researcher may be a **participant** in the situation being observed or a **nonparticipant.** Five stances toward observation have been identified: (1) complete participant, (2) participant as observer, (3) observer as participant, (4) complete observer, and (5) collaborative partner.

A **complete** or **covert participant** is a member of the group or context under study and focuses on the natural activity of the group without informing the group that it is under study. Sociological researchers have pretended to join

street gangs or pretended to be homeless in order to learn about such groups' feelings, relationships, and problems.

Another example is a researcher who becomes a certified teacher and takes a position in a school for the purpose of conducting research without telling anyone. In the Internet world, a researcher may join a listserve or chat room or virtual world in order to examine the online world of a particular subgroup. The ethics of the covert approach, however, may be questionable. Before undertaking such a study, you must submit your plan to your institutional review board (IRB) for approval (see Chapter 18).

In the **participant as observer** stance, the observer actively participates and becomes an insider in the event being observed so that he or she experiences events in the same way as the participants. The researcher's role is known to the people being observed. Anthropologists often are participant observers when they conduct a study of a particular culture. In most educational research, however, it is difficult for the investigator to pretend to be a member of a group and play the same role as the subjects who are being studied. It might be possible for a young researcher to be accepted in a group of college freshman in order to gather data on the freshman experience but not to become a participating member of a junior high school club.

In the **observer as participant** stance, researchers may interact with subjects enough to establish rapport but do not really become involved in the behaviors and activities of the group. Their status as observer/researcher is known to those under study. Their role is more peripheral rather than the active role played by the participant observer. For example, a researcher could focus on observing a vocational training class for welfare recipients or an organization such as Alcoholics Anonymous.

The **complete observer** is typically hidden from the group or may be simply in a public setting observing public behavior. For example, in studies that involve the use of one-way mirrors or hidden cameras to observe preschool child behaviors, the researcher is a complete observer. On the playground, the researcher could observe from some distance so that his or her presence is not noticeable to the children. Or a study of greeting behaviors of travelers at an airport may be conducted as a complete observer. The qualitative researcher simply observes and records events as they occur. No attempt is made to alter the situation in any way. These are considered naturalistic observations.

Simple **naturalistic observation** can take a great deal of time because you must wait for the behavior to occur naturally. For this reason, some researchers set up contrived naturalistic situations to elicit the behavior to be observed. Although the setup is contrived, the researcher tries to maintain the naturalness of the situation and makes the observations in a way not noticeable to the subjects. Hartshorne, May, and Shuttleworth (1928) used naturalistic contrived observation in their classic study of classroom cheating (see Chapter 8).

The **collaborative partner** stance described in action research (see Chapter 17) and feminist research has as a defining characteristic an equal partnership in the research process between the researcher and participants.

The degree of participation in an observation study is thus a continuum ranging from a complete participant at one end to a complete observer at the other. The observer must decide what degree of participation will provide the most

**PICTURE THIS**


Joe Rocco

appropriate data. It is easier to ask questions and record observations if members of the group know your purpose; furthermore, it may be more ethical to make people aware of what is going on. Being open, however, may present problems. Knowing they are being observed, group members may behave differently from the way they usually do, or they may not be truthful when answering questions. This impact of the observer on the participants being studied is called **observer effect** and can result in an inaccurate picture of the group and its interactions. There is a risk that the observer will destroy the very naturalness of the setting that he or she wants. **Observer expectation** may occur when the researcher knows the participants are associated with certain characteristics and may expect certain behaviors. In other words, expectations may cause you to see or interpret actions or events in a particular way.

Another problem with observation as a data-gathering tool is the possible effect that the observer him- or herself might have on the results. **Observer bias**

occurs when the observer's personal attitudes and values affect the observation and/or the interpretation of the observation. In participant observation, there may be a problem of the observer's getting emotionally involved in the group and hence losing objectivity. The nonparticipant observer can be more objective and emotionally detached from the group. A major question with observation is the extent to which the observations of another researcher might be different.

*Using Field Notes* The most common method of recording the data collected during observation is **field notes.** The researcher may make brief notes during the observation but then later expands his or her account of the observation as field notes. Notes may supplement information from other sources, including documents and interviews, or they may comprise the main research data. [See Marshall and Rossman (2006) for advice on note organizing and note management.] Field notes contain what the researcher has seen and heard. They have two components: (1) the *descriptive* part, which includes a complete description of the setting, the people and their reactions and interpersonal relationships, and accounts of events (who, when, and what was done); and (2) the *reflective* part, which includes the observer's personal feelings or impressions about the events, comments on the research method, decisions and problems, records of ethical issues, and speculations about data analysis. Field notes may include photographs and audio and video recordings.

The researcher's reflections are identified as **observer comments** (OCs) to distinguish them from the descriptive information. Exhibit 15.1 shows an excerpt from field notes collected as part of a study of students with disabilities in an urban high school. Some researchers organize field notes so as to have the descriptive information in one column and the observer comments in another column next to the descriptive notes.

The researcher's field notes present the data that will later be analyzed to provide an understanding of the research setting and the behavior of people within that setting. It can be said that the successful outcome of the study relies on detailed, accurate, and extensive field notes. Record everything you see, hear, or experience during the observation session. Observation sessions typically should last not more than 1 or 2 hours at a time; otherwise, so many data accumulate that it is difficult to record them all. Make the field notes as soon as possible following the observation, while the experience is still fresh in your mind. You may use audio and video recordings to facilitate data collection. See Bogdan and Biklen (1998) for guidelines to writing field notes.

Although field notes are the most common data collection technique used in observations, other techniques may include audio or video recordings or photographs. New computer technologies may allow a researcher to digitally record a chat room session or a virtual world interaction. A disadvantage of some recording methods is that participants may be conscious of the camera or other recording device and behave differently or may try to avoid being filmed or photographed.

**Joe McCloud**
**11:00 a.m. to 12:30 p.m.**
**Westwood High**
**6th Set of Notes**

*THE FOURTH-PERIOD CLASS IN MARGE'S ROOM*

I arrived at Westwood High at five minutes to eleven, the time Marge told me her fourth period started. I was dressed as usual: sport shirt, chino pants, and a Woolrich parka. The fourth period is the only time during the day when all the students who are in the "neurologically impaired/learning disability" program, better known as "Marge's program," come together. During the other periods, certain students in the program, two or three or four at most, come to her room for help with the work they are getting in other regular high school classes.

It was a warm, fortyish, promise of a spring day. There was a police patrol wagon, the kind that has benches in the back that are used for large busts, parked in the back of the big parking lot that is in front of the school. No one was sitting in it and I never heard its reason for being there. In the circular drive in front of the school was parked a United States Army car. It had insignias on the side and was a khaki color. As I walked from my car, a balding fortyish man in an Army uniform came out of the building and went to the car and sat down. Four boys and a girl also walked out of the school. All were white. They had on old dungarees and colored stenciled t-shirts with spring jackets over them. One of the boys, the tallest of the four, called out, "oink, oink, oink." This was done as he sighted the police vehicle in the back.

O.C.: This was strange to me in that I didn't think that the kids were into "the police as pigs." Somehow I associated that with another time, the early 1970s. I'm going to have to come to grips with the assumptions I have about high school due to my own experience. Sometimes I feel like Westwood is entirely different from my high school and yet this police car incident reminded me of mine.

Classes were changing when I walked down the halls. As usual there was the boy with girl standing here and there by the lockers. There were three couples that I saw. There was the occasional shout. There were no teachers outside the doors.

O.C.: The halls generally seem to be relatively unsupervised during class changes.

Two African American girls I remember walking down the hall together. They were tall and thin and had their hair elaborately braided with beads all through them. I stopped by the office to tell Mr. Talbot's (the principal) secretary that I was in the building. She gave me a warm smile.

O.C.: I feel quite comfortable in the school now. Somehow I feel like I belong. As I walk down the halls some teachers say hello. I have been going out of my way to say hello to kids that I pass. Twice I've been in a stare down with kids passing in the hall. Saying "How ya' doin?" seems to disarm them.

**Exhibit 15.1** Example of Field Notes

I walked into Marge's class and she was standing in front of the room with more people than I had ever seen in the room save for her homeroom which is right after second period. She looked like she was talking to the class or was just about to start. She was dressed as she had been on my other visits—clean, neat, well-dressed but casual. Today she had on a striped blazer, a white blouse and dark slacks. She looked up at me smiled and said: "Oh, I have a lot more people here now than the last time."

> O.C.: This was in reference to my other visits during other periods when there are only a few students. She seems self-conscious about having such a small group of students to be responsible for. Perhaps she compares herself with the regular teachers who have classes of thirty or so.

There were two women in their late twenties sitting in the room. There was only one chair left. Marge said to me something like: "We have two visitors from the central office today. One is a vocational counselor and the other is a physical therapist," but I don't remember if those were the words. I felt embarrassed coming in late. I sat down in the only chair available next to one of the women from the central office. They had on skirts and carried their pocketbooks, much more dressed up than the teachers I've seen. They sat there and observed.

Alfred (Mr. Armstrong, the teacher's aide) walked around but when he stood in one place it was over by Phil and Jeff. Marge walked about near her desk during her talk which she started by saying to the class: "Now remember, tomorrow is a fieldtrip to the Rollway Company. We all meet in the usual place, by the bus, in front of the main entrance at 8:30. Mrs. Sharp wanted me to tell you that the tour of Rollway is not specifically for you. It's not like the trip to G.M. They took you to places where you were likely to be able to get jobs. Here, it's just a general tour that everybody goes on. Many of the jobs that you will see are not for you. Some are just for people with engineering degrees. You'd better wear comfortable shoes because you may be walking for two or three hours." Maxine and Mark said: "Ooh," in protest to the walking.

She paused and said in a demanding voice: "OK, any questions? You are all going to be there. (Pause) I want you to take a piece of paper and write down some questions so that you have things to ask at the plant." She began passing out paper and at this point Jason, who was sitting next to me, made a tutting sound of disgust and said: "We got to do this?" Marge said: "I know this is too easy for you, Jason." This was said in a sarcastic way but not like a strong putdown.

> O.C.: It was like sarcasm between two people who know each other well. Marge has known many of these kids for a few years. I have to explore the implications of that for her relations with them.

**Exhibit 15.1** *Continued*

*Source*: From R. C. Bogdan and S. K. Biklen, *Qualitative research for education,* 3rd ed., pp. 109–111. Published by Allyn & Bacon, Boston. Copyright © 1998 by Pearson Education. Reprinted by permission of the publisher.

**THINK ABOUT IT 15.3**

Select a setting in which you regularly participate (e.g., office, classroom, or restaurant). Observe and take field notes for 15 minutes. What sort of activities does the physical setting encourage or discourage? How do people use the space? How do they interact and communicate with one another? How do people behave when they are on their own, in pairs, in groups? What themes, features, or patterns do you see in your field notes?

### Interviews

The **interview** is one of the most widely used and basic methods for obtaining qualitative data. Interviews are used to gather data from people about opinions, beliefs, and feelings about situations in their own words. They are used to help understand the experiences people have and the meaning they make of them rather than to test hypotheses. Interviews may provide information that cannot be obtained through observation, or they can be used to verify observations. For example, observing a teacher in a classroom tells us something about the behavior, but interviewing helps us to put the behavior in context and helps us understand actions and choices. The qualitative interview is typically more probing and open ended and less structured than the interview used in quantitative research but varies considerably in the way it is conducted.

The structure of the interview follows the extent to which the questions to be asked are developed prior to the interview. At one extreme is the **unstructured interview,** which is a conversational type of interview in which the questions arise from the situation. It is sometimes described as "a conversation with a purpose." The most data-dense interviews may be of this form. The interview is not planned in detail ahead of time; the researcher asks questions as the opportunity arises and then listens closely and uses the subjects' responses to decide on the next question. The subjects in the setting may not even realize they are being interviewed. Using the who, what, when, where, why, and how categories is generally enough guidance for the researcher to follow in asking questions.

At the other end of the continuum lies the more **structured interview,** scheduled for the specific purpose of getting certain information from the subjects. Each respondent is asked the same set of questions, but with some latitude in the sequence. Although the questions are structured, qualitative structured interviews differ from quantitative structured interviews. In the qualitative approach, the list of questions is generally more limited in length and most questions cannot be answered with yes or no or limited word responses.

In between the unstructured and structured interview is the **semi-** or **partially structured interview,** in which the area of interest is chosen and questions are formulated but the interviewer may modify the format or questions during the interview process. One characteristic that all qualitative interview formats share is that the questions are typically open ended (cannot be answered with a yes or no or simple response) and the questions are designed to reveal what is important to understand about the phenomenon under study.

Seidman (2006) put forth a framework for in-depth, phenomenologically based interviewing that involves conducting a series of three interviews spaced apart by days or weeks. The first interview focuses on the context of experience and is sometimes referred to as a focused life history. The second interview asks respondents to reconstruct details of experience in the present. The third interview is about reflecting on meaning.

An interview has the advantage of supplying large volumes of in-depth data rather quickly. Interviews provide insight on participants' perspectives, the meaning of events for the people involved, information about the site, and perhaps information on unanticipated issues. Interviews allow immediate follow-up and clarification of participants' responses. One disadvantage of the interview as a data-gathering tool is that interviewees may not be willing to share information or may even offer false information. Interviews require a great deal of time to conduct and later to transcribe the audiotapes or other notes. Interviewers need skill and practice to carry out a successful interview: They must be tactful, well prepared (so that they know what questions to ask), and listen well. Seidman (2006) describes three levels of listening: (1) listening to what the participant is saying; (2) listening to the "inner voice," the unguarded response that is not targeted to an external audience; and (3) listening while remaining aware of process and nonverbal cues. Careful listening is crucial to successful interviewing, and one of the most difficult skills to develop is to know when to be silent and when to interject the next question. Basic instructions are to listen more, talk less, and ask real questions. In addition, a good interviewer must remain on diligent guard to eliminate any personal biases that could threaten the accuracy of the data collected.

One of the most efficient ways to collect interview data is to use an audio recorder. This is much less distracting than taking notes, and it also provides a verbatim record of the responses. Informal interviews in the field are reconstructed later and included in the field notes. Videotaping can also be used to collect interview data. In recent years, interviews conducted via e-mail or chat areas or virtual rooms have been used in studies. Interviews can be conducted asynchronously, with the researcher and interviewee responding at different times, or synchronously as in a chat room or virtual room, where both researcher and interviewee are responding in real time. Good sources for additional information on interviewing include Seidman (2006) and Maykut and Morehouse (1994, Chapter 7).

Qualitative interviews might involve one-time interviews with a subject or subjects, multiple interviews with the same subject or subjects, or group interviews or focus groups. A **focus group,** which is like a group interview, typically centers on a particular issue; the trained interviewer elicits the views of the group members while noting interactions within the group. The assumption is that individual attitudes, beliefs, and choices of action do not form in a vacuum. Listening to others helps people form their own opinions. Focus groups are helpful because they bring several different perspectives into contact. The researcher gains insight into how the participants are thinking and why they are thinking as they do. Focus groups make more economical use of time and money than do individual interviews and are more socially oriented. Also, they are helpful when a researcher is studying a topic that is new or one for which little information is available. The topics and ideas expressed in the focus group can help the

researcher to identify questions and other important aspects of the phenomenon to pursue in the study.

A **focused interview** is much more flexible and open in form than the survey interview discussed in Chapter 14. The respondents are free to answer in their own words and can answer either briefly or at length. The questions asked may even vary from individual to individual. The responses are recorded by taking notes, either during the interview or immediately afterward, or with an audiotape. Focus groups are more socially oriented than individual interviews and can increase the sample size in the study, but they allow less control than individual interviews and data can be more difficult to analyze. Focus groups should not be used in emotionally charged environments or in environments with unbalanced power dynamics.

Focused interviews were used by Kagan, Dennis, Igou, and Moore (1993) to examine the effects of a staff development program on the professional lives of four elementary teachers who participated in it. In the program, experienced elementary teachers worked as adjunct faculty at a university college of education for 2-year terms. The teachers interviewed said that they did not learn radically new things in the program, but that it enabled them to clarify what they already knew about teaching and to recommit themselves to the role of career teacher. This qualitative research showed the potential of school–university partnerships for enhancing teachers' professional lives.

Focus groups typically consist of 6 to 12 people. The group should be small enough that everyone can take part in the discussion but large enough to provide diversity in perspective. Focus group discussions usually need to last at least 1 hour and possibly 2 hours. Groups should be homogeneous in terms of prestige and status to ensure comfort in expressing opinions. Exhibit 15.2 provides some sample focus group questions. For more information on conducting focus groups, see Krueger and Casey (2000).

These questions were designed to be used in a study of teacher understanding of a change to block scheduling in a high school.

1. Describe your involvement in making the decision and planning to implement block scheduling in the school.
2. How do you feel about the amount and type of communication you received during the decision and planning process?
3. How has the change to block scheduling affected your curriculum?
4. How has the change to block scheduling affected your classroom instruction practices?
5. How have the changes affected teacher attitudes in the school?
6. What effect do you think the change has had on students?
7. What feedback have you heard from students and parents about the change?
8. Think back to last year, before block scheduling. What is different in the school this year?
9. What things are going well or not so well with the new block schedule?
10. If you had a magic wand, what is the one thing you would change about the school?

**Exhibit 15.2** Sample Focus Group Questions

For both individual and focus group interviewing, some key rules are important. You must take care not to impose your own agenda or bias. You must be open to responses that are contrary to your own knowledge, beliefs, or perspectives. Care should be taken not to mention specific terms or to overcue interviewees. Interviewers should direct responses to concrete, detailed accounts rather than generalizations. You should also watch for discrepancies between the interviewee's verbal and nonverbal behaviors and note these. Strategies useful in interviewing include the probe and the pause. The *probe* is a comment that leads to more detail, such as "Can you tell me more?" "I'm not sure I understand" and "Can you give me an example?" The *pause* involves learning to be silent longer than the interviewee. The pause is at least 5 seconds and may be coupled with eye contact, depending on cultural context.

A series of strategies or suggestions for interviewing are provided by Seidman (2006) and include the following:

- Ask questions when you do not understand.
- Trust your instincts and follow your hunches.
- Explore the participant's experience, but beware of inserting the interviewer's agenda.
- Avoid leading questions.
- Ask open-ended questions.
- Follow up, but do not interrupt.
- Ask participants to talk as if they were someone else or respond to you as if you were someone else.
- Ask them to tell a story.
- Ask them to reconstruct rather than to remember.
- Ask for concrete details.
- Do not take the ebb and flow of the interview too seriously.
- Rarely share your own experiences.
- Avoid reinforcing responses, either positively ("OK," "yes," and "uh huh") or negatively.
- Explore laughter; it may reflect nervousness or be indicative of something else going on.
- Use the interview cautiously and avoid imposing your own interests.
- Tolerate silence.

**Focus group interviewing** is a specific approach used in qualitative research, but there are other approaches as well that are related to particular types of qualitative research. **Ethnographic interviewing,** grounded in anthropology, attempts to understand the participants' worldviews through gathering cultural knowledge and includes descriptive questions, structural questions, and contrast questions. **Phenomenological interviewing,** grounded in philosophy, attempts to examine lived experience through three in-depth interviews, with the first focused on the past, the second on the present, and the third on connecting the previous two to understand the "essential" experience. **Elite interviewing** selects

individuals based on their expertise—those who are considered particularly influential or well informed. For example, in a study of how 4-year public universities made decisions about their response to major budget cuts, Wojcik (2007) chose to interview the chief financial officers of the institutions. Interviewers of elites often face difficulty related to making contact, time constraints, strong personalities, and elite perceptions of the competence of the researcher.

Researchers need to exercise caution when using the interview as a data collection tool. Interviewing involves a relationship and as such is impacted by the level of rapport between the researcher and those interviewed. Differences in such areas as gender, race, ethnicity, class, hierarchy, status, age, and language used can influence that relationship and thus the responses. Also, there is always a danger that the researcher may impose values through the wording of questions or data interpretation.

### Documents and Artifacts

Qualitative researchers may use written documents or other artifacts to gain an understanding of the phenomenon under study. The term *documents* here refers to a wide range of written, physical, and visual materials, including what other authors may term artifacts. Documents may be personal, such as autobiographies, diaries, and letters; official, such as files, reports, memoranda, or minutes; or documents of popular culture, such as books, films, and videos. **Document analysis** can be of written or text-based artifacts (textbooks, novels, journals, meeting minutes, logs, announcements, policy statements, newspapers, transcripts, birth certificates, marriage records, budgets, letters, e-mail messages, etc.) or of nonwritten records (photographs, audiotapes, videotapes, computer images, websites, musical performances, televised political speeches, YouTube videos, virtual world settings, etc.). The analysis may be of existing artifacts or records, or in some cases the researcher may ask subjects to produce artifacts or documents, for example, asking participants to keep a journal about personal experiences, to write family stories, to draw pictures to express memories, or to explain thinking aloud as it is audiotaped.

Documents can be classified into four categories: (1) public records, (2) personal documents, (3) physical materials, and (4) researcher-generated documents. You are likely familiar with a wide range of public records. If you were interested in examining the changing role of the federal government in education, for example, you might examine such public records as the *Congressional Record*, federal reports, websites of congressional committees, agency reports, *Education Week*, and the *Chronicle of Higher Education*. Personal documents are typically first-person narratives and include such items as diaries, letters, home videos, scrapbooks, and more. Although these may be good sources of information about the individual's beliefs and perspectives, they are highly subjective and not necessarily reliable. Physical materials may include many objects, such as equipment, paintings, photographs, and other physical traces. For example, a researcher interested in student pride in a school might document the amount of trash left in hallways and classrooms. Researcher-generated documents are prepared by the researcher or for the researcher by the participants, as in the previous example in which participants are asked to keep a journal or to draw a picture.

If a document is written by someone who has had firsthand experience with the phenomenon under study, it is considered a **primary source.** For example, researchers have used diaries and letters written by pioneer women to understand what life was like for such women in the early years of our country. A **secondary source** is a secondhand description written by someone who may have heard about an event from others but did not directly experience it. You cannot assume that documents always provide accurate accounts of events or settings. Other sources of data should be used whenever possible to corroborate the conclusions presented in the document.

It is important to establish the authenticity of documents used in research. The researcher should attend to such issues as the history of the document, its completeness, and the original purpose of the document. Even public records may have built-in biases that need examination, and personal documents may be subject to deception or distortion (intentional or unintentional). Another limitation is that documents were generally not produced for research purposes and may be incomplete or unrepresentative. Despite limitations, documents are a good source of data. They can provide good descriptive information, are stable sources of data, and can help ground a study in its context.

Examining records and documents is an unobtrusive approach to qualitative research and can be conducted as part of a study that includes other forms of data collection or alone. The specific analytic approach is called **content analysis.**

In qualitative research, observation, interviewing, and document/artifact analysis are the most predominant data collection strategies. Marshall and Rossman (2006, pp. 133–135) provide a useful chart illustrating the strengths and weaknesses of the various qualitative data collection strategies. In addition, they describe several specialized data collection strategies.

A final note on data sources. Researchers are beginning to pay particular attention to visual and technology-based information and are beginning to raise questions about how we collect and use these data in what they perceive as a new environment that may confound some of the assumptions we make about qualitative research. Harper (2005) considers how technology has revolutionized the use of visual imagery, pointing to software that allows individuals to create multimedia-based information that is nonlinear. Markham (2005) calls for a need to adjust traditional research methods to new online environments but also to reassess the premises of qualitative inquiry. Markham indicates that these new communication technologies "privilege and highlight certain features of interaction while obscuring others" (p. 796). She points to issues related to determining what constitutes data in this new environment.

## ETHICAL CONSIDERATIONS IN QUALITATIVE RESEARCH

In other chapters, we discuss ethical issues pertaining to all types of educational research. Here, we consider ethical issues inherent in qualitative research.

1. *Kind of information obtained.* One ethical issue may result from the lengthy and personal interaction that the qualitative researcher often has with the subjects and the research site. The data collection methods involve personal

interviews and observations of participants during a prolonged time at a site. Depending on the nature of the research, there is a great likelihood of situations arising that present ethical dilemmas to the researcher. A researcher may come into unforeseen possession of knowledge about illegal activities or in some cases may have actually participated in the illegal behavior. For example, a researcher studying a high school may learn about the use and sale of illegal drugs. An educational researcher investigating teaching styles may observe emotionally abusive teachers. What should a researcher do with such "guilty knowledge"? Do you have an ethical responsibility to report this information to the appropriate authorities? The researcher must decide whether his or her main responsibility is to the research study, to the participants, or to society as a whole. You must balance the benefit of continuing the study after acquiring such guilty knowledge against your responsibilities as a good citizen. One rule that researchers must always follow is that they are legally obligated (mandated) to report any evidence of child abuse.

2. *Researcher's relationship to participant.* After spending a great amount of time observing or interviewing, the **researcher's relationship to participants** may gradually become less that of researcher and researched and more like friendship. Because the researcher is regarded as a friend, the participants trust him or her and may forget a research study is going on. Some field researchers say they obtain their best data at this point but at the same time are most ethically vulnerable. Through such interactions, Peshkin (1984) said, he "donned masks in order to remove the masks of those I wanted to observe and interview" (p. 258). The reader is referred to Peshkin's account of his ethical concerns about the deceptive role he played during an 18-month study of a fundamentalist Christian school.

   Somewhat related are the issues of **anonymity** and **confidentiality** in qualitative research. Participants may expect and have a right to anonymity, but it can be problematic in qualitative research. The researcher cannot promise anonymity if, as is usually the case, he or she knows the names of the participants. If a researcher has promised confidentiality, then he or she must try to keep that promise if at all possible. Be aware, however, that your records can be subpoenaed if a court discovers that you may have evidence about criminal activity. For this reason, researchers who conduct sensitive research do not keep records that identify individuals unless absolutely necessary. In field notes, they can use false names or code numbers to keep track of what information came from whom without revealing identities.

   The confidentiality issue may arise in educational research. Let us say you have conducted a study of the teachers in a particular school and assured them confidentiality. At the end of the study, the principal asks to see your data on one or more of the teachers. What would you do? Do you have a responsibility to the teacher(s) to whom you promised confidentiality, to the principal, or perhaps to the students who are exposed to this teacher?

3. *Reciprocation.* Another issue about which the researcher should be concerned is the issue of **reciprocity.** The people in the research setting have given of themselves to help the researcher, and he or she is indebted.

Qualitative researchers need to give participants something in return for their time, effort, cooperation, and just tolerating their extended presence. For example, assume a researcher has conducted a lengthy study of students' achievement, parental involvement, and the teachers in an elementary school. At the end of the study, it would be appropriate for the researcher and his or her team to give something back to the school. They might offer to provide a written report, present the findings at a school or neighborhood meeting, give advice or assistance on other research projects at the school, help with grant writing, and so forth. Marshall and Rossman (1995) write, "Of course, reciprocity should fit within the constraints of research and personal ethics, and within the constraints of maintaining one's role as a researcher" (p. 71).

4. *Getting permission to conduct research.* Like the quantitative researcher, the qualitative researcher must get approval for the project from his or her institution's Human Subjects Research Committee, especially if minors are included in the research. Beginning researchers should get a copy of the application form used at their institutions to see what questions they must answer. It is often more difficult for a qualitative researcher to inform the committee about the specifics of the research project because in qualitative research one does not always know what is going to happen. The problem you think you are going to investigate may not be the one you actually end up studying. Qualitative research plans may be altered as you get the study under way, and unanticipated ethical issues may arise that you will need to resolve. Usually, qualitative researchers inform the committee in a general way about the project. If it is necessary to change the focus or design of the study, then the researcher needs to get an amendment from the committee.

Recent discussions have raised questions about the very practices and definitions of ethics in qualitative research. Indigenous researchers, feminist researchers, and researchers from marginalized communities are contesting the current approaches as insensitive to their values and as reflecting imperialist and racist values. Smith (2005) and Bishop (2005) both discuss these tensions. As an example of some of the issues raised, Smith discusses how individual consent to participate in research may result in the "bleeding of knowledge" away from collective protection. Smith has the following to say about ethics in research:

> For indigenous and other marginalized communities, research ethics at a very basic level about establishing, maintaining, and nurturing reciprocal and respectful relationships, not just among people as individuals, but also with people as individuals, as collectives, as members of communities, and with humans who live in and with other entities in the environment. The abilities to enter preexisting relationships; to build, maintain, and nurture relationships; and to strengthen connectivity are important research skills in the indigenous arena. They require critical sensitivity and reciprocity of spirit by a researcher. (p. 97)

Bishop (2005) discusses concerns focused on issues of initiation (whose interests and methods), benefits (who will gain), representation (who decides what

constitutes an adequate description of social reality), legitimacy (what authority is claimed), and accountability (who has control). In discussing feminist approaches to research, Christians (2005) connects accepted ethical codes for research to philosophical roots of John Stuart Mill, Max Weber, and utilitarian ethics and argues that a different approach and different assumptions are needed. Lincoln (2005) adds to the discussion by challenging the current operation of IRBs. These and other authors argue that currently accepted ethical codes of conduct for research incorporate concepts of respect, beneficence, and justice that have been defined through Western eyes.

Other researchers are beginning to point to ethical dilemmas posed by new technologies. Markham (2005) points to the fact that ethical guidelines for Internet research vary in other countries. She points to some specific ethical challenges related to the online world:

- Perceptions of privacy among those using public sites
- Pseudonyms may not guarantee anonymity due to a variety of potential issues, including the ability of search engines to scour sources for specific quotes or identifiable writing styles of individuals
- Transient populations, which may mean different individuals participating at different points in time, even if the times are relatively close
- The near impossibility of verifying age or potential vulnerable populations
- The difficulty of actually obtaining informed consent

## SUMMARY

Qualitative studies are a distinctive type of research in education and the social sciences that can produce vivid and richly detailed accounts of human experience. These studies are based on a fundamentally different approach to the study of social reality from that which underlies the standard quantitative approach to the study of education. One chooses a qualitative design because it is the appropriate method for investigating the problem. Qualitative inquiries demand a set of skills that are not readily learned by examining textbook accounts of methods. Becoming proficient in this approach to inquiry requires fieldwork experience in negotiating access to a site, developing a researcher role, establishing and maintaining trust with participants in the study, conducting and recording interviews and observations, managing data, and performing data analysis. Qualitative research requires a great deal of time and effort.

### KEY CONCEPTS

anonymity
chain sampling
collaborative partner
complete observer
complete participant
comprehensive sampling
concern for context and meaning
confidentiality
content analysis
context stripping
convenience sampling
covert participant
criterion sampling
critical case sampling
data saturation
deviant sampling
discrepant case sampling
document analysis
efficiency
elite interviewing
emergent design

ethical considerations
ethnographic interviewing
extreme sampling
field notes
focused interview
focus group
focus group interviewing
focus of inquiry
generic questions
homogeneous sampling
human as instrument
inductive data analysis
informational adequacy
instrumentalist questions
intensity sampling
interaction analysis
interview
maximum variation sampling
naturalistic observation
negative case sampling
network sampling
nonparticipant
observation
observer as participant
observer bias
observer comments
observer effect
observer expectation
opportunistic sampling
partially structured interview
participant
participant as observer
particularizing questions
phenomenological interviewing
primary source
process questions
purposive samples
qualitative inquiry
quantitative research
random purposeful sampling
realist questions
reciprocity
researcher role
researcher's relationship to participants
research problem
sampling
secondary source
semistructured interview
snowball sampling
stratified purposeful sampling
structured interview
theoretical sampling
typical case sampling
unique case sampling
unstructured interview
variance questions

## EXERCISES

1. List four characteristics of qualitative research that distinguish it from quantitative research.
2. How does the role of the researcher differ in qualitative and quantitative research?
3. Explain the use of theory in qualitative and quantitative studies.
4. How are participants selected for a qualitative study?
5. Distinguish between participant and nonparticipant observation; give an example of each.
6. Compare the role of the hypothesis in qualitative and quantitative research.
7. Give an example of disguised participant observation.
8. What are field notes? What should they contain?
9. Suppose you want to determine the extent to which elementary school social studies textbooks are discussing the achievements of black Americans. How would you proceed to obtain data on this question?
10. What roles can a qualitative observer assume in a study?
11. School systems in the United States are hiring an increasing number of social workers. You have been asked to gather data on what these social workers do and whether their presence in schools makes any difference for student achievement, student self-concept, discipline, and relations with parents. How would you design a qualitative study to answer this question? Indicate the method and the data-gathering procedure that you think would be most effective.
12. List some ethical concerns that may arise in qualitative research.
13. What is the main difference between the interpretation of qualitative data and quantitative data?
14. What are three main threats to the accuracy of qualitative observations?
15. Give two examples of probes that can be used in an interview.
16. Determine whether each of the following questions would best be answered using quantitative or qualitative research methods:
    a. What goes on in an elementary school classroom during an average week and how do observed practices relate to the teacher's espoused teaching beliefs?

**b.** What is the relationship between education level and volume of reading done in various content areas?
**c.** How does team teaching influence classroom interactions?
**d.** How can a principal improve faculty morale?
**e.** What is the impact of AIDS on the U.S. economy?
**f.** What is the emotional impact of AIDS on at-risk health care workers?
**g.** Are the descriptions of people in the social studies textbooks multiculturally inclusive?
**h.** What is the relationship between years of teaching experience and job satisfaction?

## ANSWERS

**1. a.** The nature of the data used; qualitative research uses nonnumeric data.
**b.** Qualitative research design evolves as research gets under way; no a priori hypotheses are used.
**c.** Type of data analysis used.
**d.** Human instrument; the researcher is a participant in the study.

**2.** In quantitative studies, the researcher is expected to be detached and impartial and provide an objective portrayal. In qualitative studies, the researcher is often personally involved, and it is understood that subjectivity may influence the research.

**3.** Quantitative research is conducted to test theory, whereas qualitative research is more likely to be focused on developing theory.

**4.** Qualitative research uses purposive samples; subjects are selected because they can provide relevant information about the topic and setting investigated. A number of different strategies are used to get the sample of participants.

**5.** In participant observation, the researcher actively participates in the group being studied. In nonparticipant observation, the researcher does not participate in any activity but acts only as an observer. Examples will vary.

**6.** Qualitative researchers do not, typically, formulate hypotheses before the research and then proceed to test them. Hypotheses are formulated inductively as the research proceeds. Quantitative researchers, in contrast, formulate a hypothesis at the beginning of the research, proceed deductively to identify what should be observed if the hypothesis is correct, and then make the observations needed to test the hypothesis.

**7.** A researcher might take a job as a waitress or waiter in a neighborhood bar in order to study this particular social unit.

**8.** Field notes are the notes recorded in the field by a researcher doing a qualitative study. They consist of what the researcher sees or hears and interpretations of observations. They have two components: descriptive (what the researcher sees and hears) and reflective (what the researcher thinks or feels).

**9.** A content analysis would be the appropriate research design. One would obtain a sample of widely used social studies textbooks and then set up categories and proceed to go through the books making counts of the number of times the achievements of black Americans are mentioned within each of the categories. The categories might be males or females, or they might involve the fields in which the achievements were made, such as music, science, literature, and so on.

**10.** The observer can be a participant observer who engages fully in the activities being studied but is known to the group as a researcher. The researcher can be a nonparticipant observer, who observes but does not participate. In between these roles are combinations of the two extremes. The observer can also be a covert observer in which his or her identity is hidden from the group.

**11.** Answers will vary.

**12.** The researcher may face ethical dilemmas involving anonymity, confidentiality, what to do with "guilty knowledge," deception, and so on.

**13.** Quantitative researchers use statistical tests as a basis for interpreting the data.

Qualitative researchers analyze data inductively by organizing data into categories, identifying patterns and themes, and writing descriptive narratives.

**14.** Observer bias, observer effect, and observer expectation

**15.** Answers will vary. Examples include, Can you tell me more? and Could you give me an example?

**16. a.** Qualitative
**b.** Quantitative
**c.** Either
**d.** Qualitative
**e.** Quantitative
**f.** Qualitative
**g.** Qualitative
**h.** Quantitative

## REFERENCES


Bishop, R. (2005). Freeing ourselves from neocolonial domination in research: A Kaupapa Maori approach to creating knowledge. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (3rd ed., pp. 109–138). Thousand Oaks, CA: Sage.

Bogdan, R. C., & Biklen, S. K. (2006). *Qualitative research for education: An introduction to theory and methods* (5th ed.). Boston: Allyn & Bacon.

Christians, C. G. (2005). Ethics and politics in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (3rd ed., pp. 139–164). Thousand Oaks, CA: Sage.

Corbin, J., & Strauss, A. (2008). *Basics of qualitative research* (3rd ed.). Thousand Oaks, CA: Sage.

Denzin, N. K., & Lincoln, Y. S. (Eds.). (2005). *The Sage handbook of qualitative research* (3rd ed.). Thousand Oaks, CA: Sage.

Glesne, C., & Peshkin, A. (1992). *Becoming qualitative researchers.* White Plains, NY: Longman.

Guba, E. G., & Lincoln, Y. S. (1981). *Effective evaluation.* San Francisco: Jossey-Bass.

Harper, D. (2005). What's new visually? In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (3rd ed., pp. 747–762). Thousand Oaks, CA: Sage.

Hartshorne, H., May, M. A., & Shuttleworth, F. K. (1928). *Studies in the organization of character.* New York: Macmillan.

Kagan, D. M., Dennis, M. B., Igou, M., & Moore, P. (1993). The experience of being a teacher in residence. *American Educational Research Journal*, *30*(2), 426–443.

Kreuger, R. A., & Casey, M. A. (2000). *Focus groups: A practical guide for applied research* (3rd ed.). Thousand Oaks, CA: Sage.

Lincoln, Y. S. (2005). Institutional review boards and methodological conservatism: The challenge to and from phenomenological paradigms. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (3rd ed., pp. 165–182). Thousand Oaks, CA: Sage.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry.* Beverly Hills, CA: Sage.

Markham, A. N. (2005). The methods, politics, and ethics of representation in online ethnography. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (3rd ed., pp. 793–820). Thousand Oaks, CA: Sage.

Marshall, C., & Rossman, G. B. (2006). *Designing qualitative research* (4th ed.). Thousand Oaks, CA: Sage.

Maxwell, J. A. (2005). *Qualitative research design: An interactive approach* (2nd ed.). Thousand Oaks, CA: Sage.

Maykut, P., & Morehouse, R. (1994). *Beginning qualitative research: A philosophic and practical guide.* Washington, DC: Falmer.

Merriam, S. B. (1998). *Qualitative research and case study application in education.* San Francisco: Jossey-Bass.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis* (2nd ed.). Thousand Oaks, CA: Sage.

Peshkin, A. (1984). Odd man out: The participant observer in an absolutist setting. *Sociology of Education*, *57*, 254–264.

Seidman, I. (2006). *Interviewing as qualitative research* (3rd ed.). New York: Teachers College Press.

Smith, L. T. (2005). On tricky ground: Researching the native in the age of uncertainty. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (3rd ed., pp. 85–108). Thousand Oaks, CA: Sage.

Wojcik, R. (2007). *Responsiveness of public higher education institutions in Illinois to major budgetary changes: 2001–2004.* Unpublished doctoral dissertation, Northern Illinois University, DeKalb.

CHAPTER 16

# Types of Qualitative Research

What is the use of a book, thought Alice, without pictures or conversation.

Lewis Carroll

## INSTRUCTIONAL OBJECTIVES

After studying this chapter, the student will be able to:

1. Identify different types of qualitative research and compare and contrast them.
2. Define content or document analysis and give an example of its use in educational research.
3. Explain the differences between ethnography, narrative inquiry, and case study and give examples of each.
4. Define grounded theory and describe its role in qualitative research.
5. Explain phenomenological research and how it can be used in educational research.
6. Distinguish between primary and secondary sources in historical research.
7. Distinguish between internal and external criticism in historical research.

In Chapter 15, we discussed qualitative research in terms of how it differs from quantitative research, and how to carry it out. In this chapter, we focus on some of the more frequently used types of qualitative research.

Denzin and Lincoln (2008) in the introduction to their book, *Strategies of Qualitative Inquiry,* present a description of the qualitative researcher as a **bricoleur,** a maker of quilts. They describe many kinds of bricoleurs who use different techniques, methods, tools, and interpretive approaches to understand quite different and quite complex phenomena. The bricoleur produces a **bricolage,** which can be thought of as a quilt or montage or a puzzle that connects the parts to make a whole. In this chapter, we present several different approaches to qualitative research. As the field grows, new approaches continue to emerge as the bricoleurs refine their craft.

## QUALITATIVE TAXONOMIES

Different authors have produced various taxonomies for qualitative research. Creswell (2007) describes five approaches (narrative research, phenomenology, grounded theory, ethnography, and case study). The Denzin and Lincoln (2005) handbook of qualitative research discusses a variety of approaches, including case study, ethnography, critical ethnography, performance ethnography, grounded theory, testimonio, life history, narrative inquiry, participatory action research, clinical research, interpretive study, arts-based inquiry, autoethnography, and phenomenology. Marshall and Rossman (2006) consolidate the various approaches into three major genres: (1) individual lived experience genres, which include phenomenological research as well as life histories, testimonio, and some feminist inquiry; (2) the genre of society and culture, which includes ethnography and its variants as well as action research; and (3) the language and communication genre, which encompasses all approaches that examine both spoken language and text-based and other forms of expression.

**PICTURE THIS**


Joe Rocco

We have elected to focus on eight of what we believe are the most common approaches to qualitative research. We intentionally have not included action research in this list and have opted to dedicate an entire chapter to action research as a separate topic. It is important to remember that although the qualitative research approaches described here differ in context and participants, they overlap in many aspects. They all share the general characteristics described in Chapter 15. Also remember that there is an increasing variety of approaches to qualitative study and students are encouraged to review more detailed texts for more thorough discussion of these and other qualitative approaches. Table 16.1 provides, in alphabetic order, a brief review of the eight types of qualitative research discussed in this chapter.

**Table 16.1** Characteristics of Qualitative Research Types

| Type of Qualitative Research | Key Characteristics |
|---|---|
| Basic qualitative/interpretive research<br>"How are events, processes, and activities perceived by participants?" | Has its roots in the social sciences<br>Describes and interprets a phenomenon or process<br>Seeks to understand participants' point of view<br>Identifies recurrent patterns or themes<br>Can be based on a variety of disciplinary lenses<br>May use a variety of data collection techniques |
| Case study research<br>"What are the characteristics of this particular entity, phenomenon, person, or setting?" | Has multidisciplinary roots (business, law, medicine)<br>Focuses on a single unit<br>Produces an in-depth description<br>Is anchored in real life<br>Uses multiple data collection techniques<br>Provides a rich, holistic description of context, issue<br>Time spent examining the "unit" is important |
| Content analysis<br>"What meaning is reflected in these materials?" | Has its roots in communication studies<br>Uses analysis of written or visual materials<br>Describes the characteristics of the materials<br>Can be quantitative and qualitative |
| Ethnographic research<br>"What are the cultural patterns and perspectives of this group in its natural setting?" | Has its roots in anthropology<br>Studies the naturally occurring behavior of a group<br>Focuses on culture and societal behavior<br>Describes beliefs, values, and attitudes of a group<br>Observation is the primary data collection tool<br>Immersion in the site is important<br>Provides a holistic description of context and cultural themes |
| Grounded theory research<br>"How is an inductively derived theory about a phenomenon grounded in the data in a particular setting?" | Has its roots in sociology<br>Its goal is to inductively build a theory about a practice or phenomenon<br>Is "grounded" in the real world<br>Is a cyclical process of building a tentative theory and testing it against the data<br>Interviews and observation are the primary data collection tools<br>Typically involves observations and interviews with multiple participants or settings<br>Uses a coding process that ends in description and presentation of theory and propositions |

**Table 16.1** Characteristics of Qualitative Research Types (*continued*)

| Type of Qualitative Research | Key Characteristics |
|---|---|
| Historical research | Has its roots in the study of history |
| "How can historical events and reports be interpreted?" | Focuses on the past |
| | Written documents and artifacts are the primary data sources |
| | Seeks to authenticate sources and determine their consistency with other artifacts and documents |
| Narrative research | Has its roots in different humanities disciplines |
| "What is the story and its meaning as told through this person's experience?" | Focuses on stories about their lives told by individuals |
| | Stories may be spoken or written |
| | Emphasizes sequence and chronology |
| | Emphasizes the collaborative re-storying process |
| | Seeks to understand the lived experience of an individual or small group |
| Phenomenological research | Has its roots in philosophy |
| "What is the experience of an activity or concept from the perspective of these particular participants?" | Is concerned with the essence of a phenomenon |
| | Interprets the meaning of the participant's experience |
| | Includes investigator's firsthand experiences |
| | Interview is the primary data collection tool |
| | Typically interviews multiple subjects |
| | Attempts to determine the meaning of statements |
| | Provides a rich description of invariant structures (common characteristics or essences) |

## BASIC QUALITATIVE STUDIES

**Basic qualitative studies,** also called basic interpretative studies by some, provide rich **descriptive accounts** targeted to understanding a phenomenon, a process, or a particular point of view from the perspective of those involved. The central purpose of these studies is to understand the world or the experience of another. The underlying question the researcher is asking is "How are events, processes, and activities perceived by participants?" With roots in the social sciences, these basic studies may use a variety of techniques for collecting data, with analysis focused on identifying recurrent themes or patterns.

Basic interpretive studies are more simplistic compared to other qualitative approaches. They are not restricted to a particular phenomenon as in case studies. They do not seek to explain sociocultural aspects as in ethnography. They do not seek to enter the subject's conceptual world to explain the "essence" as in phenomenology. They do not seek to define theory as in grounded theory research. They do not convey life stories through narrative analysis, delve into history, or focus on analyzing content. These studies are, as the name implies, basic. They describe and attempt to interpret experience.

These studies are the most common qualitative studies and are used in a variety of disciplines, including education. They may use a variety of data collection techniques, including interviews and observations as well as review documents, and they may draw from diverse theoretical orientations. Data analysis typically involves categorization and development of patterns or themes, interpreted by the researcher through his or her own disciplinary lens. Often, these studies

may be shorter in duration than some of the other qualitative forms, with the researcher not as fully involved in the context. To reiterate, the basic goal of interpretive studies is to understand the meaning people make of their experiences, assuming that people create their own meanings as they interact with the world around them. The researcher starts from a frame of knowledge as a social construction.

Many beginning qualitative researchers conduct basic interpretive studies. Such questions as "How did teachers feel about the new curriculum?" or "What instructional strategies do students think are engaging and why?" might be answered in a basic qualitative study using interviewing or focus group techniques. For example, a researcher might interview students at various high school grade levels about their experiences in the classroom to try and understand their perceptions on instructional techniques. By searching for themes and patterns in the data, the researcher could attempt to answer the question about which strategies appear to engage the students. Or the researcher could videotape or conduct in-person observations of high school classrooms to answer this question.

Basic studies use the most common analysis techniques in qualitative research—coding and looking for recurring themes. Other qualitative approaches are rooted in specific traditions that influence the questions asked, the data collection process, the analysis approach, and how data are interpreted. More detail is provided about some of these specific approaches to qualitative research in the rest of this chapter.

## CASE STUDIES

Emerging from approaches in business, law, and medicine, a **case study** focuses on a single unit to produce an in-depth description that is rich and holistic. The underlying question is "What are the characteristics of this particular entity, phenomenon, person, or setting?" Case studies typically include multiple sources of data collected over time.

As indicated, case studies provide an in-depth description of a single **unit.** The "unit" can be an individual, a group, a site, a class, a policy, a program, a process, an institution, or a community. It is a single occurrence of something that the researcher is interested in examining. The unit of analysis determines whether the study is a case study or a different form of qualitative research. Case studies can answer descriptive questions (what happened) or attempt to explain why something happened by looking at a process. They are particularistic (focused on a particular phenomenon, situation, or event), descriptive (providing as an end result a thick rich description), and heuristic (focused on providing new insights). The researcher believes that something can be learned in this specific case.

A specific unit may be selected because it is unique or typical or for a variety of other reasons. The unit is defined within specific boundaries, referred to as a "**bounded system.**" To be bounded, the phenomenon must be identifiable within a specific context. If it cannot be described in such a way, case study may not be the best approach to study it. In some ways, case study is not so much about how the phenomenon is studied but, rather, about the choice of what to study. The end result of the research also is called a case study.

In comparing a case study with single-subject experiments (see Chapter 11), both may study a single individual. However, single-subject experiments focus on a single behavior or a very limited number of behaviors, whereas case studies attempt to describe the subject's entire range of behaviors and the relationship of these behaviors to the subject's history and environment. In a case study, the investigator attempts to examine an individual or unit in-depth. The emphasis is on understanding why the individual does what he or she does and how behavior changes as the individual responds to the environment. This requires detailed study for a considerable period of time; that is, it is a longitudinal approach. The investigator gathers data about the subject's present state, past experiences, environment, and how these factors relate to one another.

The greatest advantage of a case study is the possibility of depth; it seeks to understand the whole case in the totality of the environment. Not only the present actions of an individual but also his or her past, environment, emotions, and thoughts can be probed. The researcher tries to determine why an individual behaves as he or she does and not merely to record behavior. Case studies often provide an opportunity for an investigator to develop insight into basic aspects of human behavior. The intensive probing characteristic of this technique may lead to the discovery of previously unsuspected relationships.

However, case studies need not be limited to the study of individuals. Case studies are made of communities, institutions, and groups of individuals. A classic case study of a community was Lynd and Lynd's *Middletown* (1929), which described life in Muncie, Indiana, a typical average-size midwestern city. This first study was followed by *Middletown in Transition* in 1937. Another classic was Hollingshead's *Elmtown's Youth* (1949), which studied the life of adolescents in a small Illinois community. Other case studies have focused on communities ranging from the isolated mountain villages of Appalachia to life among ethnic and minority groups in inner sections of large cities. A more recent community case study by Matthew Corrigan (2007) examines race, religion, and economic change in the Republican South by focusing on one southern city.

Institutions such as schools, churches, colleges, fraternal organizations, and businesses have been the focus of case studies. Case studies have been made of groups of individuals such as gays, drug addicts, delinquents, street gangs, migratory workers, CEOs, medical students, teachers, and many others. Members of religious groups such as the Amish, the Shakers, and various other denominations have been studied. In her dissertation, Suzan Reiko Kobashigawa (2005) studied native Hawaiian literacy through a case study of three generations of one native Hawaiian family.

Three types of case studies have been described. The **intrinsic case study** is conducted to understand a particular case that may be unusual, unique, or different in some way. It does not necessarily represent other cases or a broader trait or problem for investigation. The case in and of itself is of interest to the researcher. In an **instrumental case study,** the researcher selects the case because it represents some other issue under investigation and the researcher believes this particular case can help provide insights or help to understand that issue. The case is illustrative of something under investigation. The **multiple** or **collective case study** uses several cases selected to further understand and investigate a phenomenon, population, or general condition. The researcher

believes that the phenomenon is not idiosyncratic to a single unit and studying multiple units can provide better illumination.

Case studies are anchored in real life and can provide rich detailed accounts of phenomena. The case study permits an in-depth examination of factors that explain the present status and that may influence change over time. Case studies may employ multiple methods of data collection and do not rely on a single technique. Testing, interviewing, observation, review of documents and artifacts, and other methods may be used. The distinction is that whatever techniques are used, all are focused on a single phenomenon or entity (the case) and attempt to collect information that can help understand or interpret the focus of the study.

The case study researcher starts with a particular concern or topic, and from that general area emerge **foreshadowed problems.** A purposeful choice is made of the bounded system to be studied, and then data are collected from multiple sources and analyzed. Two kinds of analysis appropriate for case studies have been described: **holistic analysis** of the entire case and **embedded analysis** that focuses on specific aspects of the case. Multiple case studies require analysis across site.

Researchers conducting case studies provide a detailed report that may build on narratives, vignettes, tables, charts, figures, visual displays, text, and more. Typically, the report is written to provide both an **emic,** or insider, **perspective** (the perspective of the individuals who are part of the case) as well as an **etic,** or outsider, **perspective** (the interpretations of the researcher). Assertions are made about the interpretations or meanings of the case.

Two examples of a case study approach used by doctoral students illustrate the kind of problems that may be investigated by researchers at this point in their careers. In one study (Hawkins, 2003), the researcher was interested in understanding the implementation of a particular character education curriculum in a school. The school had chosen not to use any of the standard character education curricula but instead to develop its own with the input of teachers, parents, and community members. The researcher used observations, document and record analysis, review of test data, and interviews over a 2-year period to study the phenomenon of program implementation for this unique curriculum. The researcher had an intrinsic interest in this unique case.

In a second study (McKenzie, 2008), the researcher wanted to understand what made some student support services programs more successful in motivating students to persist. To understand the characteristics of successful programs, he selected one program that had a higher rate of persistence among students who had participated compared to other programs. Through document analysis, on-site observations, and interviews with both graduates and staff members, the researcher found four key factors he referred to as FIRE: (1) providing students with financial advising and support, (2) helping students feel integrated into the academic and social aspects of the institution, (3) developing relationships between students and advisors/faculty and students and peers, and (4) providing students with opportunities for enlightenment—to see a broader world and to better understand themselves. These were suggested as components for other programs to consider in designing support services in this instrumental approach to a case study.

The case study has its weaknesses. Although it can have depth, it inevitably lacks breadth. The dynamics of one individual or one social unit may bear little relationship to the dynamics of others. Some argue, however, that what we learn in a particular case can be transferred to similar situations with the reader, rather than the researcher, determining what might apply to his or her context. The opportunities for insights in a case study are also opportunities for subjectivity or even prejudice. The preconceptions of an investigator can determine which behaviors are observed and which are ignored, as well as the way in which the observations are interpreted.

The novice researcher should develop a deeper understanding of case study research before attempting to conduct one. Merriam (1998), Stake (1995), and Yin (2003) all discuss the use of case study as a research approach.

**THINK ABOUT IT 16.1**

Read a case study (reported either in a journal article or in a book). Describe in your own words the bounded case and explain whether it was an intrinsic, instrumental, or collective case.

## CONTENT OR DOCUMENT ANALYSIS

With roots in communication studies, content analysis focuses on the characteristics of materials and asks "What meaning is reflected in these?" **Content** or **document analysis** is a research method applied to written or visual materials for the purpose of identifying specified characteristics of the material. The materials analyzed can be textbooks, newspapers, web pages, speeches, television programs, advertisements, musical compositions, or any of a host of other types of documents. Content analysis is widely used in education. The following are some of the purposes of content analysis in educational research:

1. *To identify bias, prejudice, or propaganda in textbooks.* For example, a researcher might analyze high school history texts in a particular school district to determine how often women are mentioned and how much discussion is given in each mention.
2. *To analyze types of errors in students' writings.* For example, you could look at students' written work to classify spelling or grammatical errors and their nature and frequency.
3. *To describe prevailing practices.* For example, you could identify the entrance requirements of Big Ten universities by analyzing their bulletins.
4. *To discover the level of difficulty of material in textbooks or other publications.* For example, you could ask, "What is the vocabulary level of the fourth-grade social studies textbooks in this district?"
5. *To discover the relative importance of, or interest in, certain topics.* For example, you might analyze popular educational research textbooks to see the coverage given to qualitative research and changes in that coverage over time.

Content analyses may be done in an **emergent design** framework, or they may be done in a quantitative research framework with variables that are specified a priori and numbers that are generated to enable the researcher to draw conclusions about these specified variables. For example, Allen, Allen, and Sigler (1993) investigated gender role stereotyping in children's books that had won the Caldecott Medal award. The purpose of the analysis was to determine if these books presented stereotypical or gender-dependent role behaviors to children through the characters in their text and pictures. They compared books from 1938 to 1940 and from 1986 to 1988 on 11 categories in which gender role stereotyping could occur, including the characters in text and pictures, the occupation of the main characters, whether the characters were active or passive, indoor or outdoor, in traditional or nontraditional roles, and so on. They found a weak trend toward equal representation in 7 of the 11 categories, although males still comprised the majority of the characters in each category. Males were characterized as active, outdoors, nontraditional, and in diverse occupations more often than females in both time periods. The researchers concluded that gender role stereotyping had decreased, but it remained prevalent in each category of investigation.

The preceding study illustrates the steps involved in a content analysis:

1. *Specifying the phenomenon to be investigated* (such as gender role stereotyping).
2. *Selecting the media from which the observations are to be made* (such as the Caldecott Medal books for specified time periods).
3. *Formulating exhaustive and mutually exclusive coding categories* so that the verbal or symbolic content can be counted (such as the categories in which gender role stereotyping could occur).
4. *Deciding on the sampling plan to be used* in order to obtain a representative sample of the documents (such as the sample consisting of all the Caldecott winners for two 2-year periods). You might decide to look at three issues per week of a newspaper over a period of 1 year, for instance, or every issue of a weekly newsmagazine for 1 year.
5. *Training the coders* so that they can consistently apply the coding scheme that has been developed and thus contribute to the reliability of the content analysis. Some coding is not so straightforward but may require inferences about whether a minority group, for example, is being portrayed positively or negatively. Several coders should be able to code the documents using the scheme and obtain consistent results. If the interrater reliability is satisfactory, you can proceed to the next step. If it is less than satisfactory, the coders may be useful in revising the category definitions to make them clearer and more complete.
6. *Analyzing the data,* which may involve just the frequencies and percentages in the various categories or may involve more descriptive accounts.

As the technological world changes and communication over the Internet becomes more prevalent, there are opportunities for content analysis to play a role in understanding changing communication patterns. Tremayne (2007)

edited a volume that presents results of content analysis of blogs, including a longitudinal analysis of blogs over a period of time.

Content analysis can be slow and time-consuming. Fortunately, computers can now carry out a content analysis quickly and accurately. Most universities have image scanners that can read printed pages and transfer the text to electronic files. Then, a computer program can be used to search through the scanned text and find words or phrases that meet specified criteria. You can obtain a listing of specified words, for example, and the frequency with which the words appear in the printed documents. Even more sophisticated computer programs, including ones that will be able to classify content and interpret meaning of words according to the context, can be expected to become available as research continues into the area known as artificial intelligence.

An advantage of content analysis is its **unobtrusiveness.** The presence of the observer does not influence what is being observed. You do not need to enlist the cooperation of subjects or get permission to do the study. Another advantage of content analyses is that they are easily replicated. Readers interested in further information on content analysis should read Krippendorf's (2004) introduction to this approach.

**THINK ABOUT IT 16.2**

As a way to examine changing language, find three blogs online written in the past year and identify any new words that are used. Develop codes for how these words are used (their meaning) in the context of the blog.

## ETHNOGRAPHIC STUDIES

"What are the cultural patterns and perspectives of this group in its natural setting?" is the underlying question addressed in ethnography. With its roots in anthropology and an emphasis on culture, group behavior is examined focusing on beliefs, attitudes, and values. Typically, observation is the primary data collection tool, although interviews are also used.

**Ethnography** is the in-depth study of naturally occurring behavior within a culture or entire social group. It seeks to understand the relationship between culture and behavior, with culture referring to the shared beliefs, values, concepts, practices, and attitudes of a specific group of people. It examines what people do and interprets why they do it. What are the meanings of these human actions and interactions within this context? Ethnographers typically describe, analyze, and interpret culture over time using observations and fieldwork as the primary data collecting strategies. The final product is a **cultural portrait** that incorporates the views of participants (emic perspective) as well as views of researcher (etic perspective). Ethnographic studies consider where people are situated and how they go about daily activities as well as cultural beliefs.

The ethnographic research method was developed by anthropologists such as Margaret Mead as a way of studying and describing human cultures. In Mead's classic study, *Coming of Age in Samoa* (1928), she studied adolescent girls in Samoa and analyzed the differences in development between Samoan and American girls. Anthropologists immerse themselves in the lives of the people they study, using primarily **extended observation** (participant and nonparticipant) and occasionally in-depth interviewing to gain clarification and more detailed information. The ethnographer undertakes the study without any a priori hypotheses to avoid predetermining what is observed or what information is elicited from informants. The ethnographer explores and tests hypotheses, but the hypotheses evolve out of the fieldwork. Ethnographers refer to the people from whom they gather information as "**informants**" rather than participants, and they study "sites" rather than individuals. The success of a researcher in conducting ethnography depends on the skill with which the researcher builds rapport and exhibits appropriate behaviors in the setting. The researcher must build a level of acceptance and trust with the informants.

The term *ethnography* is used to refer to both the work of studying a culture and the end product of the research. Ethnography has moved from anthropology to other disciplines, including education, where it has become a valuable tool in understanding the process of schooling. Ethnographic research on immigrant children and other minority populations in U.S. schools, for example, has helped educators understand these students' cultural backgrounds and the discontinuity that may exist between family and school culture that affects the success or failure of these diverse students in schools. Spindler and Hammond (2000) write that ethnography "can help teachers separate their personal cultural values from those of their students in order to see both themselves and their students more clearly" and to understand the voices of the "other" (p. 44).

Creswell (2007) describes two approaches to ethnography. **Realist ethnography** is the more traditional approach. In realist ethnography, the researcher tries to provide an objective account of the situation, typically from a third-person point of view. Standard categories are used, and factual information and closely edited quotes are presented as data. The researcher's interpretation occurs at the end. In **critical ethnography,** the researcher takes an **advocacy perspective** and has a value-laden orientation. The researcher is advocating for a marginalized group, challenging the status quo, or attempting to empower the group by giving it voice.

Ethnographic research has advantages and disadvantages. The main advantage is its observation of behavior in a real-life setting, the assumption being that human behavior can be fully understood only by knowing the setting in which it occurs. The main limitation is that the findings depend heavily on the particular researcher's observations and interpretations of the data. For example, researchers now suggest that Margaret Mead's interpretations of the data collected on the Samoan culture were completely wrong. Later research did not support her conclusions about the sexual behavior of young Samoan women but found almost the exact opposite. Mead was criticized for spending too little time in Samoa, seeing only what she wanted to see, and accepting the responses of a few young girls at face value without corroboration from other sources (Freeman, 1983). Later in this text, we discuss some strategies for enhancing

the validity of interpretations made of qualitative data. Other concerns that can arise in ethnography are related to confidentiality and anonymity and the issue of reciprocity and ownership of the data.

Spindler and Hammond (2000) describe some of the characteristics of good ethnography: (1) extended participant observation; (2) long time immersion at the site (a year would be a short time for a community study); (3) collection of large volumes of material such as notes, artifacts, audio- and videotapes, and so on; and (4) openness, which means having no specific hypotheses or even highly specific categories of observation at the start of the study.

Peshkin's work is an example of ethnographic research in education. In his 1991 work, *The Color of Strangers, the Color of Friends,* Peshkin presented a year-long, in-depth study of Riverview High School, a school in a working-class community in California with large numbers of students from different ethnic groups. He used participant observation and interviews to explore the role of ethnicity in this multiethnic school. Although Riverview had experienced ethnicity-based disorder and violence a few years earlier, Peshkin found the school characterized by ethnic peace. He described Riverview High School as a social success story, a place where strangers and friends were not sorted out on the basis of color. Social interactions that usually occur only within an ethnic group routinely took place across ethnic groups at Riverview. The students' level of academic success, however, did not match their success in social interactions. Peshkin found a persistently low level of academic success, especially among the black and Hispanic students.

In another ethnographic study, *The Imperfect Union: School Consolidation and Community Conflict,* Peshkin (1982) focused on a midwestern rural school district that was resisting state-mandated consolidation. After thorough study of the school district and its history, Peshkin illustrated how, when viewed through the eyes of the people in the district, their behavior was quite reasonable and not at all "irrational" as had been suggested. This is what ethnographic research attempts to do—build a complete understanding of a group from the perspectives of the members of that group. Other ethnographic research has provided insights on immigrant children and ethnic minorities in our schools, special needs children in schools, and teacher–pupil relationships.

As in case studies, a variety of data collection techniques may be used as part of the ethnographic study. Observation is the primary data collection tool in ethnography, but other common means of collecting data include interviewing, document analysis, and researcher diaries. It is not the data collection techniques that determine whether a study is an ethnography but rather the "sociocultural interpretation" that sets it apart from other forms of qualitative inquiry. "Ethnography is not defined by how data are collected, but rather by the lens through which the data are interpreted" (Merriam & Associates, 2002, p. 9). Ethnography can help educators and policymakers understand social and cultural issues that need to be addressed and provide insights into strategies that might be appropriate in a given culture or with particular marginalized groups. Consider the insights that could be gained from examining ethnographies such as the following: (1) *Othering America: An ethnography of US "drop outs" in Hawaii* (Pickering, 2007); (2) *Sleeping rough in Port-au-Prince: An ethnography of street children and violence in Haiti* (Kovats-Bernat, 2006); (3) *Capital and opportunity:*

*A critical ethnography of students at-risk* (Pasco, 2003); and (4) *Girl making: A cross-cultural ethnography on the process of growing up female* (Bloustien, 2003). It is recommended that students read several ethnographies to get a sense of how they are written.

Spradley (1980) identified the sequence of steps comprising the methodology of ethnographic research:

1. *Selecting an ethnographic project.* The scope of these projects can vary greatly from studying a entire complex society, such as an Inuit hunting group in Alaska, to studying a single social situation or institution, such as an urban bar, a fraternity, or a school playground. The beginner would be wise to restrict the scope of his or her project to a single social situation so that it can be completed in a reasonable time. A social situation always has three components: a place, actors, and activities.
2. *Asking ethnographic questions.* The researcher has questions in mind to guide what he or she sees and hears and the collection of data.
3. *Collecting ethnographic data.* The researcher does fieldwork to find out the activities of the people, the physical characteristics of the situation, and what it feels like to be part of the situation. This step generally begins with an overview comprising broad descriptive observations. Then, after looking at the data, you move on to more focused observations. Here, you use participant observation, in-depth interviews, and so on to gather data.
4. *Making an ethnographic record.* This step includes taking field notes and photographs, making maps, and using any other appropriate means to record the observations.
5. *Analyzing ethnographic data.* The fieldwork is always followed by data analysis, which leads to new questions and new hypotheses, more data collection and field notes, and more analysis. The cycle continues until the project is completed.
6. *Writing the ethnography.* The ethnography should be written so that the culture or group is brought to life, making readers feel that they understand the people and their way of life or the situation and the people in it. The ethnographic report can range in length from several pages to a volume or two. The writing must be detailed and concrete, not generalized or vague.

There are many types of ethnographic research called by a variety of names, all with the characteristics described previously but with different nuances. Among them are autoethnography (a self-examination within a cultural context), ethnographic case studies (a case study within a cultural perspective), critical ethnography (a study of a marginalized group), feminist ethnography (the study of women and cultural practices), and postmodern ethnography (a study of particular challenges or problems of society). More recently, we have seen references to confessional ethnography, visual ethnography, and online ethnography.

For the student interested in conducting ethnographic research, we recommend further study of this approach. You may find texts by Carspecken (1995), Fetterman (1998), LeCompte and Schensul (1999), and Madison (2005) useful.

**THINK ABOUT IT 16.3**

Select and read one of the ethnographies by Pickering (2007), Kovats-Bernat (2006), Pasco (2003), or Bloustien (2003). Determine whether it fits one of the definitions of a specialized form of ethnography as described previously. Try to identify one educational or social policy or practice recommendation based on the findings of the study.

## GROUNDED THEORY STUDIES

**Grounded theory** has its roots in sociology. Its goal is to inductively build a theory about a practice or phenomenon using interviews and observation as the primary data collection tools. This emphasis on theory distinguishes it from other qualitative approaches. Glaser and Strauss (1967) developed grounded theory as a way of formalizing the operations needed to develop theory from empirical data. This research approach focuses on gathering data about peoples' experiences in a particular context and then inductively building a theory "from the bottom up." The theory is grounded in the data and "conceptually dense"; that is, it presents many conceptual relationships that are stated as propositions pertaining to a particular context, situation, or experience. Grounded theory moves beyond description to generate or discover a theory that emerges from the data and that provides an explanation of a process, action, or interaction.

Grounded theory studies of a sociological nature have focused on victims of Alzheimer's disease and how families accommodate to the different stages, drug addiction in women, chronic illness, alcoholism, eating disorders, pain management, how families make decisions about placing members in a nursing home, and so on. A grounded theory study in education might focus on mainstreaming, mentoring new teachers, integrating minorities in private prep schools, classroom discipline, and many other topics. In education, an example of a grounded theory study is that of Brott and Myers (1999), who conceptualized the development of professional school counselor identity. The research sought to identify counselors' self-conceptualizations or personal guidelines that provide a framework for carrying out their professional roles. The theory developed through the study describes the context, conditions, and phases for a process called the *blending of influences*. Charmaz (2006) provides insight into theories developed using the grounded theory approach. She discusses a theory of bereavement by Hogan, Morse, and Tason as well as Jane Hood's theory of the role bargaining process and Patrick Biernacki's theory of identity transformation.

A brief look at some grounded research studies can help the reader get a sense of one of the most used approaches to qualitative research, popular in the nursing, medical, information technology, and business worlds as well as education. Pace (2004) developed a grounded theory of the "flow" experiences of web users engaged in information-seeking activities. Data were gathered through semistructured interviews with informants of varying gender, age, educational attainment, occupations, and web experience. "Flow" explained how users refocused on an enjoyable activity that is challenging. In another example, Drury, Francis, and Chapman (2008) used grounded theory to propose a three-phase

theory of disengagement for mature-aged students in their university journey. Data were gathered using interviews and focus groups with mature students.

In the role as the primary data-gathering instrument, the researcher asks questions about some event, experience, or social phenomenon. He or she collects data through interviews and observations. The personal open-ended interview is the primary method of data collection in grounded theory studies. The interviewer asks questions about what happened to individuals, why it happened, and what it means to them. Choose a sample where each individual has had the experience and can contribute to theory development. The study may include as many as 20 to 25 subjects who are interviewed on the topic until no new information is forthcoming **(data saturation).** To confirm or refute the theory that has developed, researchers sometimes interview another group that has had different experiences.

Documentary materials (letters, speeches, etc.) and literature can also be potential data sources. In reviewing text materials, it is important to identify whether the text is **extant** (those the researcher did not shape, such as letters or diaries) or **elicited** (those in which the researcher involved participants in writing, such as through an internet survey). Text used in the study must be situated in the context. The researcher must attend to who produced it, how it was produced, what its purpose was, how it was structured, how language is used, what meanings are embedded, and more. Data are analyzed by searching for similarities and differences among the participants' responses about the experience.

After forming categories having similar units of meaning, the researcher searches for underlying themes and relationships among the categories. This analysis of the data results in insights, conditional propositions, and questions that are pursued through further data collection. The researcher constructs tentative theoretical statements about the relationships among constructs, explores these theoretical propositions through further data collection, and so on. This cyclical process of testing the explanatory adequacy of the theoretical constructs by comparing with additional empirical data continues until the comparative analysis no longer contributes anything new **(theoretical saturation).** Thus, through induction and verification techniques, the researcher progressively elaborates a general theoretical statement well grounded in the data. The **constant comparative method of analysis** is typically used in grounded theory. In this method, the researcher compares units of data with each other to generate tentative categories, eventually reducing these to conceptual categories that evolve into an overall framework or theory. Generating the theory is not easy; it requires insight and understanding and, as indicated, many reviews of the data.

There has been some debate in the field about the approach to coding data and to analysis in grounded theory. Strauss and Corbin (1998) present a **systematic approach** that is different from the constructivist approach adopted by Charmaz. In the systematic approach, the constant comparative method is the primary analysis technique. **Open coding** is used to develop major or core categories with **axial coding** to develop categories around the core. Think of a wheel with a center and spokes extending. The spokes are all related to the central category. A visual model is developed called an axial coding paradigm. **Selective coding** is then used to develop propositions or hypotheses based on the model, showing how the categories are related. The resulting theory can take the form of a narrative statement, a picture, or a series of hypotheses. Table 16.2 summarizes

**Table 16.2** Description of Coding Types Used in Grounded Theory Studies

| Coding | Purpose |
|---|---|
| Open coding | It deals with labeling and categorizing phenomenon in the data. |
| | It uses the comparative method. |
| | Data are broken down by asking what, where, how, when, how much, etc. |
| | Similar incidents are grouped together and given the same conceptual label. |
| | Concepts are grouped together into categories. |
| | The purpose is to develop core concepts, categories, and properties. |
| Axial coding | It is designed to put data back together that were broken apart in open coding. |
| | It develops connections between a category and its subcategories (not between discrete categories). |
| | It's purpose is to develop main categories and subcategories. |
| Selective coding | It shows the connections between the discrete categories. |
| | Categories that have been developed to build the theoretical framework are integrated. |
| | It's purpose is to bring the categories together into an overall theory. |

the differences between the three types of coding in grounded theory as developed by Strauss and Corbin.

Charmaz (2006) takes a more constructivist approach, which emphasizes diverse local worlds and multiple realities and recognizes that the theory developed depends on the researcher's view. In her work, there is more emphasis on making visible hierarchies of power, communication, and opportunity. A key part of analysis is **memoing,** a process by which the researcher writes down his or her ideas about the evolving theory throughout the coding process. Memo writing is used as an intermediate step between coding and writing and is used to capture thoughts and perceived connections.

Charmaz (2006) describes the grounded theory research process as follows. The researcher begins with certain research interests and general concepts that are referred to as **sensitizing concepts** that provide a place to start. Data are collected through observations that focus on process and through intensive interviewing using broad open-ended questions followed by focus questions that elicit details. She warns of the potential impact of age, race, gender, and ideology on the interview process and cautions researchers to consider this during analysis. She also advises the researcher to attend to the nuances of language. During analysis, coding of special terms used by participants are called **in vivo codes.** These could be general terms that flag significant meanings, innovative terms, or words that reflect insider shorthand. These terms are characteristic of the social world being examined.

The researcher begins to examine the data early and begins **initial coding,** sorting data by attaching labels to segments. There are no preconceived categories and the researcher is open to all theoretical directions. Preliminary analytic notes, called **initial memos,** are written about the codes and include comparisons and ideas. The researcher defines the ideas that best fit and interpret the data and use these as tentative analytic categories. These early memos

help direct and focus further data collection. Additional data are gathered and analysis enters a **focused coding** phase. Significant or frequent initial codes are used to sort, and the researcher begins to synthesize and integrate theoretical ideas in an interactive process. **Advanced memos** are written that refine the conceptual categories and help to build and clarify the research. **Theoretical coding** follows focused coding as relationships among categories are posited. Certain categories are adopted as **theoretical concepts.** Further memos are written, and the sorting of memos and integrating of concepts lead to **diagramming.** A literature review is conducted after analysis in order to avoid seeing the world through extant ideas.

Charmaz (2006, pp. 18–19) suggests that the grounded theory researcher ask him- or herself the following questions as they immerse themselves in the task of developing a theory:

- Have I collected enough background data about persons, processes, and settings to have ready recall and to understand and portray the full range of contexts of the study?
- Have I gained detailed descriptions of a range of participants' views and actions?
- Do the data reveal what lies beneath the surface?
- Are the data sufficient to reveal changes over time?
- Have I gained multiple views of the participants' range of actions?
- Have I gathered data that enable me to develop analytic categories?
- What kind of comparisons can I make between data? How do these comparisons generate and inform my ideas?

In Denzin and Lincoln (2005, p. 528), Charmaz provides a set of succinct criteria for evaluating grounded theory studies: Is it credible? Is it original? Does it resonate? and Is it useful? Because of the complexity of grounded theory research, readers are directed to resources that will provide further explanation, such as Bryant and Charmaz (2007), Charmaz, (2006), Dey (1999), Glaser (2001), and Strauss and Corbin (1998).

**THINK ABOUT IT 16.4**

Using the key word or descriptor "grounded theory," search for an article that used this approach in qualitative research (educational research, if possible). Why (or why not) do you think grounded theory was the appropriate method to investigate this problem?

## HISTORICAL STUDIES

Historical studies are oriented to the past rather than to the present and thus use different data collection methods from those used in other qualitative approaches. **Historical research** is included in qualitative research because of its emphasis on interpretation and its use of nonnumeric data. Historical research is the attempt

to establish facts and arrive at conclusions concerning the past. The historian systematically locates, evaluates, and interprets evidence from which people can learn about the past. Based on the evidence gathered, conclusions are drawn regarding the past so as to increase knowledge of how and why past events occurred and the process by which the past became the present. The desired result is increased understanding of the present and a more rational basis for making choices.

The historian operates under different handicaps from those of researchers in other fields. Control over treatment, measurement, and sampling is limited, and there is no opportunity for replication. Although historians have no choice concerning what documents, relics, records, and artifacts survive the passage of time, they do have some limited control over what questions they will ask of these sources and what measures they will apply to them. When interviewing witnesses of past events and when searching the historical record, researchers can decide what questions to ask and what is to be measured. However, they can measure only those things that witnesses remember or the record contains. In descriptive and experimental research, investigators can attempt to control sampling; that is, they can decide for themselves whom they are going to study. Historians can study only those people for whom records and artifacts survive. If newspapers ignore a particular segment of a community and no other sources for that community exist, then historians cannot directly assess the contributions that a particular segment of a population made to that community. Another limitation impinging on historical researchers is that no assumption about the past can be made merely because no record can be found, nor can it be assumed, on the contrary, that a conspiracy of silence has distorted the historical record.

## PRIMARY AND SECONDARY SOURCES

The historian classifies materials as primary and secondary sources. **Primary sources** are original documents (correspondence, diaries, reports, etc.), relics, remains, or artifacts. These are the direct outcomes of events or the records of participants. Examples are the minutes of a school board meeting, an unedited videotape of a basketball game, or a collection of the artwork completed by a third-grade class. In the electronic world, these could include such things as blogs. Relics, remains, and artifacts are direct outcomes of events. Records of eyewitnesses are also classified as primary sources. If only the mind of an eyewitness comes between the event and the source of information about the event, it is still classified as a primary source. For example, an eyewitness or participant decided what would or would not be recorded in the school board minutes, and someone decided when the camera would be on or off and where it would be focused during the basketball game.

With **secondary sources,** the mind of a nonobserver comes between the event and the user of the record. If a newspaper reporter has been present at a school board meeting, the published report is a primary source. If the reporter relies on the minutes of the meeting or an interview with a participant to prepare the report, then the published report is a secondary source. Common examples of secondary sources are history books, articles in encyclopedias, and reviews of research. Historians seek to employ primary sources whenever possible.

### EXTERNAL AND INTERNAL CRITICISM

Two ideas that have proved useful in evaluating historical sources are the concepts of external (or lower) criticism and internal (or higher) criticism. Basically, **external criticism** asks if the evidence under consideration is authentic and, depending on the nature of the study, may involve such techniques as authentication of signatures, chemical analysis of paint, or carbon dating of artifacts. Suppose a historian has a letter describing Massachusetts schools that is believed to have been written by Horace Mann. Using external criticism, the investigator would ask questions such as the following: Is the paper of the correct age? Is the handwriting Mann's? and Are the point of view and the writing style consistent with Mann's other writings?

After the authenticity of a piece of evidence has been established, the historical investigator proceeds to **internal criticism,** which requires evaluating the worth of the evidence, for instance, whether a document provides a true report of an event. Such a question can best be answered by comparing it with others that throw light on an event or provide further information about an event and the people or circumstances surrounding it. In the example, the investigator would ask the following: Is Mann's description of the schools unbiased? and Does it agree with other contemporary descriptions of the schools?

Because historical research does have limitations, you could very well ask why it should be attempted. The fundamental reason is that there is no other way to investigate many questions. How else might you attempt to assess the effect of the Kent State shootings and other campus disorders in the spring of 1970? An advantage of historical research, and sometimes a reason for using this approach, is that it is unobtrusive. The researcher is not physically involved in the situation studied. There is no danger of experimenter–subject interaction, nor is there any need to get the permission of school authorities for the research. The historian locates appropriate documents, gathers suitable data, and draws conclusions at a distance from the situation being studied. In addition, historical research may provide new perspectives to a crisis situation.

The uninvolved nature of historical research may make it acceptable in an emotionally charged situation in which other types of research would be impossible. Because of its limitations, however, caution must be exercised in generalizing the results of historical research. Students who plan to do historical study should consult appropriate bibliographies and sources on historical methodology, such as Howell and Prevenier (2001), Storey (2004), and Thompson (2000).

## NARRATIVE RESEARCH

Narrative research has its roots in different humanities disciplines and focuses on stories (spoken or written) told by individuals about their lives. The researcher emphasizes sequence and chronology and a collaborative **re-storying** process. The key question is, "What is the story and its meaning as told through this person's experience?" The researcher seeks to understand the **lived experience** of an individual or small group.

A narrative can be any text or discourse. Narrative research provides accounts of human experiences by collecting and analyzing stories about lives. The stories are accounts of events and actions chronologically connected. Narrative research

evolved from diverse disciplines in the humanities, such as literature, art, and film, as well as from disciplines such as psychology and sociology. The narrative can be a first-person account or a collaboratively constructed narrative. Narrative studies use documents and other sources of personal information to illuminate a person's life. Written in a story format, these studies attempt to communicate the meaning of an experience and have become increasingly popular. Recently, educational trends in terms of promoting teacher self-reflection and an emphasis on giving teachers voice have been influencing narrative research.

Narrative research is not designed to be an historical record but, rather, it is designed to understand the perspective of the storyteller in the context of his or her life. Narrative analysis is categorized based on who was the author, whose perspective is presented, whether there is one or more narrative, the scope of the narrative (one or more episodes versus an entire life), and the conceptual framework of the researcher. How these characteristics play out help us to identify the narrative analysis as an autobiography, biography, personal narrative, life story, life narrative, testimonio, or oral history. An autobiography is written by the individual. In a biographical study, the researcher writes and records the experiences of another. A life history, life story, or life narrative is the portrayal of an entire life, whereas a personal experience story focuses on a single episode or multiple episodes. An oral history recounts personal reflections of events and their causes and effects. Testimonio is a form of testimonial narrative that has roots in Latin American perspectives. Denzin and Lincoln (2005) describe testimonio as "a first-person political text told by a narrator who is the protagonist, or witness to the events that are being reported" (p. 383).

Narrative analysis may be approached through a biographical lens, a psychological lens, or a linguistic lens. The narrative typically tells the story of an individual in the chronology of experience; is set within a personal, social, and historical context; and includes important themes in the lived experience (Creswell, 2007, p. 57). Data in narrative research may come from primary sources (direct participant recollection) or secondary sources (documents written by the participant). Data may be gathered via such methods as interviews with the individual or family members of close friends. Data may be gathered from written records such as diaries, journals, letters, blogs, created artifacts, e-mail exchanges, memos, photographs, memory boxes, audio recordings, story writing, or other personal, family, or social artifacts.

Andrews, Squire, and Tamboukou (2008) discuss the difficulty in conducting narrative research. It is not simply content based, it does not lend itself to the thematic approach, it does not focus on the analysis of elements of language, and there are not clear rules on analysis as there are in grounded theory or phenomenology. Narrative analysis attempts to capture individual representations of phenomena that are event and experience based. There are questions as to whether the stories represent internal individual states or external social circumstances. There are questions about whether the stories as told are adapted to the audience or whether the stories are shaped by the listener. Challenges are presented when the storyteller is asked to recall or reconstruct events from memory. Recollection changes over time, and although narratives can help organize memory, they may not be accurate representations of facts. We create our own stories, which help us see ourselves and our world and to make sense of self and experiences over time.

Narrative research has been used extensively in the medical community as a means to better understand patient care and health care professional's experiences and to improve care as well as to change policy and practice. In the medical arena, narrative research has made it possible to hear the stories of those suffering from HIV/AIDS and cancer; to listen to those going though the grief process, facing death, or recovering from severe burns; and to understand choices in having an abortion or choosing particular medical treatments. It has been used to understand the views of those with disabilities and those suffering from mental illness.

Narrative research can help in understanding change. It has been used to understand reconciliation and conflict. Also, it has been used to understand cultures and social worlds. In cultures that have oral traditions, narrative inquiry is an approach that is more welcome than some other approaches. It has been used to understand African American experiences. In education, narrative research has been used to examine the career trajectories of principals, life histories of girls who choose to take advanced mathematics classes, and adolescent culture. Visual narratives have combined oral stories with digital photographs that participants were asked to use to record their worlds. These visual narratives can be powerful in providing insights into lived experiences and shared understandings of a culture.

Narrative research has been claimed to be the best qualitative approach for capturing detailed stories of life experiences of either single individuals or small groups. The researcher gathers multiple types of stories. Information is collected about the context of stories in order to situate the individual stories within personal experiences (home, job, and school), within culture (ethnic and racial), and within history (time and place). The researcher then analyzes and reorganizes the data and shapes them into a framework with a chronological sequence. The story usually has a beginning, middle, and end. Narrative research typically has the basic elements found in novels. There is a conflict or struggle, a protagonist or main character, a sequence or plot, and some resolution. The setting influences the story. Within the stories may be epiphanies or turning points in the story line. In postmodern narratives, the researcher may deconstruct the stories to expose dichotomies or examine silences.

The following are the basic steps in conducting narrative inquiry:

Step 1: Identify a phenomenon that addresses a problem.

Step 2: Select an individual from whom you can learn about the phenomenon (or use your own experience).

Step 3: Collect stories from the individual that reflect personal experience. Records may also be collected.

Step 4: Transcribe the stories including the nuances noted by the researcher. Sometimes it is not what is said but, rather, how it is said or what is not said.

Step 5: Retranscribe the stories.

Step 6: Analyze the stories: Look for elements of time, place, plot, and scene.

Step 7. Build in past, present, and future; collect other field texts.

Step 8: Look for themes.

Step 9: Restory or retell the story in chronological sequence, attending to setting, characters, actions, problems, and resolutions.

Step 10: Write a coherent story in collaboration with the participant about the individual's personal and social experiences.

Step 11: Validate the accuracy of the story with the individual.

It is important for the researcher to collaborate with the participant throughout. The key distinguishing feature of the narrative analysis is the re-storying. Re-storying provides the causal link among ideas. Analysis involves understanding the narratives as a series of answers to underlying questions. What is the story about? Who was involved? When and where did it happen? Then what happened? What did you do? What finally happened? The coda then returns the story to the present—story over.

In considering the quality of the narrative research, Creswell (2007) provides some questions to consider. Is the story authentic? Faking a story is possible. Is the story "real"? Sometimes the participants may not be able to tell the "real story." Who owns the story? Does the researcher have permission to share it? Is the participant's voice lost? Does the researcher gain at the expense of the participant?

You must also be aware of how your own narrative might limit you in conducting narrative research. Our own lives influence what we hear, what we attend to, how we interpret things, what things we might miss, or our potential to ignore things that do not fit our frame of reference or worldview. The researcher's role is to listen, report, and interpret, but also to recognize his or her own role in the research process.

Students interested in pursuing narrative inquiry should read more on the subject before attempting to conduct such a study. Andrews et al. (2008), Clandinin and Connelly (2000), Clandinin (2006), Czarniawska (2004), and Riessman (2007) are good places to start. It may also be helpful to read some narrative research.

**THINK ABOUT IT 16.5**

Select a science teacher and interview him or her about past experiences and "stories" concerning how he or she came to teach science. Put the stories in chronological sequence. What are some key themes that come from the stories you hear?

## PHENOMENOLOGICAL RESEARCH

A **phenomenological study** is designed to describe and interpret an experience by determining the meaning of the experience as perceived by the people who have participated in it. What is the experience of an activity or concept from the perspective of particular participants? That is the key question in phenomenology. Rooted in philosophy and psychology, the assumption is that there are many ways of interpreting the same experience and that the meaning of the experience to each person is what constitutes reality. This belief is characteristic of all qualitative studies, but the element that distinguishes phenomenology from

other qualitative approaches is that the subjective experience is at the center of the inquiry. It is different in that phenomenology makes a distinction between the appearance of something and its **essence.** The central research question aims to determine the essence of the experience as "perceived by the participants." Phenomenology moves from individual experience to a universal essence and always asks what is the nature or meaning of something. Interviewing multiple individuals is the typical data collection approach.

As defined by Van Manen (2002), the term *essence*

> does not describe the whatness of a phenomenon but it describes the meaning relations that we maintain with the world. Essence is a relational term that refers to the intentionalities of our world, to possible ways of encountering and relating to the things of our world before and while we understand or think them in language and poetic and conceptual thought.

**Phenomenology** addresses questions about common human experience. A phenomenological researcher who was asked to study, for example, the integration of special needs children into a regular classroom would focus on asking what this experience means to the parties involved: the special needs children, other students, and the teacher. Whether or not the integration plan is working is not the important issue; what matters is how the students and the teacher experience it. "In the same way that ethnography focuses on culture, a phenomenological study focuses on the essence or structure of an experience. Phenomenologists are interested in showing how complex meanings are built out of simple units of direct experience" (Merriam & Associates, 2002, p. 7).

In the process of phenomenological research, the researcher must first identify a problem for which, in order to better understand its features or to develop better practices or policies, an examination of shared experience is necessary. Broad philosophical assumptions must be specified and examined with a view to **bracketing** out personal experience Interview data are typically collected from those who have experienced the phenomenon, but other data sources may be used, including observations, art, poetry, music, journals, drama, films, and novels. The key questions involve exploring what has been experienced in terms of the phenomenon and what contexts affected or influenced those experiences. Analysis is conducted by identifying significant statements or quotes (called **horizontalization**) and from those developing clusters of meaning and themes. A **textual description** is used to illuminate what was experienced, whereas the **structural description** illuminates the context that influenced the experience, how it was experienced, in what conditions and situations. After examining these descriptions and the researcher's own experiences, a **composite description** is written that conveys the overall essence of the phenomenon. This also is called the essential or **invariant structure.**

The following are examples of topics that have been studied by phenomenological researchers:

The meaning of aging

How children think about the environment

Math anxiety

The experience of insomnia

Forgiveness and reconciliation

Shame

Chronic pain

Disappointment in teacher–student relationships

Daydreaming

Loneliness

The experience of re-reading

The experience of studying a foreign language

In a study by Eva Alerby (2000), the thinking of children and young people was examined through drawings and interviews to understand how they think about the environment. The article attempts to visualize the way in which children and young people think about the environment. Anna Kirova-Petrova (2000) examined the experiences of loneliness among linguistically diverse students in order to address pedagological implications.

The participants in a phenomenological study are chosen because they have been through the experience being investigated and can share their thoughts and feelings about it. The distinguishing data collection method in a phenomenological study is the personal, unstructured interview. The interviews may be lengthy (1 or 2 hours), and there may be more than one interview with each participant. The researcher typically interviews 10 to 25 individuals who may come from a single site or different sites. Tape-recording the interviews facilitates the subsequent analysis.

Maykut and Morehouse (1994) write, "The human instrument is the only data collection instrument which is multifaceted enough and complex enough to capture the important elements of a human or human experience" (p. 27). Because of the importance of the interview in this research, it is critical that the investigator be a skilled interviewer. The questions used need to focus on meaning and be designed to elicit the "essence" of the experience from the perspective of the participants. The interviewer must be able to listen, prompt when necessary, and encourage subjects to expand and elaborate on their recollections of the experience.

Phenomenological researchers also typically explore their own experiences related to the topic of interest prior to conducting interviews in order to examine their own biases and assumptions. The concept of **bracketing** is used in phenomenological research. Bracketing involves the researcher intentionally setting aside his or her own experiences, suspending his or her own beliefs in order to take a fresh perspective based on data collected from persons who have experienced the phenomenon. The term was borrowed from mathematics and used by Husserl, the "father of phenomenology," who was a mathematician. The bracketing or suspension of belief is also referred to as **epoche.**

From an analysis of the interview data, the researcher writes descriptions of the participants' experiences and how those experiences were perceived. From the analysis, the researcher derives an overall description of the general meaning of the experience. This is done through a process called **reduction.** Think of reduction as a way to reflect. It is a phenomenological device that aims to bring aspects of meaning into nearness or focus. It is not so much a procedure as a certain thoughtful attentiveness.

Phenomenology is not the same as phenomenography, although the differences are difficult to grasp for novice researchers. **Phenomenography** has its roots in a set of studies of learning among university students in the 1970s concerning why some students were better at learning than others. Phenomenology is rooted in a philosophical method and involves the understanding of the essence of the phenomenon, whereas phenomenography has a distinct empirical orientation focused on investigating the experience of others and their subsequent perceptions of the phenomenon—their reflections on the phenomenon. Both approaches investigate the human experience.

For more detailed information on phenomenology, readers should consult Moustakas (1994) and Van Manen (1990). Van Manen also has a website that readers might find useful.

## OTHER TYPES OF QUALITATIVE RESEARCH

Although we have chosen to focus on just some of the types of qualitative research, we recognize that there are others and that the variety of approaches seems to be expanding. There are more approaches than we can mention here, but we discuss a few of note.

**Portraiture** is a form of qualitative research that seeks to join science and art in an attempt to describe complex human experiences within an organizational culture. The "portrait" is shaped by the dialogue between the researcher (portraitist) and the subject and attempts to reveal the "essence" of the subject and to tell the "central story." As Lawrence-Lightfoot and Davis (1997) explain,

> Portraiture is a method framed by the traditions and values of the phenomenological paradigm, sharing many of the techniques, standards, and goals of ethnography. But it pushes against the constraints of those traditions and practices in its explicit effort to combine empirical and aesthetic description, in its focus on the convergence of narrative and analysis, in its goal of speaking to broader audiences beyond the academy.... (p. 13)

The goal of portraiture is to paint a vivid portrait or story that reflects meaning from the perspectives of both the participants and the researcher. Data can be collected using in-depth interviews and observations over a period of time, which typically result in a personal relationship between the researcher and participants.

**Critical research** seeks to empower change through examining and critiquing assumptions. Questions focus on power relationships and the influence of race, class, and gender. Whereas other forms of qualitative research described in this text have as a key purpose the understanding of a phenomenon and the meanings people attach to events, the purpose in critical research is to critique and challenge the status quo. Critical research may analyze texts or artifacts such as film or other communication forms such as drama or dance to reveal underlying assumptions. **Feminist research** and **participatory research** are sometimes classified as critical research. Some resources for more information on this approach include Carspecken's (1995) text on critical ethnography, *Feminism and Method* by Naples (2003), and *Feminist Methodology* by Ramazanoglu and Holland (2002).

**Semiotics** and **discourse analysis** study linguistic units to examine the relationship between words and their meanings. Texts or signs and their structural relationships are the subject of study for semiotics and there is no neutral text. These approaches stress the system of relations between words as a source of meaning and view language as a social construction. Words are viewed as signs that bring together a concept and an image, derive their meaning from their place within an articulated system, are arbitrary (e.g., different languages use different terms for the same concepts), can be put together in combinations and patterns, and their use excludes other choices. Questions investigated through discourse analysis focus on verbal interaction and dialogue, and data collection strategies focus on recoded dialogue—text based, audio, or video recorded. One text, *Semiotics: The Basics* (Chandler, 2001) provides an introduction to semiotics. Paul Gee (1999) provides an introduction to discourse analysis.

## THINK ABOUT IT 16.6

For each description of a study below, determine which type of qualitative research is represented.

1. A study of in-church and televised worship sought to understand the perspectives of worshippers and to interpret how they describe the experience or worship through first-hand descriptions captured through interview processes
2. A study to investigate identity development of minority teachers during the first 3 years of their teaching career
3. A study of how a particular school assimilated immigrant children and transmitted values, beliefs, and customs of American society to them
4. A study to understand the experience of good supervision from the supervisees' perspective to determine what central factors need to be present for good supervision experiences to occur
5. A study to understand the communities created and sustained in cyberspace
6. A study of a nursing student who changed her major to elementary education and how memories of childhood experiences framed her adult schooling
7. A study to examine the effect of an impoverished, rural environment on the creativity of a gifted black child
8. A study of the extent to which discursive and nondiscursive elements of character education curricula reflected attention to multicultural aspects
9. A study of the lives (from their earliest memories) of inner-city African American and Latino American young men previously involved in destructive behavior (i.e., illegal drug marketing) who had made positive behavioral changes in order to examine personal and environmental transitions that contributed to the change
10. A study about student perceptions of the influence of race on professor credibility through nonparticipant observation, interviews, and questionnaires

**Answers**

1. Phenomenology
2. Grounded theory
3. Case study
4. Phenomenology
5. Ethnography
6. Narrative study
7. Case study
8. Content analysis
9. Basic
10. Basic

## SUMMARY

The major types of qualitative studies discussed in this text include ethnography, basic studies, case studies, content analysis, phenomenological studies, grounded theory studies, narrative analysis, and historical research. The three most widely used qualitative data collection methods are observation, interviews, and documents. The reader who is interested in conducting a qualitative research study or who wants more information on a specific approach is advised to consult relevant textbooks on qualitative methodology.

## KEY CONCEPTS

advanced memo
advocacy perspective
axial coding
basic qualitative studies
bounded system
bracketing
bricolage
bricoleur
case study
collective case study
composite description
constant comparative method of analysis
content analysis
critical ethnography
critical research
cultural portraits
data saturation
descriptive accounts
diagramming
discourse analysis
document analysis
elicited text
embedded analysis
emergent design
emic perspective
epoche
essence
ethnography
etic perspective
extant text
extended observation
external criticism
feminist research
focused coding
foreshadowed problems
grounded theory
historical research
holistic analysis
horizontalization
informants
initial coding
initial memo
instrumental case study
internal criticism
intrinsic case study
invariant structure
in vivo codes
lived experience
memoing
multiple case study
open coding
participatory research
phenomenography
phenomenological study
phenomenology
portraiture
primary sources
realist ethnography
reduction
re-storying
secondary sources
selective coding
semiotics
sensitizing concepts
structural description
systematic approach
textual description
theoretical coding
theoretical concept
theoretical saturation
unit
unobtrusiveness

## EXERCISES

1. How do case studies differ from single-subject experiments?
2. An investigator has a letter describing education in Britain in 1985. It is supposed to have been written by Prime Minister Margaret Thatcher.
   **a.** What question would be asked in external criticism?
   **b.** What question would be asked in internal criticism?
3. What are the advantages and disadvantages of historical research compared with other types of research?
4. Suggest a research question that could be answered best by using each of the following methods:
   **a.** Case study
   **b.** Ethnography
   **c.** Content analysis
   **d.** Phenomenological study

5. Assume you are going to prepare a qualitative research proposal for your research methods class or for a dissertation. Suggest a question that you would like to investigate and then the qualitative method that would be most appropriate to use.
6. Describe the disciplinary roots of the various approaches to qualitative research. How might these roots influence how studies using these methods are done?
7. Explain what a unit is in case study research. Give a specific example of a unit you might be interested in studying.
8. How is narrative research different from historical research?
9. What are the different kinds of coding used in grounded theory research?
10. Explain the concept of bracketing in phenomenology. What is it also called?
11. Determine which kind of case study would be conducted in each of the following:
    a. A researcher wants to understand the dramatic decline in behavior problems in a school district after it developed and implemented its own character development curriculum.
    b. A researcher wants to understand how elementary and secondary school principals in a large urban district use technology in their work.
    c. A researcher wants to understand the life of a home-schooled child.
12. Describe the difference between realist and critical ethnography. Which is more appealing to you and why?

## ANSWERS

1. Single-subject experiments focus on a single behavior or a very limited number of behaviors. The investigator introduces a specific treatment for the purpose of studying the effect of this treatment on the subject. A case study attempts to describe the subject's entire range of behavior as it occurs in a natural setting. The researcher observes the subject's behavior in relation to the influence of the physical, social, and psychological environment.
2. a. Was the letter really written by Margaret Thatcher?
   b. Does it accurately describe education in Britain in 1985?
3. One advantage of historical research is the unlikelihood of researcher or experimental interaction effects confounding interpretation of findings. A historical perspective can deal with issues and past situations that cannot be handled experimentally. The main disadvantage is the lack of experimental control, which makes unequivocal interpretation of data and generalization difficult. There is also the possibility of gathering inadequate or inaccurate information that is not verifiable.
4. Answers will vary.
5. Answers will vary.
6. Basic qualitative research—social sciences; case study—business, law, and medicine; content analysis—communication studies; ethnography—anthropology; grounded theory—sociology; historical—history; narrative analysis—humanities; and phenomenology—philosophy. Responses to how these roots influence the research may vary.
7. A unit in case study research is a single instance or occurrence of something that has boundaries so that it can be defined easily. The unit could be an individual, a group, a community, an institution, a program, and so on. Examples will vary.
8. Narrative research attempts to tell the story or "lived experience" of an individual or group from a particular perspective, whereas historical research attempts to establish facts and arrive at conclusions concerning the past.
9. The systematic approach uses open, axial, and selective coding processes. The constructivist approach uses initial, focused, and theoretical coding processes.
10. Bracketing is a process used to help the researcher set aside his or her own

experiences and beliefs in order to take a fresh perspective. It is also called epoche.

11. **a.** Intrinsic case study of the program
    **b.** Collective case study of elementary and secondary principals
    **c.** Instrumental case study of a home-schooled child
12. Realist ethnography attempts to provide an objective third-person account, whereas in critical ethnography the researcher takes an advocacy position based on values. Answers will vary on preferences and reasons why.

## REFERENCES


Alerby, E. (2000). A way of visualizing children's and young people's thoughts about the environment. *Environmental Education Research, 6,* 205–223.

Allen, A. M., Allen, D. N., & Sigler, G. (1993). Changes in sex-role stereotyping in Caldecott Medal award picture books 1938–1988. *Journal of Research in Childhood Education, 7,* 67–73.

Andrews, M., Squire, C., & Tamboukou, M. (Eds.). (2008). *Doing narrative research.* London: Sage.

Atkinson, J. W., & Birch, D. (1978). *An introduction to motivation.* New York: Van Nostrand.

Atkinson, R. (1998). *The life story interview.* Thousand Oaks, CA: Sage.

Bloustien, G. (2003). *Girl making: A cross-cultural ethnography on the process of growing up female.* New York: Berghahn Books.

Brott, P. E., & Myers, J. E. (1999). Development of professional school counselor identity: A grounded theory. *Professional School Counseling, 2*(5), 339–348.

Bryant, A., & Charmaz, K. (Eds.). (2007). *The Sage handbook of grounded theory.* London: Sage.

Carspecken, P. F. (1995). *Critical ethnography in educational research: A theoretical and practical guide.* New York: Routledge & Kegan.

Chandler, D. (2001). *Semiotics: The basics.* London: Routledge.

Charmaz, K. (2005). *Grounded theory in the 21st century: Applications for advancing social justice studies.* In N. K. Denzin and Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (3rd ed.).Thousand Oaks, CA: Sage.

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis.* London: Sage.

Clandinin, D. J. (Ed.). (2006). *Handbook of narrative inquiry: Mapping a methodology.* Thousand Oaks, CA: Sage.

Clandinin, D. J., & Connelly, F. M. (2000). *Narrative inquiry: Experience and story in qualitative research.* San Francisco: Jossey-Bass.

Corrigan, M. T. (2007). *Race, religion, and economic change in the Republican South: A case study of a southern city*. Gainesville: University Press of Florida.

Creswell, J. W. (2007). *Qualitative inquiry & research design* (2nd ed.). Thousand Oaks, CA: Sage.

Czarniawska, B. (2004). *Narratives in social science research.* London: Sage.

Denzin, N. K., & Lincoln, Y. S. (2005). *The Sage handbook of qualitative research* (3rd ed.). Thousand Oaks, CA: Sage.

Denzin, N. K., & Lincoln, Y. S. (2008). *Strategies of qualitative inquiry* (3rd ed.). Thousand Oaks, CA: Sage.

Dey, I. (1999). *Grounding grounded theory: Guidelines for qualitative inquiry.* San Diego: Academic Press.

Drury, V., Francis, K., & Chapman, Y. (2008). Letting go and moving on: A grounded theory analysis of disengaging from university and becoming a registered nurse. *Nurse Education Today, 28*(7), 783–789.

Fetterman, D. M. (1998). *Ethnography: Step by step* (2nd ed.). Thousand Oaks, CA: Sage.

Freeman, D. (1983). *Margaret Mead and Samoa: The making and unmaking of an anthropological myth.* Cambridge, MA: Harvard University Press.

Gee, J. P. (1999). *An introduction to discourse analysis: Theory and method.* London: Routledge.

Glaser, B. G. (2001). *The grounded theory perspective: Conceptualization contrasted with description.* Mill Valley, CA: Sociology Press.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory.* Chicago: Aldine.

Glesne, C., & Peshkin, A. (1992). *Becoming qualitative researchers.* White Plains, NY: Longman.

Goetz, J. P., & LeCompte, M. D. (1993). *Ethnography and qualitative design in educational research.* New York: Academic Press.

Hawkins, R. (2003). *The seven habits of highly effective students: A case study in character education and school cultural transformation.* Unpublished doctoral dissertation, Northern Illinois University, DeKalb.

Hollingshead, A. B. (1949). *Elmstown's youth.* New York: Wiley.

Howell, M., & Prevenier, W. (2001). *From reliable sources: An introduction to historical research.* Ithaca, NY: Cornell University Press.

Kobashigawa, S. R. (2005). *Native Hawaiian literacies: A case study of three generations of one native Hawaiian family.* Unpublished doctoral dissertation, Indiana University of Pennsylvania, Indiana, PA.

Kirova-Petrova, A. (2000). Researching young children's lived experience of loneliness: Pedagogical implications for linguistically diverse learners. *Alberta Journal of Educational Research, 46*(2), 99–116.

Kovats-Bernat, J. C. (2006). *Sleeping rough in Port-au-Prince: An ethnography of street children and violence in Haiti.* Gainesville: University Press of Florida.

Krippendorf, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed). Thousand Oaks, CA: Sage.

Labov, W. (1972). *Language in the inner city: Studies in the Black English vernacular.* Oxford: Basil Blackwell.

Lawrence-Lightfoot, S., & Davis, J. H. (1997). *The art and science of portraiture.* San Francisco: Jossey-Bass.

LeCompte, M. D., & Schensul, J. J. (Eds.). (1999). *Designing and conducting ethnographic research.* Walnut Creek, CA: AltaMira.

Lynd, R. S., & Lynd, H. M. (1929). *Middletown.* New York: Harcourt Brace.

Lynd, R. S., & Lynd, H. M. (1937). *Middletown in transition.* New York: Harcourt, Brace, and World.

Madison, D. S. (2005). *Critical ethnography: Method, ethics, and performance.* Thousand Oaks, CA: Sage.

Marshall, C., & Rossman, G. B. (1999). *Designing qualitative research* (3rd ed.). Thousand Oaks, CA: Sage.

Marshall, C., & Rossman, G. B. (2006). *Designing qualitative research* (4th ed.). Thousand Oaks, CA: Sage.

Marton, F. (1994). Phenomenography. In T. Husén & T. N. Postlethwaite (Eds.), *The international encyclopedia of education* (2nd ed., pp. 4424–4429). Elmsford, NY: Pergamon.

Maykut, P., & Morehouse, R. (1994). *Beginning qualitative research: A philosophic and practical guide.* Washington, DC: Falmer.

McKenzie, J. (2008). *The role of student support services programs in motivating students to persist in postsecondary institutions.* Unpublished doctoral dissertation, Northern Illinois University, DeKalb.

Mead, M. (1928). *Coming of age in Samoa.* New York: William Morrow.

Merriam, S. B. (1998). *Qualitative research and case study applications in education.* San Francisco: Jossey-Bass.

Merriam, S. B., & Associates. (2002). *Qualitative research in practice.* San Francisco: Jossey-Bass.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50*(9), 741–749.

Moustakas, C. (1994). *Phenomenological research.* Thousand Oaks, CA: Sage.

Naples, N. A. (2003). *Feminism and method: Ethnography, discourse, and activist research.* London: Routledge.

Pace, S. (2004). A grounded theory of the flow of experiences of Web users. *International Journal of Human–Computer Studies, 60*(3), 327–363.

Pasco, R. J. (2003). *Capital and opportunity: A critical ethnography of students at-risk.* Lanham, MD: University Press of America.

Peshkin, A. (1982). *The imperfect union: School consolidation and community conflict.* Chicago: University of Chicago Press.

Peshkin, A. (1984). Odd man out: The participant observer in an absolutist setting. *Sociology of Education, 57,* 254–264.

Peshkin, A. (1991). *The color of strangers, the color of friends: The play of ethnicity in school and community.* Chicago: University of Chicago Press.

Pickering, L. (2007). *Othering America: An ethnography of US "drop outs" in Hawaii.* Unpublished doctoral dissertation, University of Manchester, Manchester, UK.

Ramazanoglu, C., & Holland, J. (2002). *Feminist methodology: Challenge and choices.* Thousand Oaks, CA: Sage.

Riessman, C. K. (2007). *Narrative methods in the human sciences.* Thousand Oaks, CA: Sage.

Spindler, G. D., & Hammond, L. (2000). The use of anthropological methods in educational research: Two perspectives. *Harvard Educational Review, 70*(1), 39–48.

Spradley, J. P. (1980). *Participant observation.* New York: Holt, Rinehart & Winston.

Stake, R. (1995). *The art of case study research.* Thousand Oaks, CA: Sage.

Storey, W. K. (2004). *Writing history: A guide for students.* New York: Oxford University Press.

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: Grounded theory procedures and techniques* (2nd ed.). Newbury Park, CA: Sage.

Thompson, P. R. (2000). *The voice of the past: Oral history* (3rd ed.). Oxford, UK: Oxford University Press.

Tremayne, M. (Ed.). (2007). *Blogging, citizenship, and the future of media.* New York: Routledge.

Van Manen, M. (1990). *Researching lived experience: Human science for an action sensitive pedagogy.* Albany: State University of New York Press.

Van Manen, M. (2002). www.phenomenologyonline.com. Accessed September 5, 2008.

Yin, R. K. (2003). *Case study research: Design and methods* (3rd ed.). Thousand Oaks, CA: Sage.

CHAPTER 17

# Analyzing and Reporting Qualitative Research

Analysis is the act of giving meaning to data.

Corbin and Strauss (2008)

## INSTRUCTIONAL OBJECTIVES

After studying this chapter, the student will be able to:

1. Describe the general procedures followed in the analysis of qualitative data.
2. State the role of categories in qualitative data analysis.
3. Identify the most commonly used approaches to data analysis in qualitative research.
4. Specify the difference between data analysis and data interpretation.
5. Explain the role of technology in qualitative analysis.
6. Describe the major approaches to asserting validity in qualitative studies.
7. Describe the major strategies for enhancing reliability in qualitative studies.
8. Identify the components of a qualitative report.
9. State some of the criteria to use in evaluating a qualitative study.

Data analysis and interpretation in qualitative research are far from simple. The challenge facing the researcher is to make sense of copious amounts of data and to construct a framework for communicating the essence of what the data reveal. These data may be text based, as in notes, transcripts, and other written materials, or they may be from audio or visual sources, such as recordings, pictures, or video. Data analysis is a process whereby researchers systematically search and arrange their data in order to increase their understanding of the data and to enable them to present what they learned to others. In the previous chapter, we briefly commented on some analysis and reporting strategies associated with specific approaches to qualitative research, such as open and axial coding and the constant comparative method used in grounded theory studies or restorying in narrative inquiry. In this chapter, we focus on the basic, most common approaches to qualitative analysis and reporting. Readers should review references noted in the previous chapter for more detailed information on analysis techniques for specific qualitative approaches.

# ANALYZING QUALITATIVE DATA

**Data analysis** is the most complex and mysterious phase of qualitative research. Data analysis in qualitative research is a time-consuming and difficult process because typically the researcher faces massive amounts of field notes, interview transcripts, audio recordings, video data, reflections, or information from documents, all of which must be examined and interpreted. Analysis involves reducing and organizing the data, synthesizing, searching for significant patterns, and discovering what is important. The researcher must organize what he or she has seen, heard, and read and try to make sense of it in order to create explanations, develop theories, or pose new questions.

Qualitative analysis is messy and nonlinear. Data analysis in qualitative research is often done concurrently or simultaneously with data collection through an iterative, recursive, and dynamic process. Data collection, analysis, and report writing do not occur in distinct steps as is typical in quantitative studies. Maxwell (2005, p. 95) describes this process as follows: "The experienced qualitative researcher begins data analysis immediately after finishing the first interview or observation, and continues to analyze the data as long as he or she is working on the research, stopping briefly to write reports and papers."

All qualitative analysis involves attempts to comprehend the phenomenon under study, synthesize information and explain relationships, theorize about how and why the relationships appear as they do, and reconnect the new knowledge with what is already known. The task of analyzing qualitative data can appear overwhelming but becomes manageable when broken down into key stages. Creswell (2007) describes the **data analysis spiral.** Once data are collected, they must be organized and managed. The researcher must become engaged with the data through reading and reflecting. Then data must be described, classified, and interpreted. Finally, the researcher represents or visualizes the data for others. Creswell describes how this spiral fits with various approaches to qualitative inquiry (narrative, phenomenology, grounded theory, ethnography, and case study).In different texts, the approaches to analysis of qualitative data vary slightly, but we believe they can be described in three stages as noted in Table 17.1: (1) organizing and familiarizing, (2) coding and reducing, and (3) interpreting and representing.

## FAMILIARIZING AND ORGANIZING

The first stage in analyzing qualitative data involves **familiarization** and **organization** so that the data can be easily retrieved. Initially, the researcher should become familiar with the data through reading and rereading notes and transcripts, viewing and reviewing videotapes, and listening repeatedly to audiotapes. The researcher must be immersed in the data. Field notes, audiotapes, videotapes, observer comments, and other data must be put into a form ready for analysis. Preferably, **transcriptions** should be made of all data, including tape-recorded interviews, focus groups, video recordings, and handwritten field notes. Words should be transcribed directly to avoid potential bias in selection or interpretation that may occur with summarizing. As transcriptions are made, you should include notes that provide nonverbal information (e.g., gestures and laughter) that can give added meaning. While transcribing, do not change words or phrases to make them grammatically correct because it may

**Table 17.1** Stages of Qualitative Data Analysis

| Stage | Reference: Creswell (2007) | Marshall and Rossman (2006) | Maxwell (2005) | Wolcott (1994) |
|---|---|---|---|---|
| Organizing and familiarizing | Data managing<br>Reading/ memoing | Organizing the data<br>Immersion in the data | Reading, listening | Describe and highlight |
| Coding and reducing | Describing<br>Classifying | Generating categories and themes<br>Coding the data | Coding, memoing, categorizing | Analyze and identify patterns |
| Interpreting and representing | Interpreting | Offering interpretations through analytic memos | Connecting | Contextualize |
| | Representing, visualizing | Searching for alternative understandings<br>Writing the report | Reporting | Display findings |

inadvertently change the sense or meaning of what was said. Typically, during transcription, identifiable information is stripped to ensure confidentiality. Pseudonyms may be given.

Analysis without making transcripts is possible but not recommended. The researcher can take notes while playing a recording, for example. Although this may be less time-consuming than transcription, you are more likely to miss important information, and your analysis may have less depth and comprehensiveness. Once transcriptions have been completed, continue to read and reread the data. As you are thus familiarizing yourself with the data, write notes or memos (also called a **reflective log**) to capture your thoughts as they occur. Notes may be taken in the margins of the transcripts indicating key ideas. Once you have made notes in the margins, review them and make a complete list of the different types of information you see. This is an essential preliminary step to developing a coding scheme.

The major task of organizing the large body of information begins after familiarization. Start with creating a complete list of data sources. Files can be organized in a variety of ways, for example, by interview, by questions, by people, or by places. Photocopy all data pages and work from the copies. Always keep backup copies of the original data.

Today, most qualitative researchers rely on technologies and software programs to help them organize and manage the vast amounts of data common in qualitative research. There are a growing number of software programs available, which are continually being changed and upgraded. Students are advised to review the different products based on their needs. For example, some products support both text and multimedia materials, some are specific for content analysis, and some are designed for use with audio or video data. Often, you can access a demonstration copy from websites. The following eight criteria have been suggested to help in selecting computer-based programs for qualitative data analysis: (1) ease of use and operating system compatibility; (2) type of data

accepted (e.g., text, images, and multimedia); (3) ability to read, search, and review text; (4) memo writing functions; (5) categorization processes and ability to develop, apply, display, and revise codes; (6) analysis features that enable sorting, combining, comparing, and developing concept maps; (7) the ability to import data from and export data to quantitative programs; and (8) the ability to support multiple researchers and merge databases.

We suggest that before data are entered into a computer program, researchers prepare a log indicating the types of data according to dates, names, times, and places where, when, and with whom they were gathered.

### THINK ABOUT IT 17.1

Find a volunteer and audio- or videotape an informal interview (5 to 10 minutes) on the topic of his or her favorite learning experience. Transcribe the interview, including as much nonverbal material as possible. Then answer the following questions: (1) How long did it take to transcribe the tape? (2) What nonverbal communications were you able to include and how did they inform your understanding of the words? (3) Did you ask any leading questions or miss important cues? (4) As you listen to the tape while reading the transcript, did you change any of the words (is the transcription accurate)? (5) What key ideas did you write in the margins? (6) If you videotaped, when you reviewed the tape were there visual cues you provided to the respondent? and (7) How did the respondent's physical gestures and postures match (or not match) with verbal responses?

**Answers**

Answers will vary.

## CODING AND REDUCING

After familiarizing yourself with the data and organizing them for easy retrieval, you can begin the **coding** and **reducing** process. This is the core of qualitative analysis and includes the identification of categories and themes and their refinement. Wiersma (2000, p. 203) suggests that coding is analogous to getting ready for a rummage sale: You sort the stuff for the sale into categories—housewares, clothing, furniture, books, and so on. You might further subdivide the categories: The clothing category would include children's clothing, teenagers' clothing, and adults' clothing. Then each of these clothing categories could be further subdivided; for example, the children's clothing could be categorized as infants', toddlers', and school-age children's clothes; each of these categories could be further subdivided into boys' and girls' clothes.

Coding is about developing concepts from the raw data. The first step in coding is referred to as **axial coding, open coding, preliminary coding,** or **provisional coding.** The most common approach is to read and reread all the data and sort them by looking for units of meaning—words, phrases, sentences, subjects' ways of thinking, behavior patterns, and events that seem to appear regularly and that seem important. Each unit of meaning label should be understandable without any additional information. These initial codes are likely to be modified later. These codes may be named from actual words of respondents (**in vivo codes**) or may be

names created by the researcher to include a variety of ways an underlying concept is expressed. Or the researcher may begin with a framework for analysis, a set of a priori concepts derived from the literature that are used as codes.

When coding initially, use as many codes as needed. These will be reduced later. Labeling or coding of items is done in order to begin to recognize differences and similarities in the data. The initial coding leads to the development of tentative **categories.** These may be refined and reconceptualized as the analysis process continues. Understand that in qualitative analysis, the boundaries of the categories and themes involve interpretive judgment. As the "big picture" begins to emerge, you may rethink the categories and reestablish boundaries.

The researcher initially goes through all the data and identifies each unit (word, paragraph, sentence, etc.) with the appropriate code. As the codes are developed, some qualitative researchers count the frequency with which these codes appear, which may give some insight into the importance of that category of meaning. Others argue that is inappropriate to count in this way and may in fact obscure important concepts. Recall that the goal of qualitative coding is not to count but to break apart the data and rearrange them into categories that facilitate comparisons within and between and to develop theoretical concepts. Codes can represent information you expected to find, information you did not expect to find, and interesting or unusual information.

Corbin and Strauss (2008) and Creswell (2007) provide some analytic strategies or thinking techniques to help in developing codes:

1. *Ask questions about the data* (What does that mean? Who is involved? How does that happen? How often does it happen? etc.). The typical who, what, when, where, how, and why questions help you dig deeper into the data and provide insight.
2. *Make comparisons*. One approach to comparison is sometimes referred to as constant comparison. Each data element is compared with previous elements to look for similarities or differences. Another approach is theoretical comparisons, which focus on the property or dimensional levels. We know something because it has certain properties or elements that we recognize from the literature or from experience.
3. *Consider different meanings or implied meanings of words or phrases*. Sometimes we miss important elements because we do not consider that our interpretation of a word or phrase might not be that of the respondent.
4. *Use the flip-flop technique*. This involves consciously turning a concept "inside out" or "upside down," looking at the opposite of the term.
5. *Draw on personal experience*. Consider your own experiences and how they might help you draw insights into what the participants are describing.
6. *Wave the red flag*. Certain words when used in the data may provide a hint that further analysis is warranted—for example, words such as "always," "never," or "everyone knows."
7. *Look at language*. For example, how people use terms such as "I," "we," or "it" can provide insights into their views.
8. *Look at expressed emotion*. Emotions and feelings can be clues to meaning.

9. *Look for words that indicate time*. The use of words that reference time (e.g., "when," "after," "then," "before," and "since") can frame events, help identify context, or indicate shifts in perceptions.
10. *Think in metaphors and similes*. The words used can paint vivid pictures that help our interpretation or understanding. For example, I have often been referred to as an "energizer bunny." What words come to mind that could be used in coding meaning?
11. *Look for negative cases*. Searching for the negative can help you consciously consider alternative explanations.
12. *Consider narrative structure*. How is the story organized? How does it begin and end? These could provide clues to importance.
13. *Dismantle dichotomies*. Look for examples in which there are distinctions drawn (e.g., nature/nurture and positive/negative).
14. *Examine silences*. Look for what is not said, what is excluded.
15. *Attend to disruptions*. Are there places where something does not make sense, where the respondent does not continue, or where there are contradictions?

In addition to using the previous prompts while reading and coding data, researchers must also consider the assumptions, cultural beliefs, societal structures, and knowledge levels of the participants. Developing codes enables the researcher to physically separate material bearing on a given topic from other material and is a crucial step in organizing the data. The goal is to develop a set of codes that provide a reasonable reconstruction of the data that have been collected. Some researchers start with a smaller number of codes (**lean coding**) and then expand to a fuller set of codes.

## THINK ABOUT IT 17.2

Review your transcript from the interview conducted in Think About It 17.1. Review the responses using the analysis techniques listed just discussed.

**Answers**

Answers will vary.

After all the data are coded, the researcher places all units having the same coding together. Researchers used to do this manually by cutting with scissors according to the codes and putting material with like codes together in a marked folder. Some coded using highlighters, with a master sheet indicating which colors were connected with which categories. Now, researchers use computers to code and then connect codes. After initial coding, you begin searching for categories, themes, or dimensions. You began by taking the data apart and breaking it into small pieces. Now the task is to reduce the large number of individual codes into a manageable set of categories. Categories are one level of abstraction above the initial codes.

Once coding of a transcript is completed and all items with a particular code are placed together, review the sets of items to ensure they belong together. Then begin considering whether codes can be put together into larger categories. Some items may belong in more than one category. Once categories have been established, consider whether some categories may be linked to create major categories or themes. **Themes** are a level of abstraction beyond the categories. This process of coding, categorizing, and developing themes will be repeated for each transcript or set of data. Then merge these sets together, reviewing categories and themes. Once all coding and categorization is complete, go back and look at the original transcripts and review any areas not coded and consider whether these now fit into categories.

The categories developed from the coded data should be internally consistent and distinct from one another. Think of them as bins in which to organize the coded data. The categories one investigator uses to organize qualitative data may not be the same categories another researcher would use to organize the same data. The researcher's interests and style and the research question influence to a great extent the categories chosen. **Organizational categories** typically could have been anticipated and may have been established prior to data collection. However, these are not usually a good mechanism for making sense of the actual data. Substantive or theoretical categories help provide insights.

**Substantive categories** are primarily descriptive and not generally related to more abstract theories. Emic substantive categories are those from participants' perspectives and words. However, substantive categories are more likely to be based on the researcher's interpretation of what is going on (etic categories). **Theoretical categories** are more abstract and can be from prior theory or from inductively developed theory. They are more likely to be etic categories. Often, novice researchers use only organizational categories.

The number of categories developed will depend on the type of data collected, the focus of inquiry, and the researcher's analytic skills. If the number of categories is very large, they should be collapsed into a manageable number. In qualitative data analysis, the researcher conducts **inductive coding** of the individual data pieces and simultaneous comparison of units of meaning. Once the data pieces have been coded, they are merged into categories that are refined through several iterations. After the categories have been refined, the researcher explores the relationships or patterns across categories, identifying major themes. The integration of data into themes yields an understanding of the context and people being studied.

Whereas coding "fractures" data into small bits and categorizing re-sorts the codes into bins, developing themes involves looking for connections and relationships. Generating themes is the most difficult and challenging process of analysis. Figure 17.1 shows the levels of data analysis in a **data pyramid.**

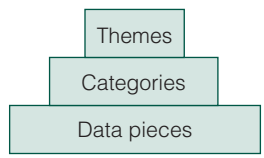



**Figure 17.1** Data Pyramid

Following are comments from a series of focus groups with Hispanic/Latino parents for a study that examined parent involvement in urban elementary schools (LaViolette, 2001). The researcher held focus groups in four different schools. The study identified points of conflict between parents and the schools, obstacles to resolution, and strategies to resolution. For each category, only one comment is shown to demonstrate coding. The letter and number combinations identify which parent (1, 2, 3, 4, etc.) in which focus group (A, B, C, D, etc.) made the comment.

**Category: Reporting Student Progress**

(Parents indicated they were not informed when their child was not being successful in school.)

J7: He's right about whenever, the teachers you know, if you want to get feedback from them as to the progress of your child. Whenever I do want to get feedback as to how my child is doing, where is it that he or she is lacking, I always call and schedule an appointment with the teachers, and that has worked for me. But it is true, I'll agree with him, if you don't ask, they don't say anything. And I do think that maybe that shouldn't be the case because by the time we dig, we find out as to how our child is doing, maybe that will be too late.

**Category: Communicating about Incidents**

(Parents were concerned about poor communication about incidents.)

J6: In my girls classroom, all those kids were playing with a used condom. They were throwing it to each other. And when I tried to talk about it, they [the school] ignored us. They didn't want to talk about it because that was a very delicate subject.

**Category: Understanding Policies, Programs, and Procedures**

(Parents indicated they did not have adequate information or understanding about policies, programs, and procedures.)

B11: The only thing is that my son and my daughter they have been here, but I'm from another country. I speak another language at home and for some reason, they put them in the ESL program, which they don't speak another language. I speak the language, but they don't. My husband is born here. He speaks good English. But when I call the school, they say yes, once you speak another language in the house, they should be in ESL class, which I didn't understand because they are good in writing and reading. I don't know why they have to go to ESL. I call them many times and I say I want to take them out of the ESL program, but they say no, once I speak another language in the house, they have to be in ESL.

The researcher later combined these three categories [(1) reporting student progress, (2) communicating about incidents, and (3) understanding policies, programs, and procedures] into a single theme: *We are not being told what we need to know*. A focus group was also held with principals, and comparisons were made between parent and principal identification of points of conflict. The researcher later represented her themes from the comparisons of principal and parent comments regarding conflict in the graphic shown in Figure 17.2.

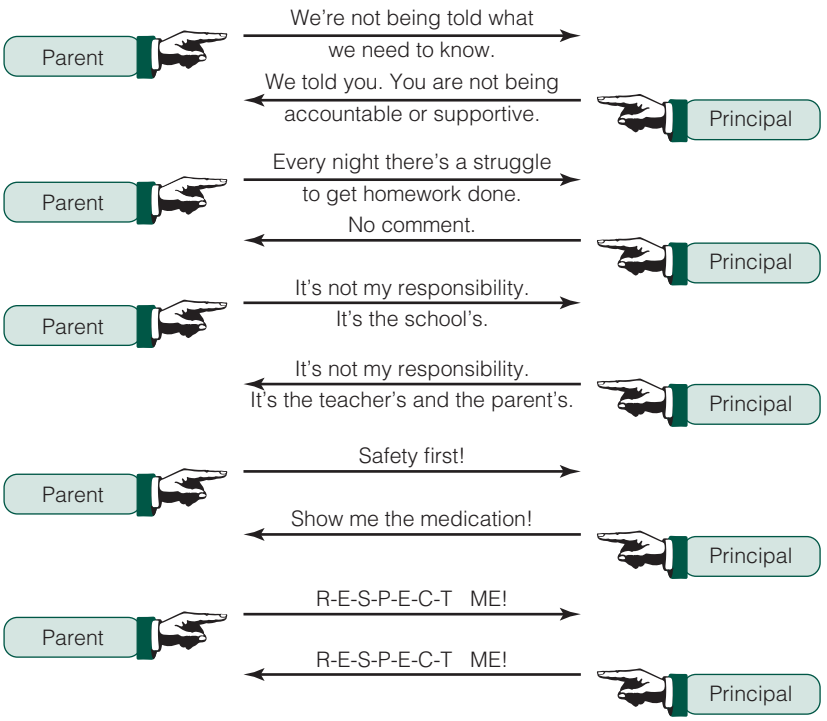



**Figure 17.2** Perceptions of Conflict Reported by Elementary School Principals and Parents in an Urban Setting

*Source:* From *School–community conflict: Perceptions of elementary-school principals and parents in an urban setting,* Dissertation by Sharon A. LaViolette, 2001. Reprinted by permission.

## THINK ABOUT IT 17.3

A pilot professional development program was presented by a university to a group of urban teachers. At the conclusion of the program, a focus group was held with the teachers to determine their perceptions of the impact of the program. A brief portion of the lengthy transcript is provided here. Review the transcript and code all the items of interest. (Fac indicates a facilitator comment. Other labels, such as A3 and B11, identify respondents.)

**Responses**

Fac: What aspects of the program have been most useful in your professional practice?

B2: There was theory with practical applications, which is extremely useful in my classroom. The program gave me a spark, made me more innovative. The different modeling of instruction was wonderful.

A2: The team working activities were most beneficial, especially when there is a class with lots of bickering. You can help kids see they can help each other. We were provided with cooperative frames rather than competitive games and it has made a difference in my classes.

B11: We learned how to be resourceful in using our schools and communities. It gave us a boost and a confidence to go out and do that.

A1: I liked the principles of brain-based learning and found ways to teach things from that perspective. Because we were taught that way and the instructor modeled it through a variety of ways, I could then connect things to my own experiences and that helped me have more ideas.

B13: Yeah, the brain-based learning and cognitive systems really caught my attention. It meant something to me personally. Through my reflective journal I learned a lot about myself. It occurred to me if I felt this way, perhaps my students felt the same way. It altered the way I might respond to a child, especially in terms of following through on work and behavior. I also learned from talking to parents. You know things theoretically, but then it's real. Now I have to figure out how to bring all this back to the classroom.

**Answers**

The codes are intended as examples. Different researchers may come up with different names for concepts contained in the transcripts.

| Comment | Code |
|---|---|
| B2 | Practical application, modeling |
| A2 | Teamwork/cooperation |
| B11 | Connection to community |
| A1 | Brain-based learning, modeling, self-reflection |
| B13 | Brain-based learning, self-reflection |

Once all data are sorted into major and minor categories, look at the range of categories and determine whether some fit together into themes. For example, in the partial transcript presented previously, there are a number of categories that might be combined into a theme called "connecting." The next step is to summarize; here, you begin to see what is in the data.

Examine all entries with the same code and then merge these categories into patterns by finding links and connections among categories. This process further integrates the data, and you can begin to make some statements about relationships and themes in the data. For example, when investigating classroom management in a large high school, you might first code the data into categories dealing with teachers' attitudes, students' attitudes, effects of certain techniques, descriptions of behavior, the setting, and so on. Then summarize by trying to find relationships among the categories. Summarizing is beginning to tell the stories and to make connections among stories. The researcher begins to make meaning of the categories and themes, to connect them. At this stage, connections may be displayed using graphs, charts, concept maps, or other visual representations of the patterns observed.

Perhaps the best known qualitative analysis strategy is the **constant comparative method,** which combines inductive category coding with simultaneous comparison of all units of meaning obtained. The researcher examines each new unit of meaning (topics or concepts) to determine its distinctive characteristics. Then he or she compares categories and groups them with similar categories. If there are no similar units of meaning, form a new category. Thus, there is a process of continuous refinement; initial categories may be changed, merged, or omitted; new categories are generated; and new relationships can be discovered. Another approach used in analysis is the **negative case analysis** or **discrepant data analysis.** Look for data that are negative or discrepant from

the main body of data collected. Negative cases contradict the main category or pattern, and discrepant cases provide a different perspective on a category or pattern. This approach provides a counterbalance to a researcher's tendency to hold on to first impressions or hunches. Some researchers revise based on these instances; others go with the substantial accumulation of positive instances.

Coding and data analysis are clearly affected by the particular qualitative approach taken. Whether coding by hand or, more likely, using a computer program, different qualitative approaches frame the codes for units, categories, and themes. Creswell (2007, pp. 170–172) provides templates of coding schemes used in computer-generated analysis. These are helpful in visually seeing how the initial codes build to the higher level of abstraction.

Once data have been completely analyzed and themes developed, the next step is interpretation, going beyond the descriptive data to extract meaning and insights from the data. You state what you found that was important, why it is important, and what can be learned from it.

## INTERPRETING AND REPRESENTING

Interpreting involves reflecting about the words and acts of the study's participants and abstracting important understandings from them. It is an inductive process in which you make generalizations based on the connections and common aspects among the categories and patterns. You may develop hypotheses that have evolved during the analysis. **Interpretation** is about bringing out the meaning, telling the story, providing an explanation, and developing plausible explanations.

Interpreting qualitative data is difficult because there are no set rules to follow. The quality of the interpretation depends on the background, perspective, knowledge, and theoretical orientation of the researcher and the intellectual skills he or she brings to the task. Unlike quantitative research, there are no statistical tests of significance to facilitate interpretation of qualitative data. In interpreting qualitative data, you confirm what you already know is supported by the data, you question what you think you know and eliminate misconceptions, and you illuminate new insights and important things that you did not know but should have known. Although interpretation is personal and proceeds without set rules, this does not mean that the qualitative researcher can rely strictly on personal feelings when interpreting the data. The interpretation cannot be just a figment of your imagination but must be supported by the data.

The approach to interpretation is clearly affected by the particular qualitative approach used in the study. In narrative inquiry, the researcher is describing the larger meaning of the story. In phenomenology, the researcher is attempting to examine the "essence" of something using textual and structural descriptions. In grounded theory, interpretation involves developing propositions. Ethnographers are trying to make sense of how a culture works. A case study develops generalizations. Each of these qualitative approaches evolved from differing disciplinary bases that clearly influence how the data are interpreted.

How the researcher represents the findings is also influenced by the qualitative approach taken and by the types of data analyzed. **Representation** involves how the data are presented. Are there graphs, pictures, diagrams, figures, or frameworks? The procedure used most frequently by qualitative researchers is to report by themes, topics, or cases and demonstrate these through descriptive detail. Some findings are best represented visually, and newer technologies are providing alternatives for how qualitative research is shared. Performance ethnography and some other approaches often present findings in alternative formats—for example, using multimedia presentations or theater.

Generally, ethnographies have narrative augmented by tables, figures, and sketches. Case studies present a detailed view of a case using narrative, tables, and figures. Grounded theory usually concludes with a visual model of a theory or a set of propositions. Phenomenology provides a narration related to the essence with discussion of meaning; it describes what the participants experienced and how they experienced it. Narrative inquiries result in life stories or similar presentations and are likely to include such elements as plot, characters, setting, and interaction and often are written chronologically.

## REPORTING QUALITATIVE RESEARCH

Remember that the purpose of the research is to make clear to others what you studied, how you studied it, what you observed, and how you interpreted it. The last step for the qualitative researcher is to represent to others what he or she saw, heard, and now understands about the phenomenon that was investigated. Although representation takes a variety of formats for qualitative researchers, the student or novice researcher is most likely expected to produce a more typical research report. These qualitative reports generally include the following elements or some variation of them:

1. Abstract
2. Introduction
3. Research design, steps to ensure credibility (validity) and dependability (reliability)
4. Methods, site and sample selection, data collection methods, data analysis procedures
5. Findings
6. Interpretations and implications
7. References
8. Appendix

### Abstract

An abstract is a very brief summary of the major aspects of the qualitative inquiry: problem, design, methods, and outcomes. Complete and concise abstracts are very helpful to readers who are conducting literature reviews.

### Introduction

In the introduction, the writer states the purpose of the research study and provides the reader with some background of the problem and the need for the study. State the focus of inquiry as a question or a statement. Indicate how you became interested in the topic and how the question evolved. Present and discuss any relevant research in this section because, typically, there is not a separate section for review of the literature. The introduction in a qualitative report may also contain an interesting story or quote to capture the reader's interest. In addition, the researcher's preliminary biases or suppositions should be revealed.

### Research Design

In the section on research design, explain the qualitative approach used and why you chose it to investigate the problem. Also discuss the steps taken to ensure credibility, dependability, transferability, and confirmability of the study, such as triangulation, audit trail, member checks, and so forth (these are discussed in the next section of this chapter). It is important to describe how bias was controlled and the limitations of the study.

### Methods

In the methods section, describe the research method that was followed to obtain the findings. The reader should not have to wonder about what was done, to whom, or how it was done. Some aspects to cover in separate parts of the methods section are as follows.

*Site and Sample Selection* The researcher describes the site of the study and the participants involved. The investigator describes the participants demographically, how they were selected, and how entry was gained into the site; gives a detailed description of the site; and so on. Researchers generally use fictitious names of people and places to protect the privacy of the participants and the site.

*Data Collection Methods* The section on data collection methods describes the methods used to gather the data (interviews, observation, document analysis, and so on) and explains why these methods were chosen. Because the *main* instrument in qualitative research is the human instrument, it is important that the writer give some personal or professional information about him-or herself that might be relevant to the inquiry. A complete discussion of the methodology is essential so that readers of the report can understand how the researcher reached conclusions and can agree or disagree with those conclusions.

*Data Analysis Procedures* The researcher should describe the approach taken in the analysis of the data. If you used the constant comparative procedure, for example, make this explicit.

### Findings

This section contains the major findings or propositions relevant to the original focus of inquiry that are revealed in the data. Findings may be reported by data collection method (survey, interviews, documents, etc.), by cases (School 1, School 2, etc.),

or by theme or topics (real-world connections, active learning, extrinsic rewards, etc.). Because of the massive amount of data in a qualitative inquiry, you cannot report everything you found. One challenge for qualitative researchers is deciding what not to include in their reports. The researcher must rank the outcomes primarily on the basis of their relevance and significance. Use direct quotes and field note excerpts to illustrate the outcomes of the study and to help the reader to vicariously experience the research setting and to better understand how the conclusions were reached. Unlike quantitative research, this section also includes your interpretations of the data. You may also relate findings to the literature.

### Interpretations and Implications

In the final section, you respond to the implicit question, So what? Here, the researcher tries to make sense of the findings. You interpret what you have found. You discuss the meaning of the outcomes reported in the previous section and state major conclusions and implications of the study. You should relate the study to previous research and suggest directions for future research efforts. This is also the section in which new or integrating theories may be proposed.

### References

A complete alphabetical listing of all works cited in the report is presented in the reference section of the report.

### Appendix

The appendix includes interview schedules and other documentation that will help the reader understand the report.

Reports of qualitative inquiry will vary in form depending on the approach used (as mentioned previously) as well as the nature of the publication in which they appear (e.g., an article prepared for the *American Educational Research Journal* will differ from a monograph or book-length study). It will be helpful to the student who is writing a qualitative report to read qualitative studies that use the same approach and to examine journal articles to see how the reports are organized. Some suggested journals that publish qualitative research include the *International Journal of Qualitative Studies in Education, Qualitative Inquiry*, and *The Qualitative Report* (an online journal). Merriam and Associates (2002) present sample articles showing basic interpretive, phenomenological, grounded theory, case study, ethnographic, and narrative analysis approaches to writing reports.

Writing is extremely important in qualitative research. Qualitative reports are generally heavily narrative in form and contain rich descriptions of setting and context. This "**thick description**" is intended to place readers vividly in the research setting so that they can follow the logical processes that the researcher went through in collecting the data. The role of the writer is to "tell the story the data tell." In contrast to the more technical and structured style of quantitative reports, a qualitative report is more like a story and may have very little, if any, technical language. A qualitative report may not follow a conventional organizational format. Literature may not be found in a separate section of the report but may be woven into the findings with the themes identified from the current study connected to the work of others.

The approaches to writing a qualitative report vary. Some are more journalistic in style, interpreting the connections among events and people. Some approach it as a translator would, interpreting others' worlds. Some appear as realist tales presenting observations and descriptions in great detail. Some may appear as a historical report, following a chronology. Some may read as confessional tales from the author's viewpoint. Some use techniques found in drama and other artistic modes. Qualitative reports convey the participants' thoughts, feelings, and experiences in their own words as much as possible. Note that it is acceptable in qualitative reports for writers to refer to themselves in the first person, to distinguish their opinions from those of the participants. Qualitative reports more frequently incorporate visual representations (computer-generated graphics, pictures, videos, audio, etc.) as well as rely extensively on quotes to help the reader "see" and "experience" the participants' world.

## TECHNOLOGY IN QUALITATIVE ANALYSIS

Analysis of qualitative data can seem overwhelming. Analysis by hand in earlier years relied on literally cutting and pasting from printed text and the use of color-coded highlighters. Today, most qualitative researchers use technologies, although technology can never replace the human researcher in terms of thought process and interpretation. Using a computer can save an enormous amount of time in compiling and sorting data. For example, you can take a laptop computer into the field to write notes during observations. You can use word processors to transcribe and enter data. Databases and spreadsheets can help sort and organize data. Graphics software can help show visual representations of how data connect.

Technology can be used as a source of data (audio, film, visual digital materials, and web-based information), as a data collection tool (e.g., videotaping, audiotaping, e-interviews, and digital pictures), and as an analysis tool. Technology can be used to organize and manage data or for actual data analysis. Lee and Fielding (1995) introduced the term *computer-assisted qualitative data analysis software* (CAQDAS). The computer is a tool that can be used for executing specific mechanical aspects of the research process and to make your work faster, more accurate, and less tedious. However, the computer does not think for the researcher. The researcher must decide what to enter into the computer and what to ask the computer to do. The computer cannot reflect or transform data into meaningful findings.

Word processors (e.g., Microsoft Word) and common databases (e.g., Access) can be used to record and manage qualitative data. Consider how the "search" function in Word can be used to locate key words in a text. More recent technologies can help examine features and relationships in text; help develop conceptual maps and models; and support rich text, diagrams, and the incorporation of videos, images, movies, and multimedia. Others facilitate collaboration among researchers and data sharing.

CAQDAS is commonly used for the following:

1. *Data storage and management*. Technology can help with the storage and management of the data. There are software packages in which the researcher can enter raw data directly or import data from word-processed documents. Some software will handle visual materials (photographs,

diagrams, video, web page links, and multimedia). Most will let you annotate and edit text after it has been entered. Some have indexing capability and allow you to add your own identification information (name, date, etc.) to better organize the data. In addition to storing data, the researcher can store memos and documentation of insights.

2. *Data searching and retrieval*. Searching for and retrieving data can be simplified using technology. Most qualitative analysis software allows you to search text data for particular words or phrases. The programs can count specific words or phrases to assist in analysis. Most use Boolean operators (AND, OR, and NOT) to refine searches, and typically these programs retrieve data in context (e.g., showing you the word within a phrase or paragraph) and include identification codes. Test as well as image segments can be located and coded.
3. *Coding*. Although the human researcher must identify the codes, technology can assist in the mechanics of coding. Sections of data can be highlighted in the text and assigned codes. In much widely used qualitative analysis software, coded items are stored as nodes that can be searched like documents and can be given titles and descriptions by the researcher. Technology can assist with locating segments that relate to multiple codes, making comparisons among code labels, looking for interrelationships among codes. Templates for coding can also be adapted to the particular qualitative approach being used.
4. *Developing and testing theories*. Qualitative software can aid in developing theories by helping to show relationships between codes. Relationships can be displayed through graphical representations. Some use trees or mind maps that can help provide overviews of how the data are connected.
5. *Visual representation*. Software can help conceptualize different levels (codes, categories, and themes). Concept mapping, diagrams, and newer abilities to incorporate multimedia allow better visualization.
6. *Writing reports*. Software programs can also be used to produce reports or to create printouts of an entire data set in one node or code, which can then be included verbatim or quotes can be pulled out to use in the text. Many of the programs include fields for memos as the researcher documents his or her thinking process, and these memos can be printed out and analyzed as well. Forms can also be created in many of the products.
7. *Collaboration and linking*. New technologies allow for multiple researchers to share data and provide for collaborative tools. Formats such as HTML and PDF can help link a wide range of materials with different formats. Data collection and analysis can occur in a distributed format.

Computers have revolutionized the handling and manipulation of qualitative data. However, beware of letting the computer guide the analysis rather than the researcher thinking about the data. Some writers in the field believe that computers are most beneficial to the experienced researcher. They believe that to really understand the data, beginning researchers need to "eyeball" and frequently reexamine the data rather than depend on what the computer provides.

Novice researchers sometimes make the mistake of assuming that the technology tool will actually conduct the analysis and provide for interpretation. That is not the case. Much as statisticians may want students to actually do computations by hand before using computers, qualitative researchers need to understand the qualitative analysis process before relying on technology. The researcher must also understand the differences in the various software programs available. For example, some may be appropriate to use in developing a grounded theory study but not for a different approach. Just as good writing is not guaranteed by using a good pen or a computer, neither is good qualitative analysis guaranteed by the use of a technological tool. It is the uniqueness of the human mind that enables the research to come to conceptual conclusions and interpret meaning from data. Use of complex analysis tools may create an illusion of analytic sophistication.

Table 17.2 indicates the capabilities of four of the more popular programs used in qualitative analysis.

Lesser known, but also used software programs include HyperQual 2, Ethnograph, CI-Said, Qualrus (supports text and multimedia sources and uses artificial intelligence for coding text), WordStat (for content analysis), Leximancer (focuses on key themes, concepts, and ideas in unstructured text), TextAnalyst (for semantic analysis), and Annotape (for recording and transcribing audio).

**Table 17.2** Qualitative Analysis Software

| | ATLAS.ti | QSR NVivo | Hyper-RESEARCH | MAXQDA |
|---|---|---|---|---|
| Platform | Windows PC | Windows PC | Windows or Mac | PC |
| Data capability | Text, graphic, audio, visual, codes, memos, findings | Text, graphic, audio,visual, codes, memos, findings | Text, graphics, audio, video, multimedia | Text, memos, images, video segments |
| Analysis | Search, retrieve, browse, code, annotate, compare, drag and drop | Analyze, manage, shape, code, compare, manipulate data | Code, retrieve, build theories, conduct analysis | Hierarchical code system, theory development, data/code/category system |
| Visual capacity | Build visual networks, concept maps | Graphic displays of codes and categories | Visual diagrams, can create a transfer of video and audio data | Visual diagrams |
| Import/ export data | Yes | Yes | Yes | Yes |
| Collaboration possible | Yes | Yes | Unknown | Yes |
| Other | Less memory needed compared to others | Combines features of N6 and NVivo2.0 Uses multiple languages | Easy to use Advanced multi media capabilities | Can use weighting to indicate relevance |
| Available | Scientific Software Development, Germany | QSR International, Australia | Research Ware, United States | VERBI Software, Germany |

**Table 17.3** Advantages and Disadvantages of Technology

| Advantages | Disadvantages |
|---|---|
| It can help store, manage, and organize data, including text, audio, video, and multimedia. | Researchers must first learn how to operate the program since they range in difficulty. |
| It can aid with searching and locating material easily (word, phrase, segment, etc.). | It may create a distance between the researcher and the data (researcher not as immersed). |
| It can provide visual representations, concept mapping, and diagrams that help with seeing relationships. | It may inhibit how the researcher thinks about the data and may stifle creativity. |
| It allows easier access to memos and insights from the researcher during the process. | The researcher must be careful to select the technology appropriate to the qualitative approach used (e.g., ethnography versus grounded theory). |
| It enhances ease of collaboration among researchers and facilitates data sharing. | It may make novices think that coding alone is analysis or that use of technology takes care of issues of rigor. |
| It makes new forms of data collection and analysis available as well as new ways to present research. | It may provoke ethical issues with storage of visual, audio, and multimedia data. |
| It allows analysis of larger volumes of data | It may expand data collection and coding beyond any real benefit. |

Because technology changes rapidly, it is not worthwhile to make a recommendation about a particular program here. We advise students to learn more about the capabilities of particular programs and to consider the time it may take to learn a program's functions when making decisions about use.

Various authors and researchers have discussed the advantages and disadvantages of using technology for collecting, analyzing, and storing qualitative data. Table 17.3 describes some of these.

## RIGOR IN QUALITATIVE RESEARCH

Dey (1993) lists six questions you should ask to check the quality of your data:

1. Are the data based on your own observation or hearsay?
2. Is there corroboration by others of your observations?
3. In what circumstances was an observation made or reported?
4. How reliable are the people providing the data?
5. What motivations might have influenced a participant's report?
6. What biases might have influenced how an observation was made or reported?

These questions seem to refer to whether the data are valid and reliable, or, in other words, is the quality of your data deemed rigorous? Although the terms *validity* and *reliability* have most commonly been associated with **rigor** in quantitative

**Table 17.4** Standards of Rigor for Research

| Quantitative | Qualitative | Issue Addressed |
|---|---|---|
| Internal validity | Credibility | Truth value |
| External validity | Transferability | Generalizability |
| Reliability | Dependability or trustworthiness | Consistency |
| Objectivity | Confirmability | Neutrality |

research, the concepts of making valid inferences from data and the consistency of the data are also important issues in qualitative research. Qualitative researchers may use different terms to apply to these criteria. Although the underlying concepts related to rigor are similar, there are differences in the nature of the data and the philosophical assumptions on which quantitative and qualitative research are based that have led to different terminologies. We discuss the qualitative terms most commonly used to address these issues of rigor and then strategies used in qualitative research to address concerns. Sometimes a single strategy may address more than one issue of rigor. Table 17.4 summarizes the standards of rigor used in quantitative and qualitative research and the issues of rigor addressed.

## CREDIBILITY

Validity cannot be taken for granted. It is relative—related to the purpose and circumstances of the research. The integrity of qualitative research depends on attending to the issue of validity. Validity concerns the accuracy or truthfulness of the findings. The term most frequently used by qualitative researchers to refer to this characteristic is **credibility.** How confident can you be in the researcher's observations, interpretations, and conclusions? Are they believable (credible)?

Credibility in qualitative research concerns the truthfulness of the inquiry's findings. Credibility or truth value involves how well the researcher has established confidence in the findings based on the research design, participants, and context. The researcher has an obligation to represent the realities of the research participants as accurately as possible and must provide assurances in the report that this obligation was met. The term *credibility* in qualitative research is analogous to *internal validity* in quantitative research.

A number of methods have been identified in the literature for enhancing the credibility (internal validity) of qualitative studies. These methods may be categorized according to five types of evidence: structural corroboration, consensus, referential or interpretive adequacy, theoretical adequacy, and control of bias.

### Evidence Based on Structural Corroboration

Eisner (1998) defines **structural corroboration** as a "means through which multiple types of data are related to each other to support or contradict the interpretation and evaluation of a state of affairs" (p.110). The use of multiple sources of data, multiple observers, and/or multiple methods is referred to as

**triangulation.** Structural corroboration uses different sources of data (data triangulation) and different methods (methods triangulation). A combination of data sources, such as interviews, observations, and relevant documents, and the use of different methods increase the likelihood that the phenomenon under study is being understood from various points of view. In **data triangulation,** the researcher investigates whether the data collected with one procedure or instrument confirm data collected using a different procedure or instrument. The researcher wants to find support for the observations and conclusions in more than one data source. Convergence of a major theme or pattern in the data from these various sources lends credibility to the findings. **Methods triangulation** uses more than one method (e.g., ethnography and document analysis) in the study. The assumption is that the combination of methods results in better evidence. When these different procedures or different data sources are in agreement, there is corroboration. For example, when interviews, related documents, and recollections of other participants produce the same description of an event or when a participant responds similarly to a question asked on three different occasions, one has evidence of credibility.

### Evidence Based on Consensus

Validity based on consensus is defined as "agreement among competent others that the description, interpretation, evaluation, and thematics" are correct (Eisner, 1998, p. 112). This type of validity is primarily demonstrated through two methods: peer review and investigator triangulation. In **peer review,** also called **peer debriefing,** the question is asked, "Given the evidence presented, is there consensus in the interpretation?" Colleagues or peers are provided with the raw data along with the researcher's interpretation or explanation. Discussions then determine whether the reviewer(s) considers the interpretation to be reasonable, given the evidence. **Investigator triangulation** involves having multiple researchers collect data independently and compare the collected data. The key questions here are "Did what was reported as taking place actually happen? Did the researcher accurately report what was seen and heard?"

If multiple investigators agree in their description of the context, in their description of events, and in their reporting of what was said, internal validity is enhanced. Using investigator triangulation makes it less likely that outside reviewers of the research will question the data. Several human instruments working together are usually better than one; furthermore, team members act as peer reviewers or "peer debriefers," keeping one another honest. A beginning researcher who must work alone on a dissertation can ask an outside person to be a debriefer; that is, the outsider can go through the researcher's data periodically and point out bias when necessary.

### Evidence Based on Referential or Interpretive Adequacy

**Referential** or **interpretive** evidence of validity refers to "accurately portraying the meaning attached by participants to what is being studied by the researcher" and "the degree to which the participants' viewpoints, thoughts, feelings, intentions, and experiences are accurately understood . . . and portrayed" (Johnson & Christensen, 2000, p. 209). Two primary strategies are used to enhance referential adequacy: member checks and low-inference descriptors.

**Member checks** (participant feedback) ask the question, "Do the people who were studied agree with what you have said about them?" At the end of the data collection period, the researcher may ask participants to review and critique field notes or tape recordings for accuracy and meaning. Or the researcher's sharing his or her interpretations of the data with the participants can help clear up miscommunication, identify inaccuracies, and help obtain additional useful data. In member checks, the researcher solicits feedback from the participants about the study's findings. Has the researcher accurately described and interpreted their experience? Feedback from the participants may help the researcher gain further insight and/or call attention to something that he or she missed. Furthermore, through member checking, the researcher demonstrates courtesy to the participants by letting them read what has been written about them.

Using many **low-inference descriptors** such as verbatim or direct quotations helps the reader experience the participants' world. Using tape recorders or video cameras enables the researcher to use these descriptors. Thick, rich description also helps the research convey an understanding of the study's context. These descriptions are very detailed, helping the reader "see" the setting, or if reporting themes from interviews, using the actual words of the respondents.

### Evidence Based on Theoretical Adequacy

**Theoretical adequacy** or **plausibility** concerns the degree to which a theoretical explanation developed from the study fits the data and is defensible. There are three key strategies for promoting theoretical adequacy: extended fieldwork, theory triangulation, and pattern matching.

**Extended fieldwork** provides time for the researcher to observe a full range of activities in order to identify patterns and relationships and their typicality, which will contribute to valid interpretations. An extended time in the research setting enables the researcher to gain the participants' trust and thus to obtain more detailed and honest responses. Credibility is threatened by errors in data collection caused by participants providing socially acceptable responses or acting differently because of the researcher's presence. Use of extended fieldwork can reduce this possibility because it provides the researcher more opportunities to check perspectives and allows informants to become accustomed to the researcher's presence.

**Theory triangulation** involves consideration of how the phenomenon under study might be explained by multiple theories. Considering different theories, the researcher may gain better insights. **Interdisciplinary triangulation** uses other disciplines to infer processes and understandings of the findings.

**Pattern matching** involves making predictions based on theory and testing those predictions. Based on theory, the researcher predicts a certain pattern of results. Then the researcher determines the degree to which the patterns found in the data match the predicted pattern. If the predictions occur, the researcher has strong evidence to support his or her explanations.

### Evidence Based on Control of Bias

**Researcher bias** is a source of invalidity in qualitative studies. Bias may result from selective observations, hearing only what one wants to hear, or allowing personal attitudes, preferences, and feelings to affect interpretation of data.

**Table 17.5** Approaches to Enhancing Credibility in Qualitative Studies

| Criterion | Strategies |
|---|---|
| Structural corroboration | Methods triangulation<br>Data triangulation |
| Consensus | Peer review/peer debriefing<br>Investigator triangulation |
| Referential or interpretive adequacy | Member checks/participant feedback<br>Low-inference descriptors/thick, rich description |
| Theoretical adequacy | Extended fieldwork<br>Theory triangulation<br>Interdisciplinary triangulation<br>Pattern matching |
| Control of bias | Reflexivity<br>Negative case sampling |

The most common strategy to control for bias in qualitative studies is reflexivity. **Reflexivity** is the use of self-reflection to recognize one's own biases and to actively seek them out. The researcher should refer to his or her journal reflections during the process of data analysis.

Another strategy used to control for bias is **negative case sampling,** in which researchers intentionally seek examples of the opposite of what they expect. To avoid the appearance of bias, researchers should show that they have searched for and explained any discrepant or contradictory data. Table 17.5 summarizes the strategies used for enhancing credibility in qualitative studies.

## TRANSFERABILITY

**Transferability** is the degree to which the findings of a qualitative study can be applied or generalized to other contexts or to other groups. In quantitative research, the term *external validity* is used to refer to the generalizability of the findings. Although the qualitative researcher typically does not have generalizability as a goal, it is his or her responsibility to provide sufficiently rich, detailed, thick descriptions of the context so that potential users can make the necessary comparisons and judgments about **similarity** and hence transferability. This is referred to as **descriptive adequacy.** The researcher must strive to provide accurate, detailed, and complete descriptions of the context and participants to assist the reader in determining transferability.

Qualitative inquirers argue that it is possible to apply qualitative findings to other people, settings, and times to the extent that they are similar to the people, settings, and times in the original study. Transferability of a set of findings to another context depends on the similarity or "goodness of fit" between the context of the study and other contexts. The transfer is made by the potential user of the findings, who must compare and decide on the similarity of the two contexts. This contrasts with quantitative research, in which the original researcher makes generalizations. One strategy to enhance transferability is to include **cross-case comparisons.** The researcher may investigate more than one case. If findings are similar, this would increase the possibility of transferability of findings to other

**Table 17.6** Approaches to Enhancing Transferability in Qualitative Studies

| Criterion | Strategies |
|---|---|
| Descriptive adequacy | Thick, rich description |
| Similarity | Cross-case comparisons<br>Literature comparisons<br>Describing limitations |
| Limiting reactivity | Reflective statement<br>Detailed description of methods |

settings or contexts. In some cases, even a single case can be compared with other cases in the published literature that might demonstrate transferability.

Be aware that there are threats to transferability, such as **selection effects** (the fact that the constructs being investigated are unique to a single group), **setting effects** (the fact that results may be a function of the specific context under investigation), and **history effects** (the fact that unique historical experiences of the participants may militate against comparisons). The researcher should recognize limitations of the study in the description. Detailing of circumstances helps the reader to understand the nature of the data and what might be peculiar to your particular study.

**Reactivity** (the effect of the research itself) might also limit transferability. Although eliminating the influence of the researcher may be impossible in a qualitative study because the researcher is the key data collection instrument, the researcher can help the reader understand the potential influence by describing his or her own biases through a reflective statement and providing detailed descriptions of such things as observation strategies and interview questions. Reactivity is a more serious threat in studies using interview techniques. Table 17.6 summarizes the strategies used to enhance transferability.

## DEPENDABILITY

Qualitative researchers speak of **dependability** rather than reliability. Recall that reliability in quantitative research has to do with consistency of behavior, or the extent to which data and findings would be similar if the study were replicated. However, unlike quantitative research, in which tight controls enhance replicability, qualitative studies expect variability because the context of studies changes. Thus, consistency is viewed as the extent to which variation can be tracked or explained. This is referred to as *dependability* or **trustworthiness.** Some strategies to investigate dependability are using an audit trail, replication logic, stepwise replication, code–recoding, interrater comparisons, and triangulation. To enhance reliability, the researcher wants to demonstrate that the methods used are reproducible and consistent, that the approach and procedures used were appropriate for the context and can be documented, and that external evidence can be used to test conclusions.

### Documentation

One of the best ways to establish dependability is to use an **audit trail.** Audit trails provide a mechanism by which others can determine how decisions were made

and the uniqueness of the situation. It documents how the study was conducted, including what was done, when, and why. The audit trail contains the raw data gathered in interviews and observations, records of the inquirer's decisions about whom to interview or what to observe and why, files documenting how working hypotheses were developed from the raw data and subsequently refined and tested, the findings of the study, and so forth. The researcher must keep thorough notes and records of activities and should keep data well organized and in a retrievable form. He or she should provide information on the sample of people studied, the selection process, contextual descriptions, methods of data collection, detailed field notes, tape recordings, videotapes, and other descriptive material that can be reviewed by other people. Using the audit trail as a guide, an independent third-party auditor can examine the inquirer's study in order to attest to the dependability of procedures employed and to examine whether findings are confirmable—that is, whether they are logically derived from and grounded in the data that were collected. A complete presentation of procedures and results enables the reader to make a judgment about the replicability of the research within the limits of the natural context.

### Consistent Findings

Dependability can be demonstrated by showing consistent findings across multiple settings or multiple investigators. **Replication logic,** which involves conducting the study in multiple locations or with multiple groups, is suggested for determining dependability of a study. According to this logic, the more times a finding is found true with different sets of people or in different settings and time periods, the more confident the researcher can be in the conclusions. **Stepwise replication** is another technique suggested for enhancing dependability. In this strategy, two investigators divide the data, analyze it independently, and then compare results. Consistency of results provides evidence of dependability.

### Coding Agreement

**Intrarater** and **interrater agreement** are strategies for assessing dependability (reliability). An intrarater method is the **code–recode strategy:** A researcher codes the data, leaves the analysis for a period of time, and then comes back and recodes the data and compares the two sets of coded materials. Because much qualitative research involves observation by multiple observers, some researchers suggest interrater or interobserver agreement methods for assessing dependability. For example, a researcher might randomly select a transcript and ask a peer to code the transcript using the coding labels identified by the researcher. The second coder would be free to add other codes he or she might identify. After the peer completes coding of the transcripts, the results are compared to the original coded transcript to determine whether both coders labeled components of the transcript the same.

### Corroboration

Triangulation, which we have previously discussed, is also used to establish the dependability of qualitative studies. If multiple data sources or multiple methods result in similar findings, it enhances the reliability of the study. Table 17.7 summarizes strategies to enhance dependability in qualitative studies.

**Table 17.7** Approaches to Enhancing Dependability in Qualitative Studies

| Criterion | Strategies |
|---|---|
| Documentation | Audit trail |
| Consistent findings | Replication logic<br>Stepwise replication |
| Coding agreement | Code–recode/intrarater agreement<br>Interrater/interobserver agreement |
| Corroboration | Data triangulation<br>Methods triangulation |

## CONFIRMABILITY

**Confirmability** in qualitative research is the same as the quantitative researcher's concept of objectivity. Both deal with the idea of **neutrality** or the extent to which the research is free of bias in the procedures and the interpretation of results. Because it may be impossible to achieve the levels of objectivity that quantitative studies strive for, qualitative researchers are concerned with whether the data they collect and the conclusions they draw would be confirmed by others investigating the same situation. Thus, in qualitative studies, the focus shifts from the neutrality of the researcher to the confirmability of the data and interpretations.

The audit trail is the main strategy for demonstrating confirmability. By providing a complete audit trail, the researcher enables another researcher to arrive or not arrive at the same conclusions given the same data and context. Other strategies used to enhance confirmability include triangulation of methods, peer review, and reflexivity—all discussed previously. Table 17.8 summarizes the strategies used to enhance confirmability.

In summary, support for validity and reliability of qualitative studies requires well-documented research and rich description. As you consider rigor in qualitative research, Daniel and Onwuegbuzie (2002) suggest that you ask these questions:

1. Have you shown the cohesiveness of the evidence? Are related pieces of data used to form a narrative argument and examined collectively to form evidence of cohesiveness?
2. Have data inconsistencies been examined? Have you looked consciously for negative cases and inconsistencies in findings?
3. Have you considered alternate explanations? Are there contradictions in the broader patterns within the data in which data from one source do not line

**Table 17.8** Approaches to Enhancing Confirmability in Qualitative Research

| Criterion | Strategies |
|---|---|
| Documentation | Audit trail |
| Corroboration | Triangulation<br>Peer review |
| Control of bias | Reflexivity |

up with data from another? Might such contradictions indicate systematic misunderstandings of the data or suggest a need to develop a new theory?

4. Do you have confidence (surety) in the results? Have your data been recorded accurately and have codes been applied uniformly and consistently?
5. Have you achieved the elusive goal of data collection? In positivist research this refers to obtaining an accurate understanding of the true score. In qualitative research, the question is whether you have accurately captured the social understanding or social reality underlying events, activities, and behaviors.
6. Is there adequacy of evidence? Similar to the positivist concept of domain sampling, what is the degree to which narrative descriptions provide an adequate view of the social phenomena of interest. Is there adequate thickness of description?

### THINK ABOUT IT 17.4

For each strategy listed, indicate which concern related to rigor it will help address. Note that some strategies address more than one concern.

1. Audit trail
2. Cross-case comparisons
3. Data triangulation
4. Extended fieldwork
5. Interrater agreement
6. Investigator triangulation
7. Member checks
8. Methods triangulation
9. Negative case sampling
10. Peer debriefing
11. Reactivity
12. Thick, rich description

**Answers**

Answers are contained in the tables in this chapter.

## EVALUATING QUALITATIVE REPORTS

The following are some of the general criteria used to evaluate qualitative reports:

1. Is the research question stated, and does the researcher make clear the conceptual and/or theoretical framework for the study?
2. Does the researcher show the relationship between the study and what is known in the literature?
3. Does the researcher indicate how and why the site and/or participants were selected for the study? Does the researcher explain the extent to which participants are representative?
4. Are the data collection methods explained so that the reader can judge if they were adequate and appropriate to the question? Does the researcher

explain his or her role as participant observer or nonparticipant observer, interviewer, and so on?

5. Does the researcher explain the data analysis procedures used?
6. Are the strategies used to enhance the credibility, transferability, dependability, and confirmability of the data (i.e., triangulation, audit trail, etc.) described?
7. Are the descriptive data separate from the interpretations? Is there abundant raw data presented (quotes, etc.) to demonstrate the findings?
8. Is there evidence that the researcher maintained ethical standards? Are personal biases and assumptions expressed? Have steps been taken to guard against value judgments in data collection and analysis?
9. Does the study answer the research question and suggest further questions for investigation?
10. Does the researcher make explicit the significance of the study?
11. Does the report qualify any generalizations that were made, and does it help the reader understand how what was learned might be transferred to another similar situation?
12. Is the study reported in a way that is accessible to others?

The reader also needs to recognize that in choosing to use certain qualitative approaches, more targeted questions may need to be considered in evaluating quality. Creswell (2007) provides a chapter on some of the differences in evaluating specific types of qualitative research. We have summarized some of these suggestions on distinguishing good qualitative research in Table 17.9. Reviewing these questions as well as the previous ones as you write your own report will help you do a better job.

In 2004, the National Science Foundation produced a report, *Workshop on Scientific Foundations of Qualitative Research* (Ragin, Nagel, & White, 2004), that investigated the characteristics of strong qualitative research and provided recommendations on how to strengthen qualitative methods. Their recommendations (on page 17 of the report) are worth noting here as we conclude this chapter. They indicate that these recommendations should be used both to improve the quality of qualitative proposals and reports and to evaluate the quality of the research conducted. In strong qualitative research, the researcher should:

- Write clearly and engagingly for a broad audience.
- Situate the research in relationship to existing theory.
- Locate the research in the literature specifying comparable cases and building on others' findings.
- Articulate the potential theoretical contribution by indicating what gaps in theory might be filled.
- Clearly outline the research procedures including details about where, when, who, what, and how the research will be conducted.
- Provide evidence of the project's feasibility, including permission to access the site and human subjects approval.

**Table 17.9** Evaluation Criteria for Specific Approaches to Qualitative Research

| Narrative | Phenomenology | Grounded Theory | Ethnographic | Case Study |
|---|---|---|---|---|
| Does the researcher focus on an individual and collect stories about a significant issue in his or her life? | Does the researcher convey an understanding of the philosophical basis of phenomenology? | Did the researcher study a process, action, or interaction as a key element in the theory? | Did the researcher clearly identify a culture sharing group and cultural themes for examination? | Did the researcher clearly identify a case that exhibited intrinsic merit in understanding the research issue? |
| Does the researcher develop a chronology that connects aspects of the story and re-story the story of the participant? | Does the researcher concisely articulate the phenomenon under study? | Did the researcher use a coding process that moved from the data to the theory? | Did the researcher provide a detailed description of the cultural group and the themes that derived from understanding that group? | Did the researcher provide a clear and detailed description of the case? |
| Did the researcher tell a persuasive story in a literary fashion? | Did the researcher use data analysis procedures recommended for phenomenology? | Did the researcher present the theoretical model in a figure or diagram? | Did the researcher identify issues that arose in the field (e.g., relationships and sensitivity, etc.)? | Did the researcher identify themes that are related to the case? |
| Did the researcher build themes from the story and conduct a broader analysis? | Did the researcher convey the essence of the experience and include both a description of the experience and the context? | Did the researcher provide propositions to connect the categories in the theoretical model and present questions for further study? | Did the researcher provide an overall explanation of how the group works? | Did the researcher appropriately make assertions or generalizations from the case analysis? |
| Did the researcher reflexively bring him- or herself into the study? | Was the researcher reflexive throughout the study? | Did the researcher self-disclose his or her own stance in the study? | Did the researcher self-disclose his or her position in the research? | Did the researcher self-disclose his or her position in the study? |

- Provide a description of the data to be collected including kinds of evidence to be gathered, different modes of data collection, and different places data will be obtained.
- Explain the plan for data analysis, including management of the data and procedures for making sense of the information obtained.

- Describe the strategy to refine concepts and construct theory as the investigation continues.
- Include plans to search for and interpret disconfirming evidence, alternative explanations, and unexpected findings.
- Provide an assessment of the possible impact of the researcher's presence and biography on the research from the point of problem selection through data analysis to address potential bias of results.
- Provide information about replicability and suggest ways others might reproduce the research.
- Describe the data archive (audit trail) that will be left for others to use and how you will maintain confidentiality.

**THINK ABOUT IT 17.5**

Using your library or through online services, locate and read the article listed here. Use the general criteria as well as the criteria for case studies and make an analysis of the quality of this study.

Tracy Smith and David Strahan (2004). Toward a prototype of expertise in teaching: A descriptive case study. *Journal of Teacher Education, 55*(4), 357–372.

## SUMMARY

Qualitative studies typically result in a massive amount of data. Analyzing the data is thus a tedious and time-consuming process. In general, the first step in analysis is to organize the data by searching for words, phrases, and events that appear regularly and putting those with similar units of meaning into a category. This continues until one has a set of categories that provide a reasonable reconstruction of the data collected. Next, the researcher compares and tries to find connections and common themes among the categories. Interpretation is a highly personal activity in which the researcher explains the meaning of the data and why it is important. Technologies are being used extensively in qualitative study to collect, organize, analyze, interpret, and report.

The exact format of the report depends on the nature of the study (whether it is a dissertation, journal article, or perhaps a presentation) and on the approach to qualitative research used (phenomenology, narrative, or case study). Validity and reliability are important concepts in qualitative as well as quantitative research. Qualitative researchers use the terms credibility, transferability, dependability, and confirmability instead of internal validity, external validity, reliability, and objectivity, respectively. A variety of methods are used to support the validity and reliability of qualitative data and their interpretation. These include peer review, member checks, triangulation, maintaining an audit trail, seeking negative case evidence, and keeping a research journal.

Qualitative reports are written differently in style than quantitative reports, but many of the components covered are similar. Both pose research questions, connect the research to the literature, describe methods of collecting and analyzing data, present findings, and discuss the implications of the study. Many of the criteria for evaluating the quality of qualitative research are also appropriate in quantitative research.

## KEY CONCEPTS

audit trail
axial coding
category
code–recode strategy
coding
confirmability
consensus
constant comparative method
credibility
cross-case comparison
data analysis
data analysis spiral
data pyramid
data triangulation
dependability
descriptive adequacy
discrepant data analysis
extended fieldwork
familiarization
history effects
inductive coding
interdisciplinary triangulation
interpretation
interpretive adequacy
interrater agreement
intrarater agreement
investigator triangulation
in vivo codes
lean coding
low-inference descriptors
member checks
methods triangulation
negative case analysis
negative case sampling
neutrality
open coding
organization
organizational categories
pattern matching
peer debriefing
peer review
plausibility
preliminary coding
provisional coding
reactivity
reducing
referential adequacy
reflective log
reflexivity
replication logic
representation
researcher bias
rigor
selection effects
setting effects
similarity
stepwise replication
structural corroboration
substantive categories
themes
theoretical adequacy
theoretical categories
theory triangulation
thick description
transcriptions
transferability
triangulation
trustworthiness

## EXERCISES

1. Explain how data analysis in qualitative and quantitative studies differs.
2. Briefly describe the constant comparative method for analyzing qualitative data.
3. One of the criticisms of qualitative research is that it is subject to bias. Do you agree? What does the qualitative researcher do to rebut this kind of criticism?
4. Would you agree with a friend who believes a qualitative study would be easier for a dissertation than a quantitative study? What are some differences you might point out?
5. How does a qualitative report differ from a report of quantitative research?
6. What is the first step the researcher should take in analyzing qualitative data?
7. What are the three levels of qualitative data analysis?
8. What are the different types of triangulation that can be used to enhance the rigor of a qualitative study?
9. What terms are used in qualitative inquiry to refer to issues of internal and external validity?
10. Name and describe at least two strategies that can be used to enhance the dependability (reliability) of a qualitative study?
11. Describe one strategy for controlling researcher bias in a qualitative study?
12. What is the difference between a member check and a peer debriefing?
13. What evaluation criteria might vary depending on the type of qualitative study conducted?
14. Explain the ways in which technology might be used in qualitative research and some of the advantages and disadvantages of using technology.

## ANSWERS

1. Data analysis in qualitative research is inductive; it is a process of categorizing data that are in the form of words, synthesis, and finding meaning. Data analysis in quantitative research is deductive; it involves testing hypotheses using numerical data and statistical tests.
2. It is a process in which the researcher takes each new unit of meaning and compares it to other units of meaning (categories) and then groups it with categories that are similar. It is a way to reduce the data and make it more manageable and conducive to interpretation.
3. A number of procedures are used to contribute to a credible study. Among them are triangulation, audit trails, member checks, and working with a team (using more than one human instrument is better than one).
4. Answers will vary. In some ways, qualitative research may be more difficult than quantitative research: The design is much less prescriptive and structured, data collection is time-consuming, and data analysis is more tedious and involved. Writing a qualitative report is likely to take much more time than writing a quantitative report. One aspect that may make qualitative research easier for some is the lack of statistics.
5. A qualitative report is narrative in form and rich with descriptions, quotations, and so forth. One can use the first person, which is never used in quantitative reporting.
6. The researcher should begin by familiarizing him- or herself with the data by reading and rereading transcripts, listening to audiotapes, or watching videos.
7. Data pieces, categories (minor), and themes (major)
8. Data triangulation, methods triangulation, and investigator triangulation
9. Credibility and transferability
10. Answers may vary. Examples include audit trail, replication logic, and intra- and inter-rater agreement.
11. Answers may include reflexivity or negative case sampling.
12. In a member check, the researcher goes back to the participants in the study and asks them to review the findings and conclusions in order to provide a check on whether the researcher "got the story right." In a peer debriefing, the researcher provides the raw data along with interpretations to a second researcher (peer) in order to check whether others would make the same interpretations given the data.
13. Answers can be found in Table 17.9.
14. Technology can be used for collecting data as well as for data storage and management, searching, coding, developing themes and visual representations, reporting, and collaborating. Advantages and disadvantages are provided in Table 17.3.

## REFERENCES


Corbin, J., & Strauss, A. (2008). *Basics of qualitative research* (3rd ed.). Thousand Oaks, CA: Sage.

Creswell, J. W. (2007). *Qualitative inquiry & research design* (2nd ed.). Thousand Oaks, CA: Sage.

Daniel, L. G., & Onwuegbuzie, A. J. (2002). *Reliability and quantitative data: Are psychometric concepts relevant within an interpretive research program?* Paper presented at Mid-West Education Research Association, Chattanooga, TN. ERIC Document Reproduction No. ED471306.

Dey, I. (1993). *Qualitative data analysis*. New York: Routledge.

Eisner, E. W. (1998). *The enlightened eye: Qualitative inquiry and the enhancement of educational practice*. Upper Saddle River, NJ: Merrill-Prentice Hall.

Johnson, B., & Christensen, L. (2000). *Educational research: Quantitative and qualitative approaches*. Boston: Allyn & Bacon.

LaViolette, S. A. (2001). *School-community conflict: Perceptions of elementary school principals and parents in an urban setting*. Doctoral dissertation, Northern Illinois University, DeKalb.

Lee, R. M., & Fielding, N. G. (1995). User's experiences with qualitative data analysis software. In U. Kelle (Ed.), *Computer-aided qualitative data analysis: Theory, methods and practice* (pp. 29–40). London: Sage.

Marshall, C., & Rossman, G. B. (2006). *Designing qualitative research* (4th ed.). Thousand Oaks, CA: Sage.

Maxwell, J. A. (2005). *Qualitative research design: An interactive approach* (2nd ed.). Thousand Oaks, CA: Sage.

Merriam, S. B., and Associates. (2002). *Qualitative research in practice: Examples for discussion and analysis*. San Francisco: Jossey-Bass.

Ragin, C., Nagel, J., & White, P. (2004). *Workshop on scientific foundations of qualitative research*. Washington, DC: National Science Foundation.

Wiersma, W. (2000). *Research methods in education: An introduction* (7th ed.). Boston: Allyn & Bacon.

Wolcott, H. F. (1994). *Transforming qualitative data: Description, analysis, and interpretation*. Thousand Oaks, CA: Sage.

CHAPTER 18

# Action Research

Action research is a powerful tool for studying and improving one's practice.

Hendricks (2009)

## INSTRUCTIONAL OBJECTIVES

After studying this chapter, the student will be able to:

1. Define action research and its underlying assumptions.
2. Give examples of areas in which action research can be used in schools.
3. Compare and contrast action research with other types of research.
4. Explain the action research process.
5. Define a problem that could be examined through action research and identify appropriate data sources to be used in investigating the problem.
6. Give examples of approaches to data analysis and interpretation used in action research.
7. Describe what should be included in an action plan.
8. Explain why action research is useful to the field of education.
9. Apply criteria for assessing the quality of an action research report.

The two words themselves, *action* and *research*, indicate the two core components of this approach. **Action research** is about taking action based on research and researching the action taken. Action research has been used in a variety of settings, including schools, hospitals, health clinics, community agencies, government units, and other environments. It can be used to enhance everyday work practices, to resolve specific problems, and to develop special projects and programs. Action research is based on the premise that local conditions vary widely and that the solutions to many problems cannot be found in generalized truths that take no account of local conditions. In this chapter, we focus on the use of action research in education.

**PICTURE THIS**


Joe Rocco

## DEFINING ACTION RESEARCH

In education, action research can be applied to such areas as curriculum development, teaching strategies, and school reform. Action research in schools is also called **practitioner research, teacher inquiry,** or **teacher research,** although the process can certainly be used by other school personnel besides teachers, including coaches, counselors, principals, superintendents, librarians, technology specialists, and other education professionals. The goal of action research in education is to create an inquiry stance toward teaching where questioning one's own practice becomes part of the work and of the teaching culture.

Good teachers have always engaged in a form of action research, although they may not have called it that. Good teachers engage in *reflection*, a key component of action research. But action research is more than reflection. It emphasizes a systematic research approach that is cyclical in nature, alternating between action and reflection, continuously refining methods and interpretations based on understandings developed in earlier cycles.

Mertler (2009) lists a set of characteristics to help us determine what is and what is not action research. Table 18.1 summarizes these and other characteristics from the literature.

**Table 18.1** Characteristics of Action Research

| Action Research *Is* | Action Research *Is Not* |
|---|---|
| A process to improve education by incorporating change and involves educators working together to improve their own practice | The usual thing that teachers do when thinking about teaching |
| Persuasive and authoritative because it is done by and for educators | Acceptance of solutions posed by the experts |
| Collaborative and encourages educators working and talking together in empowering relationships, including educators as integral, participating members of the process | Done to or by other people outside of the setting |
| Practical and relevant and allows educators direct access to research findings | Theoretical, complicated, or elaborate |
| A way to develop critical reflection and open-mindedness | A way to provide conclusive evidence |
| A planned, systematic, and cyclical approach to understanding the learning process and to analyzing educational places of work | Relying on tradition, gut feelings, and common sense |
| A process that requires testing of our ideas about education | The implementation of predetermined answers to educational questions |
| A justification on one's teaching practices | A fad |

There are three main characteristics of action research:

1. The research is situated in a local context and focused on a local issue.
2. The research is conducted by and for the practitioner.
3. The research results in an action or a change implemented by the practitioner in the context.

Here is an example of action research by a teacher: Mr. Rodriguez teaches middle school social studies in an alternative school attended by students who have been unsuccessful in the regular school. He has noticed in his classes that the students seem bored and unmotivated and that their grades in social studies are generally poor. He believes that social studies will help students become more productive citizens and is concerned that they are not learning. Mr. Rodriguez decides to study the problem to determine whether he can find ways to better engage the students in the learning process and thus improve their learning. (*Characteristic 1*: Mr. Rodriguez has engaged in a process of reflection and has identified a problem in his own context of practice, his classroom.)

Mr. Rodriguez decides to keep a journal for 2 weeks to note his observations about student behavior. He also develops an interview protocol and conducts a series of interviews with the students, asking them to talk about times when they liked learning and times when they did not. He asks them about their school experiences, particularly in social studies. By doing these things, Mr. Rodriguez has engaged in a qualitative research process. (*Characteristic 2*: He is conducting his own research for his own purposes, to improve his teaching.)

Mr. Rodriguez reviews his findings in light of the literature and concludes that some specific changes in his teaching strategies, such as using more project-based group work, might make a difference. He implements those teaching strategies and gathers additional evidence, including a review of student performance, to determine whether the strategy worked. He continues his journaling to determine

whether he observes differences in student behavior. Based on these findings, he will continue to make decisions about his practice. (*Characteristic 3*: He implements changes in his classroom based on his findings.)

Mr. Rodriguez is doing action research. There is more about Mr. Rodriguez's research later in this chapter.

## APPROACHES TO ACTION RESEARCH

Although action research has gained considerable attention in the field of education, the concepts are not new. Kurt Lewin, considered the father of action research, is credited with coining the term in the 1940s, primarily associated with social change efforts. In education, some trace the conceptual roots to the progressive views of John Dewey. Today, action research has gained popularity in the United States and elsewhere and is seen as important in the work of improving schools. In this text, we focus on the practical application of action research in schools rather than the philosophical roots of the approach. Mr. Rodriguez's study at the alternative school is an example of **practical action research.**

Within action research, there are four approaches that differ in their purposes and goals. The methods used for data gathering, analyzing, and interpreting are essentially the same, but the reader may consider the various types to determine which aligns best with his or her goals. Hendricks (2009, pp. 9–10) summarizes these four approaches to action research. Table 18.2 is based on her descriptions.

## BENEFITS OF ACTION RESEARCH IN EDUCATION

The following are some of the many benefits of encouraging action research in educational settings:

- Professionalizing the work of educators and promoting professional development

**Table 18.2** Approaches to Action Research

| Approach | Who's Involved | Purpose/Goal |
|---|---|---|
| **Collaborative action research** | Involves multiple researchers. In education, this may include school and university personnel or teachers and school administrators. | To share expertise and foster dialogue among stakeholders |
| **Critical action research** | Involves wide collaboration. In education, this may include university researchers, school administrators, teachers, and community members. | To evaluate social issues and use the results for social change |
| **Classroom action research** | Involves teachers in their classrooms; can involve groups of teachers examining common issues. | To improve classroom practice or to improve practices in the school |
| **Participatory action research** | Involves collaboration among stakeholders in a social process. | To explore practices within social structures (emancipatory); to challenge power differences and unproductive ways of working (critical); and to change theory and practice (transformational) |

*Source*: From Cher Hendricks. *Improving Schools Through Action Research: A Comprehensive Guide for Educators,* 2nd edition. Published by Allyn and Bacon/Merrill Education, Boston, MA. Copyright © 2008 by Pearson Education. Reprinted by permission of the publisher.

- Empowering teachers and giving educators a voice in the field
- Developing knowledge directly related to practice and focusing on improving practice
- Promoting reflection and use of information for better decision making
- Fostering an openness to new ideas and encouraging creativity
- Encouraging collaboration and the development of learning communities
- Encouraging rethinking about how teachers' and students' work is evaluated
- Providing rich sources of data that can be used for school improvement
- Revitalizing professional lives, making work interesting and rewarding
- Allowing articulation of choices made and methods used
- Increasing understanding and respect among teachers, students, parents, and administrators

A key benefit of action research is reducing the gap between theory and practice. There are several reasons for this acknowledged gap. Research may be written in ways that make it inaccessible to practitioners (i.e., use of jargon or focus not relevant to practice) or difficult to translate to local practice. Ways that research findings are provided to practitioners may not be sensitive to the time demands of practice. Johnson (2008, p. 33) also references the **"Moses effect"** as a contributor to the gap when "researchers hand down research edicts from on high with the expectation that teachers will be passive receivers of these edicts" without valuing teachers' views or recognizing their concerns.

## ACTION RESEARCH COMPARED TO TRADITIONAL RESEARCH

Some have suggested that action research is a new genre of research, different from the quantitative and qualitative approaches with which we are already familiar. Action research may have different purposes, different incentives, and different audiences compared to other forms of research, but it uses the same methodologies we have already discussed. Action research starts, as does all research, with a problem to be solved. However, the problem may not be well defined and is often referred to as a "focus" rather than a problem.

The key difference is in the primary purpose of action research, which is to take action to solve a local problem or to improve a practice. It is not intended to create theories or to be generalizable. Action research is a practical tool for solving problems experienced by people in their professional lives. It is empirical insofar as it requires people to define and observe the phenomena under investigation.

Although action research may use many of the methods, procedures, and concepts associated with quantitative research, it is usually more closely aligned with qualitative methods. While action research involves systematic inquiry, it has some distinctive characteristics that bear consideration. Table 18.3 compares typical action research with the more formal traditional approaches to research you have studied in this textbook.

**Table 18.3** Comparison of Quantitative Research and Action Research

| Quantitative Research | Action Research |
|---|---|
| **Purpose/Goals** | |
| Produces objective knowledge that can be generalized to larger populations in order to predict future events based on a preexisting set of conditions. | Interprets events and enables individuals or groups of people to formulate acceptable solutions to local problems. |
| Tests theories in order to explain the nature of the world or the nature of reality. | Seeks viable, sustainable, effective solutions to common problems. |
| Interprets phenomena to add to the knowledge base. | Interprets phenomena to adapt practice. |
| Investigates larger issues with intent to generalize. | Identifies and addresses local problems with little regard for generalizability. |
| **Approach** | |
| Develops a study based on what is known in the professional literature. | Develops a study based on experience and valuing practitioner perspectives. |
| Uses extensive literature review and relies on primary sources. | Uses cursory literature review and relies mostly on secondary sources. |
| Objectivity and value neutrality or control of bias are the ideals. | Subjective or authoritative judgments of individuals are respected. |
| Uses more rigorous procedures and typically has a longer time frame. | Uses less rigorous, looser procedures and more typically has a shorter time frame. |
| Studies smaller units of a phenomenon. | Complexity is embraced. |
| **Methods** | |
| Uses quantitative approaches to measure and predict variables, and tests of statistical significance to study cause-and-effect relationships. | Uses data from a variety of sources to help understand context or the effect of interventions. |
| Follows carefully prescribed procedures. | Does not necessarily follow carefully prescribed procedures. |
| Focuses on conventional research rules and concepts of measurement. | Focuses on the experienced reality of the day-to-day. |
| Experimental design is considered the "gold standard" for quantitative research. | Asserts that assigning a student to a control group when the researcher believes that the treatment is superior is to deny students the best possible instruction. |
| Uses random sampling when possible in quantitative studies. | Chooses participants based on the intentions of the study. |
| Uses primarily professionally developed instruments. | Uses primarily teacher-developed or convenient instruments. |
| Data analysis relies on statistical techniques and complex qualitative analysis. | Data analysis focuses on practical significance and reports raw data. |
| Emphasizes theoretical significance and increased general knowledge. | Emphasizes practical significance and increased knowledge about a particular context. |
| **Role of Researcher** | |
| Research is often conducted by outsiders—scholars, researchers, and professors. | Research is conducted by insiders—the practitioner or teacher. |
| There is a clear distinction between researcher and subjects. | There is little distinction between the researcher and the subjects, and participants may engage in the process. |
| Formal training is required to conduct studies. | Little formal training is required to conduct studies. |
| The researcher reports findings to professional audiences. | The researcher shares informal findings with peers or sometimes with professional audiences. |

**THINK ABOUT IT 18.1**

For each of the following statements, indicate "A" for action research or "Q" for quantitative research:

1. Occurs within a natural environment
2. Values collaborative construction of interpretation
3. Tests theories
4. Values stakeholders' and practitioners' perspectives
5. Randomly assigns subjects to experimental and control groups
6. Seeks solutions to local problems
7. Seeks objective knowledge that can be generalized
8. Maintains a clear distinction between researchers and subjects
9. Uses purposeful samples
10. Emphasizes the use of professionally developed instruments

**Answers**

1. A; 2. A; 3. Q; 4. A; 5. Q; 6. A; 7. Q; 8. Q; 9. A; 10. Q

## THE ACTION RESEARCH PROCESS

The action research process involves reflecting, planning, acting, and observing. As shown in Figure 18.1, Stringer (2008) refers to the process in three steps: look, think, act. The "look" phase involves systematically gathering information and data. In the "think" phase, information is analyzed and reflected upon. In the "act" phase, solutions are devised and implemented.

Although the action research models described in the literature differ in some ways, they appear to have common elements, as can be seen in the work of Stringer (2008), Johnson (2008) and Mertler (2009). We describe the process more fully:

1. *Reflect*. Experience and perceptions are used to identify an area of focus based on a problem. Time is taken to review what is already known about the problem or focus area and to learn more about the problem. This may include a review of the literature. This phase is also referred to as *identifying and limiting the topic*.

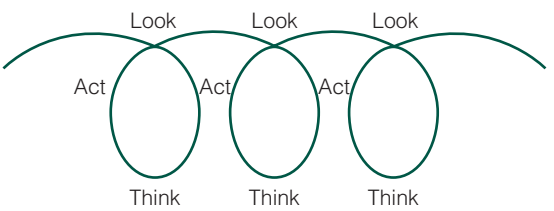



**Figure 18.1** Action Research Process

Source: From Ernie Stringer, *Action Research in Education,* second edition. Published by Allyn and Bacon/Merrill Education, Boston, MA. Copyright © 2008 by Pearson Education. Reprinted by permission of the publisher.

2. *Plan*. A plan is developed for taking action and/or for gathering information and data in order to observe or capture the experience or monitor the practice. It is in this phase that the research questions and methods are explicated.
3. *Act*. The researcher implements the plan or changes a practice and collects data. Data may be collected from a variety of sources.
4. *Observe*. The researcher synthesizes and analyzes the data. Key issues related to the problem are identified. This leads to reflection once more.
5. *Reflect*. The researcher reflects on and interprets the information and communicates or reports it to others. A new understanding of the nature of the problem is developed. Actions are taken and a new area of focus is identified.
6. *Plan*. A new plan of action is developed to resolve or investigate the problem.
7. *Act*. A new action is taken and data are collected (perhaps the same types of data or perhaps something different).
8. *Observe*. The new data are analyzed, synthesized, and interpreted. The researcher then spirals back into the reflection process.

An example of this systematic process is the case of Professor Wang, who teaches an online course in curriculum and instruction.

1. *Reflect*. Professor Wang is not satisfied with the level of reflection evident in her online students' postings in the discussion area (*identifying an area of focus*). She briefly reviews the literature to see what others have learned about improving student communication online (*reviewing what is already known*).
2. *Plan*. Using a categorization framework she found in a journal article, she *develops a plan* to look at the class data.
3. *Act*. She saves the students' online responses over a 2-week period and also interviews two of the students (*data collection*).
4. *Observe*. The data indicate that, in general, low levels of reflection are the norm (*data analysis*).
5. *Reflect*. She decides that students must not understand what is expected (*interpretation*). She *discusses her findings* with her departmental colleagues. Based on the data and what she read in the literature, the professor believes that the students need a rubric to explain the levels of communication and that some modeling might be useful.
6. *Plan*. Together with her colleagues, Professor Wang develops a rubric to use in online discussions and includes samples of the appropriate responses for the different levels.
7. *Act*. Professor Wang uses this rubric in her next class and again collects data.
8. *Observe*. Professor Wang analyzes students' online responses. That will spiral her back into the process as she observes the students' pattern of behavior to determine whether the level of communication has improved. She will again reflect on what she learns in order to make new plans and take new actions.

# ACTION RESEARCH PROBLEMS

In action research, as with all research, the first step is to determine what is to be investigated. Your focus should be on your own practice, involve something within your locus of control, be about something you feel passionate about, and provide answers to something you would like to change or improve.

## CATEGORIES OF ACTION RESEARCH PROBLEMS

We discuss problem areas in five categories applicable in educational contexts. These include both classroom-based problems and problems beyond the classroom.

First are problems arising from a desire to *improve student learning*. Questions asked by the researcher could revolve around perceived needs, such as improving the physical classroom environment for learning, improving the interpersonal interactions among students, or developing students' capacity to reflect. Questions might focus on helping an individual child or group of children or on understanding the teaching and learning context. For example, how might I best organize my first-grade classroom to promote the use of reading materials?

Second are problems arising from a desire to *improve curriculum*. Questions asked by the researcher could arise from such things as how to integrate subjects, how to best construct the curriculum, or how to embed technology use in the curriculum. Questions might focus on enriching the curricula or on developing content knowledge. For example, how can Excel (a popular computer program) be used to teach mathematics concepts?

Third, problems might arise from a desire to *adapt instructional or assessment strategies*. Questions might deal with fostering active learning, guiding student self-evaluation, or implementing a specific instructional approach. The researcher might want to experiment with new teaching strategies or techniques. For example, if I use the triarchic model of teaching—focused on analytic, creative, and practical thinking—will students' problem-solving abilities improve?

Fourth are problems arising out of a desire for one's own *professional development* or to search for connections and meaning in one's work. Questions might deal with analyzing one's own beliefs or personal style of teaching or gaining understanding about who or what influenced you in the development of your practice. The researcher may feel a desire to explore the relationship between beliefs and classroom practices or to examine the intersection of personal and professional identities. For example, what is my primary curriculum ideology and how is my instructional practice informed by it?

Fifth are problems arising from issues in the larger *school or community context*. Questions might deal with schoolwide program development, implementation, or evaluation; ways to engage families and community members in the school; or approaches to resolving tensions between groups in the school or community that impact the functioning of the school.

These five areas in which problems or focus areas may be identified in education are not necessarily discrete. They may intersect and overlap in many ways. In the classroom context, action research may be focused on the student, the teacher, the curriculum, the instructional practices, or the intersections of these areas.

**THINK ABOUT IT 18.2**

Listed below are some actual action research problems that have been investigated and written about. How do problems fit into the previously discussed categories using "S" for student learning, "C" for curriculum, "I" for instruction, "T" for teacher professional development, and "B" for broader context?

1. Mrs. Watson has noticed that some students have more difficulty memorizing their multiplication tables than others and wants to find a better way to identify those likely to have problems.
2. Mr. Rodriguez has noticed that students he teaches in the alternative school appear bored in his social studies class and wonders what instructional strategies could be used to engage them.
3. Ms. Thomas wonders how inquiry science materials could be integrated into her biology laboratory.
4. Mr. Baker is interested in knowing more about how other algebra teachers in the district teach algebra.
5. Mrs. Abbot wants to know how to create a more supportive classroom environment for one of her English language learners after noticing the child seems to have a high level of anxiety.
6. Ms. Carpenter is wondering which computer software package might be easier for students to use in developing creative projects in her art class.
7. The school principal is concerned about how community members view the new school policy allowing seniors to have independent physical education contracts that allow them to count activities such as skateboarding that occur outside of school hours.

**Answers**

1. S; 2. I; 3. C; 4. T; 5. S; 6. C; 7. B

## STRATEGIES FOR IDENTIFYING THE PROBLEM

A problem indicates a discrepancy between what is and what you would like it to be. So how do you identify the specific problem in your context? Several strategies have been recommended.

### Reflection

**Reflection** is one strategy for identifying problems. Think about your own setting and consider what is working well and what might need improvement. Think about what intrigues you about your teaching? What would you like to know more about? What values do you hold? What are your understandings of theories that impact practice? How did you arrive at your beliefs about teaching? Some recommend keeping a daily reflective journal to see if trends emerge. Others recommend conducting a reflective interview in which teachers talk through their concerns with one another. These reflective dialogues can be conducted face to face or using technology. "What if" questions may be asked as part of the reflection process and may help elicit images of what *could be*. For example, what if I begin teaching the English-speaking children in my first-grade class some basic Spanish words? Would that reduce the anxiety I sense in my English language learners?

Reflective journals can be kept in notebooks, diaries, or electronic form. Hendricks (2009) suggests several reflective **journaling** techniques that can assist in the reflection process:

Write information or jot notes as soon as you can to help jog your memory later. Even writing a word or two can help you remember important thoughts or concerns later.

Set aside time each day to write and review and to expand the detail.

Use prompts to jumpstart your writing.

Include information about the context.

Document actions you might consider taking or outcomes you might like to see.

Review the journal regularly to help you see themes or patterns.

Use technology in collaborative reflection activities (audio, video, e-mail, listservs, and chat rooms) so that there is a record that can be retrieved and reviewed.

### Description

**Description** is another strategy for determining and focusing on the problem to be investigated. Insights can be gained by describing the who, what, when, where, how, and why of a situation. These descriptions come from observations. Begin by describing the situation you wish to change or improve, describing the evidence you have that there is a problem, and describing critical factors you believe affect the situation. Other options include listing incidents or intriguing observations you have had; describing experiences; examining textbooks, curriculum documents, and lesson plans; examining demographics of students and perhaps recording observations about one subset of students; and comparing your own instructional delivery with best teaching practices advocated by leading national organizations. For example, you might keep a journal for 1 week about the interactions you observe between English-speaking children and English language learners in your class, describing the type of interactions that occur such as asking for assistance, who is engaged in the interactions (Joe and Danilo), when and where the interactions occur (during free reading time in the language center at the computer station), and your interpretation of the situation. (Danilo appears frustrated as Joe does not understand his questions about how to start the computer program.) Mertler (2009, p. 50) uses description as a way to limit the focus of study. For example, he asks

- What evidence exists that what you perceive as a problem really is a problem?
- Who are the students or groups affected by the problem?
- How or how often is the material, concept, or skill currently taught?
- How is mastery of the material, concept, or skill currently assessed?
- Where is the material tight in the scope and sequence of the course?
- When during the year is the material taught?

### Explanation

**Explanation** is a third strategy for trying to determine a specific problem for investigation. This strategy involves hypothesizing about how and why critical factors affect a situation. Sagor (2000) suggests using a technique called the "priority

pie" as a mechanism to help identify variables you believe are most relevant to an issue and to help clarify personal beliefs about the relative importance of those variables. Drawing a pictorial representation or graphic representation helps to illustrate the relationships between the variables. Once a pictorial representation is developed, ask two questions: Is this relationship meaningful? and Are you uncertain about it? If the relationship is not meaningful or you are already certain, do not waste your time investigating it. For example, Mr. Rodriguez might think that several things contribute to poor student engagement in his social studies class. He may believe that some of these factors are a greater problem than others, and he could assign percentages to determine the ones on which he really wants to focus. He might identify the following: previous failure in social studies (10 percent), no support from home (5 percent), poor reading ability (15 percent), peer pressure/culture (30 percent), intrinsic interest (15 percent), extrinsic rewards (5 percent), and few opportunities to engage in active learning (20 percent). These could be represented in an illustration (Figure 18.2).

### Literature Review

Conducting a limited **literature review** can also help in developing your explanation and clarifying the research question. Reviewing the literature helps in assessing what, if anything, other researchers have found out about the topic and what theoretical perspectives relate to the topic, as well as providing promising practices. Most people gather information from relevant sources, for example, before making a major family purchase such as a home. In action research, this review is generally more limited than that considered appropriate for more formal studies.

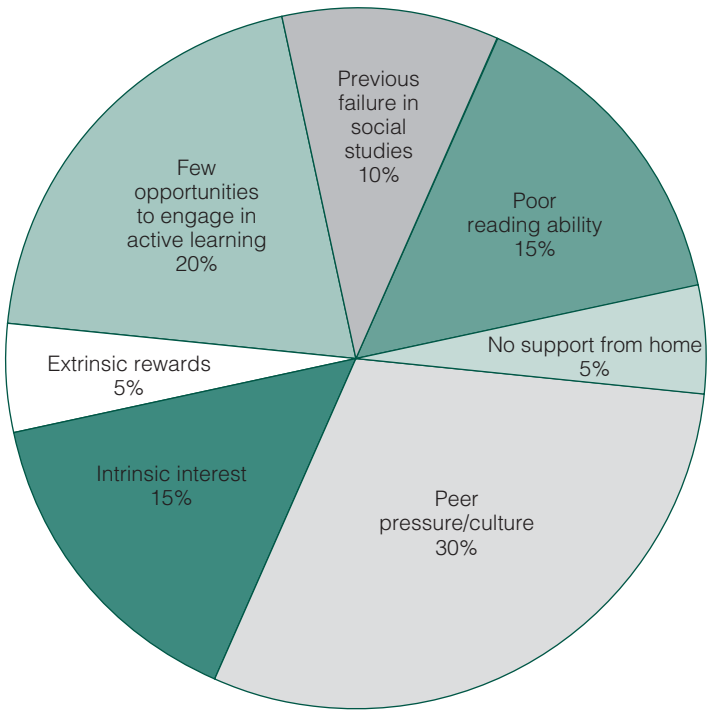



**Figure 18.2** Variables Believed Relevant to Failure in Social Studies

There are a variety of sources of literature that might be reviewed in preliminary phases of action research or later in the process as the work evolves. (See Chapter 4.) Primary sources are direct reports of original research. Secondary sources summarize primary materials. Professional literature provides perspectives from experienced professionals. Practice literature advocates particular approaches to practice. Institutional reports are done by government or other agencies. Informal reports may come from a variety of institutions (school district, clubs, committees, etc.). Media materials may also provide background information (videos, televised documentaries, and websites). In reviewing a source of literature, Hendricks (2009) recommends considering its relevance (Does it provide information that can help inform my study?), credibility (Do the conclusions appear to be supported by the data?), and similarity (Is there similarity with my setting or my participants?).

### Brainstorming

Johnson (2008) advises that if all else fails, simply **brainstorm** by drawing a line down the center of a blank sheet of paper and listing on the left side any topics of interest that come to mind. Then talk to others about some of these ideas and continue to develop the list. Once you have the topic list, on the right side begin to list specific questions for each topic.

**THINK ABOUT IT 18.3**

Use Johnson's brainstorming process to develop a list of topics you might be interested in investigating. Review these with your colleagues.

## ACTION RESEARCH QUESTIONS

As a first step in action research, the researcher must determine the focus or problem and may also at this stage identify a specific research question. A research question can help identify the variables under consideration and determine the type of data that will need to be collected. Often, the problem or focus that the researcher identifies can lead to several different research questions. For example, Principal Talbot was concerned that so few Hispanic/Latino parents attended school functions. One research question might be, "What are the factors that Hispanic/Latino parents indicate prevent them from attending school functions?" A different research question related to the same problem might be "Does sending Spanish-translated invitations directly to the home (rather than the typically English notes sent home with the students) result in an increase in Hispanic/Latino parent attendance at school functions?" In action research, the research "questions" are not always worded as actual questions.

As you develop specific research questions, you should consider several factors as suggested by Hendricks (2009): Are there potential political or personal pitfalls that could be associated with my investigation of this question? Am I interested in something specific (a particular intervention) or more general? Who will be involved in developing the research questions; just me or will others be included? Am I asking insider questions or outsider questions? Although most believe only

insider questions should be asked in action research, there are some questions driven by external factors (e.g., state or federal government) that could be considered. Finally, you should ask questions that cannot be answered with "yes" or "no" answers, that can be answered with data, and that can be answered within your limitations.

### THINK ABOUT IT 18.4

Return to the problems in Think About It 18.2. Try to create action research questions related to problems 5 and 6.

**Answers**

Problem 5. What incidents in the classroom lead to behaviors that indicate anxiety in child X?

Problem 6. Do students using X software have lower frustration levels than students using Y software?

Note: There are other possible research questions that could be developed from these problems. These are only examples. You may wish to consider developing research questions for the other problems in Think About It 18.2.

## DATA COLLECTION FOR ACTION RESEARCH

In action research, as with other types of research, different research questions require different research approaches. Both quantitative and qualitative approaches may be used in action research, and one approach is not better than the other. However, when you review action research studies, you will find that qualitative strategies are more often used. The focus is on data that are readily accessible to the practitioner.

### USING MULTIPLE SOURCES OF DATA

**Triangulation** is important in action research. Using multiple sources of data and avoiding reliance on a single source enhances corroboration of the findings. Triangulation may incorporate two or more sources of information. If multiple sources lead to the same conclusion, a stronger case is made. Discrepancies in the findings derived from the various sources lead to learning as the researcher tries to provide explanations for the discrepancies and pose new questions. Table 18.4 provides a sample triangulation matrix showing how multiple data sources can be used to examine a single problem.

### DATA COLLECTION STRATEGIES

Before collecting data, you should know why you are collecting the data, what exactly you are collecting, where and when you will collect the data and for how long, who will collect the data, and how the data will be analyzed and the findings shared. Types of data that can be collected and used in action research projects have been categorized in various ways by different authors. We use the categories described by Mills (2003) as the *three E's*: **experiencing, enquiring,** and **examining.**

**Table 18.4** Triangulation Matrix for Project Research Questions

| Research Question | Data Source 1 | Data Source 2 | Data Source 3 |
|---|---|---|---|
| Can I motivate my distance education students to participate in online discussions? | My own reflections | Student surveys | Logs showing student time online |
| What is the quality of the discussion online? | Instructor assessment using a rubric | Student self-assessment using the same rubric | Identification of levels of communication based on coding of the printed discussions |
| Do students feel comfortable expressing their thoughts online? | Student surveys | Interviews with a random sample of students | Online focused discussion in a chat room |

### Experiencing

First, data may be gathered through the researcher's own experience. This category focuses on observational data that may be recorded in various ways. **Field notes** are the most common data collection strategy used in action research to provide a record of what is going on during an observation. Field notes can include descriptions of places (locations, physical layouts, etc.), people (individuals, types, positions, etc.), objects (buildings, furniture, equipment, materials, etc.), acts (single actions that people take), activities (sets of related acts), events (sets of related activities), purposes (what people are trying to accomplish), time (times, frequency, duration, sequencing, etc.), and feelings (emotional orientations and responses). Other observational data recording strategies are listed in Table 18.5.

### Enquiring

Second, data may be collected by asking participants to respond in some manner—that is, enquiring of them. The most common action research strategy for collecting enquiring data is through interviews. It may be useful to review Stringer's (2008) discussion on questioning strategies in action research. During the first phase of study, **grand tour questions** that are global allow participants to describe something in their own terms. ("Tell me about your school?") Subcategories of grand tour questions include the following:

- Typical questions that ask participants to describe how something typically happens ("Describe a typical science class.")
- Specific questions that ask about specific events or activities ("Tell me about what happened in science class yesterday.")
- Guided tour questions that ask participants for an actual tour while providing details about the people and activities in the setting ("Can you show me around the playground?")
- Task-related questions that aid in description ("Can you draw me a map or picture of that?")

In the second phase, **extension questions** or mini-tour questions ask for more detail. ("Can you tell me more about that?") Extension questions have the same subcategories of questions as described under grand tour questions; however, these questions are derived from the initial responses to the grand tour questions.

In the third phase, **prompt questions** are used so that more details are revealed. There are three subcategories of prompt questions. Extension questions ask for more detail. ("What else can you tell me about that?") Encouragement questions or statements push the respondent to continue. ("Please go on. Yes?") Example questions do just what the name implies—ask for a specific example. ("Can you give me an example?") Other ideas for collecting enquiring data are listed in Table 18.5.

### Examining

Third, data may be collected through examining artifacts and other materials that already exist or that are routinely collected in the setting. Student records and teacher records are useful sources of information. Refer to Table 18.5 for more ideas.

**Table 18.5** Data Collection Strategies

| Experiencing | Enquiring | Examining |
|---|---|---|
| **Field Notes** | **Interviews** | **Student Information** |
| Places | Focus groups | Basic student records |
| People | Informal face-to-face | Attendance data |
| Objects | Structured face-to-face | Existing work samples |
| Acts | Telephone interviews | Portfolios |
| Activities | Internet-based interviews | Test scores |
| Events | List serves, chatrooms | Individual education plans |
| Purposes | e-mail, discussion boards | Progress reports |
| Time | Oral histories or stories | Reading records |
| Feelings | **Written Responses** | **Teacher Records** |
| **Observations** | Attitude scales | Teacher plan books |
| Active observation | Questionnaires | Written lesson plans |
| Participant observation | Rating scales | Teacher correspondence |
| Passive observation | Participant journaling | Grade books |
| Observation logs | Participant logs | **Teaching Materials** |
| Mapping or diagramming | **Performance Measures** | Curriculum guides |
| Audio- or videotaping | Portfolio development | Textbooks |
| Observation checklists | Standardized tests | Teacher manuals |
| Dialogue scripting | Work samples | Children's literature |
| Shadow studies | Conduct activity | **District/School Artifacts** |
| **Other Collection Strategies** | **Other Collection Strategies** | Memos |
| Diaries | Community meetings | Parent newsletters |
| Journaling | | Minutes and official reports |
| Photographing | | Policies and procedures |
| Rating scales | | Evaluation reports |
| | | Press accounts |
| | | Public relations materials |
| | | **Other Archival Sources** |
| | | Films |
| | | Photographs |
| | | Maps |

Recall that Mr. Rodriguez has noticed that students he teaches in the alternative school appear bored in his social studies class and wonders what instructional strategies could be used to engage them. Mr. Rodriguez could briefly examine the literature on student engagement (find one or two sources). He could make written observations of student behavior in his social studies class to determine if there is a pattern. He could individually interview students and ask them to describe their experiences in social studies. He could videotape the class to observe his own teaching practices and students' responses. He could conduct focus groups with the students to ask about their interests and their experiences in school. Based on the data collected, Mr. Rodriguez might then try some different instructional strategies and determine if students demonstrate more engagement.

**THINK ABOUT IT 18.5**

See if you can identify at least two data collection strategies for the two action research questions identified in Think About It 18.4.

**Answers**

Problem 5. Observation logs, interviews, anxiety measures

Problem 6. Videotaping, student surveys, interviews

Note: These are examples. Other strategies could be used.

## RIGOR AND ETHICAL TREATMENT IN ACTION RESEARCH

Regardless of the data collection method used, the researcher must strive to ensure the rigor of the process and ensure that the participants are treated ethically. No research, including action research, is worth doing unless it is done well. Action research in schools should be held to high standards.

First, as educators, we have an obligation to students. Actions based on poorly done research can place students at risk. Second, educators need personal and collective efficacy. Third, there is a need to add to the professional knowledge base in education, whether that knowledge is added informally through sharing with colleagues or more formally through publishing.

### RIGOR IN ACTION RESEARCH

Action researchers should be concerned about the issues of **rigor** or quality addressed by other researchers: validity, credibility, reliability, dependability, neutrality, confirmability, and transferability. These concepts are covered in other chapters of this text. There are a few comments about rigor in action research, however, worth noting here.

Action research in schools often relies on **authentic student work,** which Sagor (2000) compares to primary source materials and claims enhances credibility. **Credibility** is described as the researcher's ability to take into account the complexities that present themselves in a particular setting and to deal with patterns

not easily explained. Mills (2003) asserts that if action research is to be viewed as credible, the solution to the problem (the planned intervention) must actually solve the problem.

Being able to generalize is not a primary goal of action research; rather, the primary goal is to understand what is happening in a specific context and to determine what might improve things in that context. Action researchers believe that everything is context bound and that the goal is not to develop a generalizable statement but to provide rich and detailed descriptions of the context so that others can make comparisons with their contexts and judge for themselves whether the findings might apply (be transferable). Hendricks (2009, pp. 113 and 115) provides summary tables addressing rigor in action research and strategies to enhance it, which are combined in Table 18.6.

**Table 18.6** Strategies for Increasing Rigor in Action Research

| Type of Rigor | Focus | Strategies |
|---|---|---|
| Truth-value validity ("Are my results accurate?")<br>Process validity ("Did I use appropriate processes to study the problem?") | Accuracy of facts and findings; correct interpretations and conclusions | Persistent and prolonged observations<br>Triangulation<br>Accurate data recording<br>Member checks<br>Peer debriefing<br>Negative case analysis<br>Biases made clear |
| Outcome validity ("How will I use the results?")<br>Catalytic validity ("Are my views or practices changed?") | Ability of study to increase understanding, resolve problems, and transform practices | Presentation of results<br>Continuous, ongoing reflective planning |
| Applicability/transferability ("How might the results be useful to others?")<br>Consistency/dependability ("Would I get the same results with other students?") | Usefulness of the results in different settings and contexts and with different individuals | Thick description of the setting, study, and participants<br>Ongoing investigation |
| Neutrality/confirmability ("How do I show results do not reflect my personal desires or biases?") | Evidence that results are accurate and not a result of researcher bias | Peer debriefing<br>Accurate data recording<br>Member checks<br>Triangulation<br>Biases made clear<br>Audit trail |
| Democratic validity ("In what ways did I involve stakeholders?") | Evidence that stakeholders collaborated and/or were given a voice in the process | Peer debriefing<br>Member checks<br>Audit trail |
| Dialogic validity ("Do these results make sense to others?") | Sharing research findings, seeking feedback about the accuracy of interpretations and conclusions | Peer debriefing<br>Member checks<br>Presentation of results |

*Source*: From Cher Hendricks. *Improving Schools Through Action Research: A Comprehensive Guide for Educators,* 2nd edition. Published by Allyn and Bacon/Merrill Education, Boston, MA. Copyright © 2008 by Pearson Education. Reprinted by permission of the publisher.

Action researchers need to understand and reveal their biases, as do other researchers. Like all other researchers, action researchers do not collect data to simply validate existing practices, do not ignore discrepant data, and do not review only literature that supports a particular view. They are open to reflection.

### ETHICS IN ACTION RESEARCH

Ethics in action research is based on the same principles as ethics in other research. The first principle is that participants should not be wronged in the name of research. It is just as important to abide by ethical guidelines in action research as it is in other forms of research. However, action researchers may have some unique challenges to overcome in ensuring ethical principles are upheld. Action research is often more open ended and may change as the researcher focuses differently on the problems in the context. In action research, there is little distance between the researcher (e.g., the teacher in the school setting) and the subjects (the students in the teacher's classroom).

There is some argument in the field of action research about the need for **informed consent.** The case made is that the subjects are actually the researchers (the teachers) when the purpose of action research is to improve one's own performance. Proponents argue that students in classes, for example, are simply and naturally living through the teacher researcher's instruction—the same instruction they would have been receiving without action research. However, most experts in the field believe it is wise to obtain permission, particularly if the researcher wishes to present or even publish the findings to others outside of the local context. If you are using the information as part of a university requirement, it is best to review the institutional review board (IRB) requirements and submit a plan for approval. If conducting action research in a school or district, you are advised to consult with the appropriate offices for information on approved procedures.

Given the scrutiny education is under at the current time, action researchers in schools would be prudent to obtain **parental consent.** Sagor (2000) recommends a generic letter to parents that conveys four key points: (1) The teacher is conducting research for him- or herself that is intended to benefit the children in the class; (2) the research will not mean different things will happen to some of the children, and nothing will be granted or denied due to the research; (3) the teacher might use the child's work, words, or ideas in reporting on the research; and (4) there will be no negative consequences if permission is denied. Data about that child will then not be used in the study.

## DATA ANALYSIS IN ACTION RESEARCH

What do you do with the data you have worked hard to collect in a dependable, accurate, reliable, and correct manner? **Data analysis** involves reviewing the data while they are being collected and attempting to synthesize and make sense out of what is observed.

Much of what has been written about analysis in action research mirrors strategies used in qualitative research, although the researcher should always

remember that appropriate analysis will depend on the question asked and the method of data collection used. Action research data analysis has been described as a search for patterns or trends in the data to answer two questions: What is the story told by the data? and What might explain this story or what factors influenced the story? The action research analysis process has been compared to trying to put together a jigsaw puzzle without the box that shows the completed picture and with some pieces missing.

## CODING

One key analysis strategy often described in action research is **coding** as typically described in qualitative research. First, the researcher breaks down and categorizes the data into manageable segments **(open coding).** Then, the researcher puts the data back together again, making connections between and across categories **(axial coding).** Sometimes, the researcher has a clear and selective focus and is systematically reviewing the data for that specific category **(selective coding).**

For example, in the case of Mr. Rodriguez trying to teach social studies to the students in his alternative school, Mr. Rodriguez interviewed the students and coded their comments. Mr. Rodriguez tried to make sense of the comments and coded them as noted in Table 18.7. He has tried to summarize what the students were saying in terms of a category (*open coding*).

Next, Mr. Rodriguez tries to collapse the many categories into fewer categories by attempting to identify commonalities (*axial coding*). He recalls reading an article on student engagement in the journal *Educational Leadership* (Newmann & Wehlage, 1993) in which the authors described a concept known as authentic work (e.g., meaningful versus nonsensical). Authentic work is identified by its connection to the real world, its ability to provide a sense of ownership for students, and its use of extrinsic rewards and intrinsic interests. Mr. Rodriguez notes several of the comments he coded in the open-coding phase (real-world connections, intrinsic interest, and extrinsic rewards) seemed to fit with the description in the literature of authentic work. He combines the coded comments into a broader category called *authentic work*.

Mr. Rodriguez then has several statements left coded as *active learning strategies*. He believes this is an important concept and is different from authentic work. So, he begins looking for other examples in his interview data that describe active learning (*selective coding*).

## STAGES OF ANALYSIS

There are two stages of action research analysis, description and **sense making.** During the description stage, you review the data and ask yourself what did you see and what was happening (e.g., when Mr. Rodriguez was conducting the initial coding). During the sense-making stage, you try to consider how the pieces fit together and what stands out (e.g., when Mr. Rodriguez interpreted how the pieces fit together). During the sense-making stage, data may be organized in different ways based on such things as chronology, people, key events, key issues (as in Mr. Rodriguez's case), episodes, and the like. Organizing units emerge from the data as they are grouped and sorted into themes.

**Table 18.7** Coding of Student Comments

| Student Comments | Coded category |
|---|---|
| "Usually we like did chapters and assignments in the book, but one time I remember to this day, we had an assignment like we had a certain amount of money, like $2,000, and we had to get an apartment and a uniform for the job, and pay the electric bill and get food for the month and all this stuff. And we had to actually go to a grocery store and find out what stuff costs instead of just reading about it in class." | Real-World Connections |
| "He'd give us Snickers or candy or take us to lunch or something if we did really good." | Extrinsic Reward |
| "Sometimes he would bring in a movie or let us bring in a movie and bring pop and chips for everyone if the whole class did well." | Extrinsic Reward |
| "I think you should teach about other people instead of about the stuff you always see in books like Ancient Rome and Greece and all that, like teach about stuff from another country that people never heard of and never thought of before and I think they'd be more interested because it's something new." | Intrinsic Interest |
| "The teachers just sit there and talk and make you take notes. I would rather like them to explain it in more detail and like tell stories." | Active Learning Strategies |
| "Teachers should have us participate and like give us rewards and stuff." | Extrinsic Rewards |
| "In the history books, I really didn't like the stuff about Columbus and slavery and stuff, but we never really talked about anything about Latinos and Hispanics and stuff and that would be more interesting for a lot of Hispanics." | Intrinsic Interest |
| "I think the problem is that sometimes its stuff, material that isn't relative to our lives, personally. On areas that really do pertain to our lives, you know, are relevant to what we're doing today, that's what you should emphasize." | Real-World Connections |
| "The first thing he'd do was to talk about what was in the newspaper and we'd talk about what was in the newspaper. Then he'd ask us what was going on in school and at home and how that connected. Then after we got done reading and discussing he'd like everybody loosen up and move to a group and help each other out and we'd be like laughing. Most teachers won't do that, they want you to do your own stuff and get your own work finished, but he would say 'you can all work together' like he knew you was all thinking and really helping each other." | Active Learning Strategies |

## DATA INTERPRETATION IN ACTION RESEARCH

**Data interpretation** focuses on the implications or meanings that emerge from the analysis. Interpretation is used to help make the experiences being studied understandable, using description and conceptual frameworks or theories. In the interpretive step, you ask how the patterns in the data inform your thinking and what the patterns might mean. Themes, metaphors, similes, typologies, and vignettes may emerge in the interpretation stage. Claims that you make based on your interpretations must be supported by the data. Asking interpretive questions (why?) may help visualize the situation.

### USING VISUALS

**Concept mapping** can be used to plot elements diagrammatically so you can visualize what different components of the situation relate to the problem under investigation. **Problem analysis** using visuals of antecedents and consequences

can also be helpful in interpretation. Similar to concept mapping, problem analysis identifies antecedents to the existing problem (what led up to it) and consequences that derive from the problem. Concept mapping does not try to identify what came first and what came later.

For example, let us consider the problem posed by Mr. Rodriguez in his social studies class: lack of student engagement in academic work. He might create a concept map as shown in Figure 18.3. Or Mr. Rodriguez might develop a problem analysis to look at the antecedents and consequences related to the problem of academic disengagement as pictured in Figure 18.4.

## REFLECTING

The interpretation phase of action research is a process of ongoing reflection and is the most challenging aspect of action research. The researcher continuously reviews the data as the action research process unfolds, remembering that any interpretations reached and conclusions arrived at are not for all time, are not generalizable, and are certainly not conclusive.

During interpretation, the researcher considers a number of questions: Would I do anything different next time? What did I learn about my practice? Did I answer my question? Do I need to look at additional or other questions? Did I collect the right data to answer my question? Are there other things I should collect in the

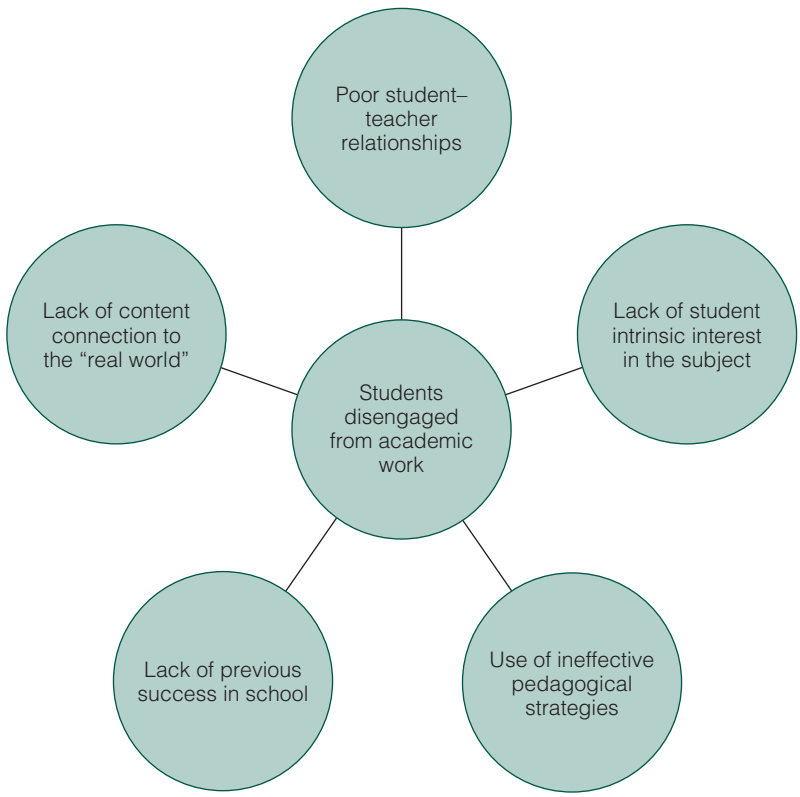



**Figure 18.3** Concept Map Showing Potential Factors Leading to Student Disengagement from Social Studies

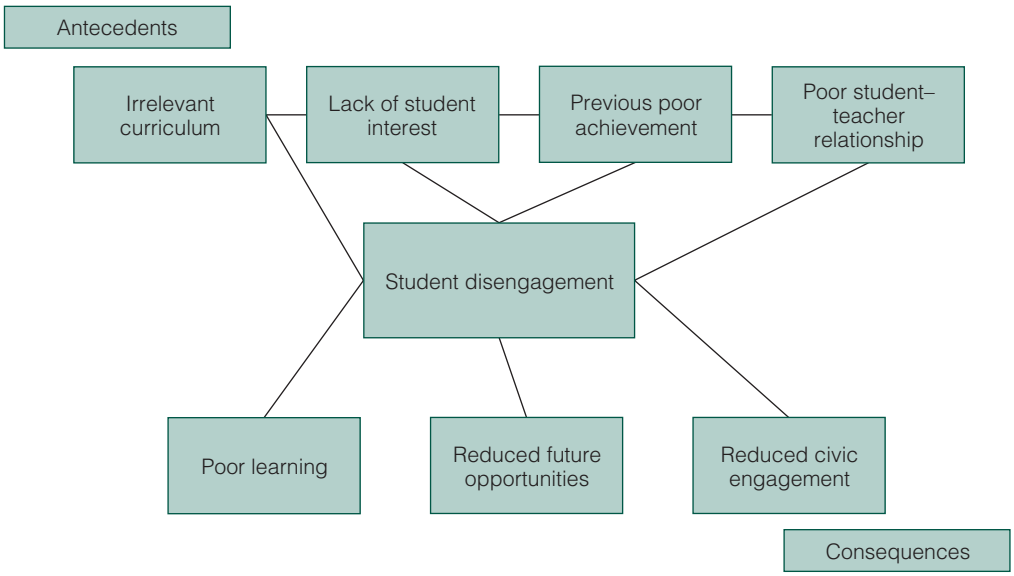



**Figure 18.4** Antecedents and Consequences of Student Disengagement from Social Studies Work

future? Reflecting on these and other questions helps the researcher consider both the next steps in terms of actions and the next steps in continuing to examine practices.

Data interpretation in action research is about making educated guesses or reasonable inferences. Once drawn, the interpretation can be connected with personal experience and contextualized. The interpretation provides a rationale for action planning. After interpretation, the researcher must decide what the implications are for practice.

## ACTION PLAN

The most important step in action research follows analysis and interpretation. That step is acting on the knowledge you have gained. What do you believe is an effective choice or course of action based on what you now know? What will you do differently? Did you discover a new problem? Does something need modification? But before you act, you must develop an **action plan.** That plan can be formal or informal, depending on the nature of the research project.

The goal of action research is to take action based on the findings of the study. The findings provide insights for changing practice. The purpose is to formulate practical solutions to the problems that have been the focus of research. The action plan sets priorities for projected actions in order of importance; sets goals for the actions required; lists objectives required to accomplish the goals; stipulates the sequence of tasks; specifies who will carry out tasks; identifies who must be informed or consulted for permission; determines the time frame for task completion; and calculates materials, equipment, and funds required to complete

the task. Although this might sound daunting, in action research this process is generally less formal. For Mr. Rodriguez, the action plan might focus on identifying different teaching strategies (addressing active learning) to incorporate into his teaching repertoire and/or on reviewing the curriculum to determine how historical perspectives from other cultures (addressing intrinsic interest) might be included.

Given what is known from the research, the researcher must determine what precisely to do, what is the course of action. This step also returns the researcher to the problem formulation step (remember the cycle discussed previously?). Action researchers often use a matrix or guide to complete the action plan. Action plans can be developed at different levels: by the individual, by a team, or at a school or larger organizational level.

Recall the problem in the distance education class mentioned previously. The individual instructor could develop a plan to address the problem. A group of faculty who share similar concerns could decide to collaborate in addressing the problem. At a campus level, a faculty development center could implement training to address the problem. Table 18.8 is an individual action plan for addressing this problem.

**THINK ABOUT IT 18.6**

Review Mr. Rodriguez' data interpretation and develop an action plan matrix (as in Table 18.8) for him to implement in his class.

**Table 18.8** Action Plan

| Key Finding | Action | Specific Tasks | Who Responsible | Who to Consult or Inform | When (Time) | Where | Resources Needed |
|---|---|---|---|---|---|---|---|
| Students in the online class exhibited lower levels of communication patterns (initiating and supporting) rather than higher level communication (challenging, summarizing, and monitoring). | Explicitly teach the five levels of communication to the students. | Develop a handout for the students. | Teacher | Institution Policies | August | Office | Paper |
| | | Model the different levels. | Teacher | Students | First week of fall classes | During class time | None |
| | | Students have practice session. | Students | Supervisor | Second week of fall classes | In lab | Access to website<br>Lab fee<br>Graduate assistant to help monitor |

# REPORTING ACTION RESEARCH

Besides developing and implementing an action plan, it is also important for action researchers to share their findings with others. In school settings, it is important to invite collegial dialogue about the implications of the research. You want other educators to hear what you have learned, and you want to hear their reactions to the action plans you have developed. Informal strategies for sharing findings might include simple conversations with peers. More formal report writing might be shared with the school board or presented at a conference.

## COMPONENTS OF THE REPORT

In determining how the action research will be reported, a key consideration is the audience: Who are you reporting to and why? The answers to these questions determine the format, length, and amount of detail to include in the report. The research may be reported in written form or through oral presentation, interactive sessions (e.g., group work), and video and electronic media. Recently, we have seen presentations using artistic and dramatic performances. Most written action research reports have sections similar to other reports but tend to be much shorter.

An outline for an action research report may include an area of focus statement or purpose statement, a summary of the related literature (usually very limited), the research questions, a description of any intervention or innovation, a description of data collection strategies used, data analysis and interpretation, and the action plan, as well as reporting the results obtained after implementing the action plan. Those familiar with the sections included in other types of research reports will find this outline very familiar.

## PUBLISHING AND JUDGING REPORTS

Some action researchers may be interested in publishing their results, although this is not typically the goal in action research. There are several action research journals to which researchers may wish to submit such a report: *Action Research International, Educational Action Research, The Qualitative Report, The Ontario Action Researcher*, and the *Action Research Electronic Reader*. Many more mainstream journals also routinely publish action research, such as the *Journal of Research in Science Teaching*, the *Journal of Research Technology in Education, Mathematical Teaching in the Middle School, Professional School Counseling Journal, Teacher Education Quarterly, Teaching Exceptional Children*, and *The Reading Teacher*.

If action research is to be shared, there should be criteria for judging the quality of those reports. Some argue that the criteria for action research should be based simply on whether the study produced change that resulted in a solution to the problem. Mills (2003) provides a checklist for the practitioner to judge the effectiveness of the educational action research effort (Table 18.9).

Mr. Rodriguez might answer these questions in the following manner. His research led to action. To address intrinsic interest, he reviewed his curriculum and added historical perspectives from other cultures, particularly the African American and Latino cultures, representing the backgrounds of the majority of

**Table 18.9** Mills Criteria for Judging Action Research

| | |
|---|---|
| Interactability of reform | Does your action research lead to action? |
| Audience | Who is the intended audience for the report? |
| Format | Have you presented the report in an acceptable format? |
| Prejudices | Have you shared prejudices that may affect your findings? |
| Professional disposition | Has the research contributed to your professional disposition? |
| Reflective stance | In what ways has it contributed to your reflective stance? |
| Life-enhancing | How have your efforts enhanced the lives of students? |
| Action | What action have you taken? |
| Action–data connection | How is the action connected to data analysis and interpretation? |
| Impact | How will you monitor the effects of your practice? |
| Changes | What would you do differently next time? |
| Colleague response | How did your colleagues respond to your findings and the actions recommended? |

*Source*: Adapted from Geoffrey Mills, *Action research: A guide for the teacher researcher*, 2nd ed., 2003, p. 157. Reprinted by permission of Pearson Education, Upper Saddle River, NJ.

his students. He shared his findings with the other teachers in the alternative school, both in a verbal report at the next teacher in-service day and via a 5-page summary report. He discussed with his colleagues his biases when he began the study, thinking that the major reasons for lack of engagement in the class were problems with the students, not with his teaching. He shared that the research experience and the findings have changed his dispositions and his approach to classroom teaching. Participating in the study also has made him more reflective about his practice. Mr. Rodriguez took action to change his curriculum and hopes that by creating a more inclusive curriculum, students will be more engaged and be more successful in their academic careers. The action Mr. Rodriguez took was connected to the data that indicated a need for more authentic work in the classroom, an element of which was work that held intrinsic interest for the students. His interpretation of student data was enhanced by his connecting it to a framework he read about in the literature. He plans to continue monitoring what is happening in the classroom through ongoing observations and a focus group with students next semester, as well as reviewing student progress reports. He will probably not repeat the individual interviews because they were very time-consuming. Mr. Rodriguez's colleagues are impressed with his findings, and together they begin a study group to examine practices in the school, focusing on ways to create more authentic work for students. They encouraged Mr. Rodriguez to publish his findings in an action research forum.

Other criteria the reader might consider in determining the quality of an action research report include Hendricks' (2009) Checklist for Action Research Reports and Johnson's (2008) Criteria for an Action Research Report.

## INCREASING USE OF ACTION RESEARCH IN EDUCATION

Advocates for action research in education and cases of teachers doing research date back more than 50 years, but there has been a recent surge of activity in this area. Professional organizations, graduate programs, and state and federal government are increasing their focus on action research and support for teacher researchers. The National Board of Professional Teaching Standards promotes approaches such as action research and requires teachers to read the professional literature in their fields and engage in data gathering, analysis, and reflection. Why action research now?

Sagor (2000) makes the case that the focus on action research can help to professionalize teaching, enhance teacher motivation and efficacy, better meet the needs of diverse learners by searching for solutions to ever-changing problems, and help schools achieve success in a standards-based environment. The standards movement and its demands for high accountability, high standards, and quality assessments, often with expectations set in legislation, presents a challenge to educators. One reason for action research gaining priority in the United States has been the passage of the No Child Left Behind legislation that pushes educators to study the effects of teaching on student learning. Demands are placed on educators to be more deliberate in documenting and evaluating their efforts.

Action research has been held out as a mechanism both for individual teachers to improve their practice and for entire schools to implement reform (Thomas, 2005). It has been closely tied to what has been called teacher research and has increasingly been incorporated into degree work, certification programs, professional development, and school reform. For example, in master's degree programs throughout the country, action research projects are an option for a culminating experience in place of more traditional requirements, such as comprehensive examinations or theses. In one professional development program designed to assist teachers in becoming certified by the National Board of Professional Standards, action research is a key component and is tied to teachers' reflections on their practice. Proponents of action research in education assert that one key purpose of action research is to improve the practice of the individual teacher by building reflective practitioners. Teacher action research is a vehicle that can be used by teachers to untangle some of the complexity that occurs in the profession, raise teachers' voices in discussion of educational reform, and transform assumptions about the teaching profession. Action research honors the skills of educators and their inclination to influence their environment with an aim toward improving it.

### ACTION RESEARCH IN PROFESSIONAL DEVELOPMENT AND SCHOOL IMPROVEMENT

Tying action research to staff development and professional growth evaluation is recommended in the literature as a part of changing the school culture and is already in practice in some schools. Action research is increasingly being used for

district, state, and national school reform initiatives. In a growing number of school districts, part of the school improvement plan calls for teachers working in teams to gather data as the school implements new strategies in order to examine the impact on students and to make recommendations for changes in implementation. Using action research in schools can help teachers develop a common focus, build a professional culture and eventually a community of learners, and promote organizational learning.

## STUDY GROUPS

Incorporating action research through **school-based study groups** can provide a means for collaboration as well as a source for energy and support. Dana and Yendol-Silva (2003) describe the following ways to collaborate: (1) In *shared inquiry*, teachers define and conduct a single research project together; (2) in *parallel inquiry*, teachers conduct two parallel but individual studies, working to support each others' individual endeavors, at times collecting data for each other, and discussing findings together; (3) in *intersecting inquiry*, teachers explore different questions but about the same topic; and (4) *inquiry support* is a method of inviting others to serve as critical friends to help formulate meaningful questions, design the project, and aid in collection and analysis of data. Collaborating in action research makes education a less isolating activity and enables teachers to support and help one another. It moves the conversations to discussions of best practices and interpretation of data and enables better informed decisions to be made. It broadens the perspectives of educators by expanding the range of experiences on which to draw.

## ACTION RESEARCH AND PROFESSIONAL DEVELOPMENT SCHOOLS

The field of education has seen growth in the creation of **professional development schools** (PDSs). The National Council for the Accreditation of Teacher Education (NCATE) describes professional development schools as innovative institutions formed through partnerships between professional education programs and prekindergarten to grade 12 schools with a fourfold mission: (1) enhanced student achievement, (2) teacher development, (3) inquiry directed at the improvement of practice, and (4) the preparation of new teachers. NCATE has developed standards for PDS work that include a focus on collaborative inquiry to identify and meet students' learning needs, effect the learning of teacher candidates, and determine the professional development agenda for teachers in the schools.

The National Association for Professional Development Schools also has developed a set of nine elements for PDS work. Reflective practice and deliberative investigations of practice are included in these elements.

PDSs are often compared to teaching hospitals, hybrid institutions created in the early 20th century. As practicing professions, both teaching and medicine require a sound academic program and intense clinical preparation. The teaching hospital was designed to provide such clinical preparation for medical

students and interns; a PDS serves the same function for teacher candidates and in-service faculty. Both settings provide support for professional learning in a real-world setting in which practice takes place. Collaborative action research, in which researchers from both the university and school mutually define problems in the school and work to mutually solve them, is recommended in PDSs and is occurring in PDS sites throughout the country. In the action research process in PDSs, students benefit from improved practice, teachers develop research skills, university faculty develop field-based methods, and both cultures are professionally renewed.

### CHALLENGES

Of course, there are challenges to implementing action research in schools. These challenges can include lack of resources, the amount of support received, resistance to change, the nature of colleagueship in the school, reluctance to interfere with others' professional practice, reluctance to admit difficult truths, the amount of encouragement received from school leaders, the climate for risk taking in the school, and the commitment for making time for action research endeavors. However, most in the education research field would agree that using research methods to inform practice is a positive trend in schools and is worthy of support.

## RESOURCES FOR MORE INFORMATION

The following are some online resources available for those interested in action research:

| | | |
|---|---|---|
| ALARPM | Action Learning and Action Research Association | www.alarpm.org.au |
| ARN | Action Research Network | http://actionresearch.altec.org |
| AROW | Action Research Open Web | www2.fhs.usyd.edu.au/arow |
| ARR | Action Research Resources | www.scu.edu.au/schools/gcm/ar/arhome.html |
| CARE | Centre for Applied Research in Education | www.uea.ac.uk/care |
| CARN | Collaborative Action Research Network | www.did.stu.mmu.ac.uk/carnnew/ |
| CARPP | Centre for Action Research in Professional Practice | www.bath.ac.uk/carp |
| PARnet | Participatory Action Research Network | www.parnet.org |
| Queen's University | Action Research at Queen's University | http://educ.queensu.ca/~ar |
| TAR | Teacher Action Research | www.edchange.org/multicultural/tar.html |

## ACTION RESEARCH EXAMPLE

The following example of an action research report is taken from *Networks, 10*(2) (2008).

# *Elementary Professional Development within a 'Practical' Action Research Effort to Improve Student Literacy*

**T. G. Ryan, A.M. Aquino, D. Berry, K. Clausen, R.L. Wideman**

### PURPOSE

The purpose of this inquiry was to support and augment the action research efforts of elementary teachers who were attempting to enhance literacy outcomes in their respective classrooms. Included are elementary teacher insights, university-based facilitator views, and principal perspectives that together complete a picture of our professional development efforts. Together the data provide an overview of an action research effort, wherein praxis was noted as a necessary element to assume 'practical' investigative roles. Praxis herein is the deliberate, informed, planned, and systematic action which is the critical underpinning of all action research efforts. The action in this case was directed towards improvement and implementation of an instructional initiative. This outcome brings with it an immense level of significance in that all educators seek to improve educational outcomes personally, professionally, and politically; therefore a report such as this may be viewed as an essential tool to refine educational practice.

What follows are several rudimentary understandings within the action research landscape. Specifically, a discussion of what it means to be a reflective teacher is followed by discussion of group action research that leads into the context of this inquiry. The questions raised and our methods to address these probes are detailed within our methodology. The analysis and interpretation of evidence guide us to an informed view of professional development. Finally, our conclusion is laid out as are the future plans for this enterprise. Ultimately, all participants agreed to work together to realize increases in student literacy.

### INTRODUCTION

#### A Reflective Teacher

Reflective professional development is never complete; there is always something else to consider, and often the process of reflection within action research is a social enterprise, as Carr and Kemmis (1986) point out:

> Action research is simply a form of self-reflective enquiry undertaken by participants in social situations in order to improve the rationality and justice of their own practices, their understanding of these practices, and the situations in which the practices are carried out. (p.162)

Self-reflective teachers will improve and change if something doesn't meet their expectations. It seems logical to make changes until you are satisfied with the outcomes, engaging in particular actions in order to ensure desired outcomes. These actions combined with reflection become praxis, the fundamental concept of action research. Praxis—deliberate, informed, planned, and systematic action—is a critical underpinning in all action research efforts. Action is usually aimed at improvement and, at the same time, may be intended to implement a new theory, program, or initiative. The teacher acting as "the action researcher is interested in the improvement of the educational practices in which s/he is engaging. He [sic] undertakes research in order to find out how to do his job better—action research means research that affects actions" (Corey, 1949, p. 509). Admittedly, a teacher's action may be solely individualistic within his or her classroom and aimed at improving some aspect of practice; yet there are many school and social implications of such change which can impact the larger community of teachers in a school.

**Reflective Teachers: A Group Endeavour**

When a group of teachers undertake substantive actions in order to achieve better results, as is often the case in some 'team' oriented schools, momentum and commitment build within a school and the larger community. Yet, we need to be reminded that "action research combines a substantive act with a research procedure; it is action disciplined by inquiry, a personal attempt at understanding while engaged in a process of improvement and reform" (Hopkins, 1993, p. 44). Indeed, it is informed action that underpins the very nature of action research (Altrichter, 2005).

> Action research is, therefore, a deliberate way of creating new situations and of telling the story of who we are. Action research consists of deliberate experimental moves into the future, which change us because of what we learn in the process. (Connelly & Clandinin, 1988, p. 153)

Perhaps action research is "best thought of as a large family, one in which beliefs and relationships vary greatly.… [I]t is a group of ideas emergent in various contexts" (Noffke, 1997, p. 306).

**Our Inquiry**

In our project we embraced the elementary classroom practice of literacy development and revisited several issues (i.e., assessment/evaluation, instruction, planning) during our interviews and daily praxis. Our work was led by classroom teachers and merely supported and facilitated by both local school administration and university-based faculty. Our social practice was aimed at improvement and our actions were cyclical so as to plan, act, observe, and reflect recursively. As we collaboratively acted, we understood that each of the participants may realize greater understanding and control of their learning. We also realized that our efforts could give way to deeper commitments. As McNiff, Lomax, and Whitehead (1996) argue,

> To be action research, there must be praxis rather than practice. Praxis is informed, committed action that gives rise to knowledge rather than just successful action. It is informed because other people's views are taken into account. It is committed and intentional in terms of values that have been examined and can be argued. It leads to knowledge from and about educational practice. (p. 8)

Therefore, without praxis (informed, committed action), classroom practice may stagnate and remain ill-conceived and narrow. One of our goals as professors was to nurture, support, and enhance teachers' development as they attempted to improve literacy and outcomes in their classroom while addressing three questions:

1. Can a group of teachers engaged in a practical action research project improve student literacy?
2. Can a group of educators maintain the necessary praxes required to enhance literacy and development in classrooms?
3. What unexpected outcomes will surface as a result of our efforts to complete a practical action research project?

In addition the university facilitators asked:

1. Can we realize the development of action theory (new understanding) and action practice (application) as it relates to teacher development and student growth?
2. Can we nurture recursive cycles of theorization/reflection and application in order to professionalize practice?

We believed from the onset that "action research is one way of restoring and enhancing professional confidence.… [W]e must, however, be aware of problems associated with too prescriptive a framework for action and the values that are embedded within it" (Hopkins, 1993, p. 56). What seems fundamental to action research is that it involves participants talking about everyday things in the life of education and unpacking them for their historical and ideological baggage (Connelly & Clandinin, 1988; Noffke, 1995; Woolhouse, 2005). The conversation can be considered the action since "conversation can play a significant role in the establishment and sustention of collaborative action research groups, and … can lead to the generation of new knowledge and understanding" (Feldman, 1999, p. 129). Our position was that most studies derive most of their action and knowledge from participants' conversations and not so much from the actions in the classroom. The conversations themselves were the "'glue' for maintaining the integrity of the group" (Feldman, 1999, p. 129) and facilitated openness to new possibilities. Communication was critical to significant action in the future, as the participants used their new understanding to develop new praxes.

## METHODOLOGY

### Our Context and Action Research

One can argue that the educational context should define the nature of praxis, and similarly, the educational context should define the nature and conduct of the action research group. Action research roles are embedded in social

contexts by the very purpose of the action research. As the context or setting changes, so can the purpose of the inquiry and the way it is conducted. Therefore, the role and commitments of the participant and facilitators are very much tied to context, setting, and purpose. Our study was located in a Catholic school of approximately 250 elementary students from a residential area of Central Ontario. The school contained nine classrooms, including a kindergarten classroom, Primary Learning Assistance Centre (LAC) for students with severe exceptionalities, library, gymnasium, and large schoolyard. Staff, including teachers, assistants, office, and custodial, totaled 29, and the teachers had a wide range of teaching experience and qualifications.

**Participants and Data**

Participants included two classroom teachers and one school principal. Within our inquiry, it was important to note that at the classroom level, assessment methods in all grades included variety of strategies such as observation, self- and peer evaluation, projects, portfolios, presentations, and classroom tests. As well, provincial testing was completed annually by the EQAO (Education Quality Accountability Office), which annually assesses Reading, Writing, and Mathematics for all students in grades three and six. Results were sent to our school each Fall. Students were assigned a score between level one (the lowest) and level four (the highest). Students performing at level three are meeting provincial expectations. Details are noted in the conclusion.

It is important to include such contextual information since any effort to reduce and decontextualize the social world is to misrepresent the situation that is the focus of the study, argument, or question (King, 1988). The real strength of action research is its capacity to recognise the complexity and uncertainty of educational contexts.

The following descriptions, using pseudonyms, were extracted from interviews to provide academic background for Ann, Pam, and our principal (Ruth).

**Teacher: Ann**

> My experience is mainly in grade six. I have taught for four years now—[grades] one, five, six, two years of six/seven and this is my first year of a straight six. So my experience remains in six and [I] worked with some split grades. This is my first year at this school. I've been in High Bay. I have had a lot of identified students. I've had maybe ten that weren't students. I've had maybe ten that weren't [identified] out of twenty six. So a lot of experience in special needs but no formal training, no Special Ed aside from teachers' college. My qualifications…I have my junior, intermediate, and senior qualifications. My teachable subjects are English and History. I intended to go into high school, but I took my junior in the spring and then got a job, and I've really enjoyed Elementary. My areas of enjoyment are language and when a student gets it, like the moment that you see that they have comprehension and they get it, that's my favorite part about teaching. (Interview, October, 14, 2004, p.1)

**Teacher: Pam**

I have three years teaching. My first year was mixed with contracts and supply (part-time/substitute) teaching. I have had two years as a full time teacher. This is my second year here at this school. I had a contract for [grades] three/four at *Northern* Public School. As an LTO (Long Term Occasional/Part-time teacher), yes. When that ended at Christmas, I got on right away with both Boards [Districts] to supply [substitute teach]. Now I teach…last year was four/five split grade, this year it's a five/six split grade. Most of the children are from last year, so it is okay. I have my special education (my specialist, my three parts). I have a Master of Arts in Education from *Central* University. I did that as a social worker. So prior to being a teacher I have a social worker background and I was also a banker. I was five years with the Bank of Canada. So I sort of have an eclectic background which I think is helping me big time in teaching for stress, organizing, and for pure enjoyment. I really did not think I could be a teacher until I had my daughter. When I had my daughter, I realized children are just little people like us and opened up the world of teaching to me, and I went back to school to teachers' college for a year, and here I am. (Interview, October 13, 2004, p. 1)

**Principal: Ruth**

As a teacher, I had six years experience, junior kindergarten right through to grade eight, and my last four years of the six being in the intermediate. This is my first year as an administrator; therefore, I have special education experience, [experience as] a religious education specialist as well as a lot of experience and enjoyment in professional development like delivering workshops. I was a teacher prior to stepping into the administrative role. I am currently finishing my Masters of Education with action research in a different area. And that sums it up. (Interview, November 9, 2004, p. 1)

**Our Questions and Protocol**

To illuminate literacy praxes, promote reflection, and inspire action, interview questions were 'Grand Tour' questions (Spradely, 1979) that invite a range of perspectives. This approach enabled teachers to describe their experiences in their own terms. For example, open-ended probes beginning with the words "tell me" to describe a "typical day" or "class" suggested to interviewees that a general or global response was expected. Additional prompts and cues allowed further depth and breadth to surface (Stringer, 1996, p. 67). Context-specific questions helped to guide, simplify, and contextualize the action research effort. To ensure objectivity, interviewers avoided conflict, when possible, with interviewees (McNiff et al., 1996; Stringer, 1996). Transcriptions, teacher-researcher's notes, participant observations, and student created artefacts provided further data. In addition, data were collected by tape recording discussion groups and general contextual notes were kept of school visits. All participants used written products (notes, memos, and daybooks) to communicate and document their thoughts. Written products detailed the classroom teachers' reflections

on their literacy praxes. These reflections included the formulation of ideas and changes in practice (Woolhouse, 2005). This inquiry revealed a pattern of internal growth and transformation that was recorded and documented. A summative final visit produced a thirty-minute videotape of each classroom and teacher to augment data collection. "Videos, however, do not reveal 'the facts' or 'the truth', they still provide only partial information…" (Stringer, 2004, p. 83). Indeed, the multiple sources of collected data captured multiple images and enhanced the validity of the inquiry (McTaggart, 1996), as we pieced together change, images, and views.

**Interviews**

Thirty-minute after-school interviews were essentially open and involved a coherent discussion of literacy praxes. A general framework of questions was used to ensure the "…deliberate establishment of an 'audit trail' of data…" (McTaggart, 1996, p. 13). In our second round of interviews during March of 2005, the questions listed in Appendix A were revisited. Generally, the openness of the interviews allowed extensive contextual data to be collected. Five interrelated contexts or situations were recognized: the classroom, personal, social, historical (teacher background), and political (King, 1988). As a result, greater sensitivity was achieved reflecting the uniqueness of each teacher's educational situation.

**Data Collection**

Ruth, the principal in our study, and teachers entered the Practical Action Research Project to improve student literacy and change praxes which ultimately enhanced literacy in classrooms. Informal and formal meetings of teachers occurred monthly (totaling 10) throughout the year at the school level. Weekly communication generally occurred between teachers and researchers. University researchers provided assistance, facilitated, and supported the process each term as requested by school personnel or as planned. The interviews/meetings (two per term) involved the recording of interviews, sharing of information, identification of learning resources including authorities on literacy development, and periodic feedback that was directed towards constructive and facilitative ends. The university researchers kept notes of meetings, and this data was shared via emailed transcripts and periodic conferences throughout the project. It was understood by all participants that the school-based teacher-researchers would continue tracking students' literacy development using school board approved qualitative and quantitative assessment methods including the DRA (Developmental Reading Assessment) (Pearson Education, 2003) and the PM Benchmark Books (Ontario Ministry of Education, 2003). Each teacher conducted action research in their classroom, which included collecting and analyzing baseline data, identifying an area of concern and a research question, planning and implementing action to address the question, and collecting data to illustrate the impact of the project. Teachers, as noted earlier, recorded their action research experience through written products, classroom-teaching activities, student product analysis, and classroom-based assessment and evaluation activities.

The university researcher collected data once per term and informally via weekly communication with participants (i.e., email, telephone). Ruth (the principal) was interviewed as the study progressed, and similar communication modes were used to remain in contact. The university researcher then began to analyze the data and produce interim analyses.

**Data Analysis**

Each of the three school-based educators was responsible for a different area in the school; each became a case study. The fourth case included the respective university-based facilitator cast in the role of action research participant, collaborator, supporter, associate, and consultant. Each case became a documented body of knowledge and, when possible, was triangulated using multiple sources of information. McNiff et al. (1996) explain:

> Because action research is case study research its usefulness is for others who can see its implications to their work. It is important that the effectiveness of action research is demonstrated in terms of an improvement in the quality of the lives of people whom it is supposed to be affecting. Case studies to show this are only just appearing in the public domain which shows the effectiveness of action research approaches in concrete terms. (p. 23)

In this action research, a systematic approach and careful consideration of literacy development were required to ensure that all participants' concerns were given due thoughtfulness. Indeed, as our action research process unfolded, it was clear that participants were committed to finding a better way to teach; they were self-motivated. With this in mind, the data reflected sincere reflection; participants were not motivated by certification, advancement, or economic reward. Moreover, the process was recognized and endorsed by the board and the school principal as a valuable professional development experience for participants.

As the data were transcribed, we began to skim the data set and reflect on what we had sensed during interviews, conversations, telephone calls, video, and emails. Further reflection brought to mind recurring key terms, concepts, and words that characterized our interactions. We used the word *term* to describe something that is unambiguous whereas the supporting words could be defined in several ways. The key terms (themes) were used to search through a transcript and highlight occurrences (see Figure 1). We could then assemble a frequency checklist that raised the profile of a key term to that of a theme. "Simply put, themes are recurring patterns, topics, viewpoints, emotions, concepts, events, and so on" (Bailey, 2007, p. 153), which informed and guided our research. We believed the conception of data analysis articulated by Sagor (1992) suited us best in this endeavour:

> Data analysis can be most simply described as a process of sifting, sorting, discarding, and cataloguing in an attempt to answer two basic questions: What are the important themes in this data? and (2) how much data support each of these themes? (p. 48)

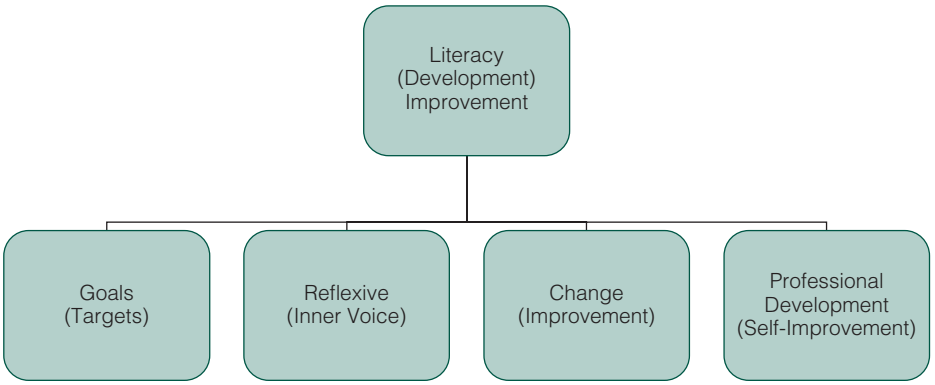



**Figure 1** Emergent themes as a result of action research praxis.

Key terms led to themes that were then used to label boxes in our diagram. Each key term was given a color as the data were skimmed, sifted, and sorted. A link to a key term often surfaced, and the data were highlighted in the corresponding color. In each box, coordinates were noted such as the date and page number. This way, if we were looking for data concerning "change," we could search the transcript to locate color and source. Data surfaced from one-to-one interviews, document analysis, casual visitations, and both informal and formal written and verbal communications. Located in a specific matrix box would be the source's location by date and page number so we could quickly locate the information. The summative video-tape was also used to capture a permanent record of summary evidence.

**Results and Findings**

During our inquiry the research landscape shifted and matured as we realized our goals. Findings were detailed in the themed sections. For instance, change was witnessed as goals were achieved, and these results encouraged participants to scrutinize and share practice insights recursively.

**Themes**

Within each case study, we sought the frequency of key terms and colored these accordingly. For instance, 'goals' were mentioned often and became a centerpiece of discussion. Other key terms included "change," "reflexivity," and "professional development." The evidence that follows demonstrates that we had many findings. These following excerpts illuminate our themes.

**Goals**

The university facilitators were targeting two elements. First, "we were looking for the development of action theory (new understanding) and action practice (application) as it relates to teacher development and student growth.... [Second], we hoped to realize a recursive cycles of theorization/reflection and application in order to professionalize practice" (Ryan, Journal September, 30, p. 2). In the autumn of 2004, interviews immediately revealed several themes. For example, during our first one-to-one interview, Pam addressed the need to seek goals in her classroom:

> I usually use the group approach, and it works well because they help each other. So they are in groups all day long. Well, I am hoping that through modeling of a peer and from what they see going on in my teaching that they will get into this act of literacy and know where they are heading with the literacy center. (Interview 1, October 13, 2004)

Pam wanted students to support each other and, through her modeling, hoped that other goals would be realized. Similarly during the first interview with the principal, two themes goals and growth–were noted:

> I really hope to cultivate a professional learning community through this process, and I think my interests and passions have to do with [a] number [of things], one is literacy and two, is bringing the kids to their full potential, to make them believe in themselves and making sure that we've got programs and [that] our school is supported to move that learner along. (Interview 1, November 9, 2004)

Clearly our administrator was hoping to 'cultivate' literacy in order to realize the 'full potential' of each student in these classes. These administrative goals were viewed as essential to our action research efforts since support and team growth can only broaden and deepen the results of our work. Ann, a second teacher, wanted to diagnose areas of need and then move to address these needs as necessary. For example, Ann detailed her situation this way, "I'm just trying to work with individual students to determine where their weaknesses lie in literacy and how I can give them personalized activities so that they can grow" (March 3, 2005, p. 2). Ann wanted to nurture and support literacy and felt she needed to define needs before moving forward with literacy efforts. Her approach was linked to her need for strategy and organization in her classroom. Early in our inquiry Ann openly suggested,

> I hope that I get to learn more about my strengths and weaknesses, areas where I can improve. But also maybe I'll find something that can help my students that I wouldn't have thought of before. I also hope to be able to collaborate with others and find out things that they've learned and so on. So the other team members of this project, I hope to learn from them. (Interview, October 13, 2004)

This need to improve and grow was common and united our participants. Our efforts to develop a community of practice "may be the single most important way to improve a school" (Sergiovanni, 2000, p. 139). Participants applauded our efforts to support their work, explaining that they had not experienced professional development like this before.

**Change**

Ruth, the principal, was also hoping to learn, grow, and change, and this complemented the professionalization that we hoped to facilitate through our project. As evident in the following excerpts, our situation was ideal given the position of our administrator.

> I want to be able to learn how to be a better administrator and I think it's critical that the [role of the] administrator of a school, it's twofold: you're the manager of the building, but you're also the instructional leader. And I think what I'm trying to do here is put the

> instructional leadership first because what happens in the day, you get so busy doing other things that your time is consumed by everything else. (Interview 1, March 12, 2005)
> And if we have to change our system, we have to change our system. It's not the kids that always have to change. We have to go and meet the child wherever they are. And I think that's been my whole passion, is bringing literacy to children, [bringing] children to literacy because you have to meet somewhere in the middle. (Interview 1, March 5, 2005)

Our efforts to reflect were not limited to self-analysis; we encouraged participants to share their own experiences in order to justify and construct images that could be understood by others. We know we learn from others; however, we also learn from ourselves by talking and interacting with others (Bruner, 1990). The process of reflecting with others augments our ability to change and shape our educational philosophy, which then impacts pupil growth. Ann struggled to deal with change within our research effort and to define her next steps:

> I just don't know where to go from here. I am continuing to collect and continuing to gather information and work towards their weaknesses to help strengthen them, but is there an end or should I just do what I can until the end of the year and then start again? How do I, through collecting data, like work samples and so on, is that enough? Where am I going next, that's kind of what it is. (Interview, March 3, 2005, p. 1)

This openness was key to our progress and professional development as we made public our concerns and inner voices. Ann continued to find her way and address student development by suggesting,

> I think they [students] are always changing to a point. Just the awareness that most students really have specialized needs in a variety of different areas has really impacted me this year. Like I have been aware of different levels with IEPs and so on, but to have students not on IEPs with such a vast array of strengths and weaknesses and trying to just, like, choose certain things to work on, and how do I teach to the group when there is all these individual weaknesses or strengths that I need to draw on. (Interview, March 3, 2005, p. 2)

This self-questioning within a community of learners (action researchers) is the most basic element required for continual, substantive school improvement (Mitchell & Sackney, 2001). Each participant voiced their endorsement of the research process, with Ruth noting that "We need to continue this journey as I need to have input every day."

**Reflexivity**

A reflexive educator cultivates an inner voice. In our study this emerged from discussions (both formal and informal). As we were attempting to improve literacy, we discussed how teaching is a time-deficit profession where most educators have high expectations. What exactly is reflexivity? Many have tried to define the term, and yet some of these definitions only lead us to questions. For instance, Nightingale and Cromby (1999) suggest, [R]eflexivity requires an awareness of the researcher's contribution to the construction of meanings throughout the research process, and an acknowledgment of the impossibility of remaining "outside of" one's subject matter while conducting research. Reflexivity then, urges us to explore the ways in which a researcher's

involvement with a particular study influences, acts upon and informs such research. (p. 228)

In other words, reflexivity is the condition of taking account of the personality and presence of the researcher within the investigation. May (1998) adds,

> The concepts of reflexivity may be a way of bringing qualitative methods to account for themselves in a way that goes some way to satisfy the demands of scientific method. This is generally a matter of questioning how the processes of research and analysis have an effect on research outcomes. This whole process of self-examination has become known as "reflexivity." (p. 22)

In our inquiry we noted several instances where both the inner voice and the researcher role surfaced. For instance, Ruth explained,

> I think I go about my own P.D. in a very public way, because I'm a very public learner.… Because I constantly question myself and I'm constantly looking for a better way to do things. When I'm sitting with kids and I'm listening to them read and I'm also looking at the text saying: Is this text appropriate for the child? And is this text too hard? Not just in the words, but in the way they've set it up on the page and the way they've got their quotation marks—could they have changed the text to make it a little easier for the children. (Interview 1, November 9, 2004)

This inner dialogue becomes buoyant and surfaces during conversation; that is the centerpiece of action research. Being reflexive is useful. When the thoughts are made public, growth and understanding become more visible than if a person simply archives his or her conversations, attitudes, and understandings. Another participant, Pam, explained,

> I don't feel I am alone out there with this isolated question that I have. But yet I have a whole different group of children than what maybe these teachers have been working with, so that helps me in looking at what is good for the students here. What else is helping me is the students. They are giving me back lots of feedback, telling me if we're on the right track. (Interview, March 3, 2005, p. 1)

This reflective evidence was required in order for our learning community to move forward from the comments, insights, and questions. Our strategic and purposeful discussions led to change that was an essential aspect of our dialogic learning. Participants summarized this aspect of our study by suggesting that "within a group the risk of speaking about what is within is somehow diminished" (Video-record 1, March 22, 2005).

**Professional Development**

Ruth suggested,

> [It's] [m]y number one motivation, because I think action research is the sole professional development that I've been exposed to that's really increased my level of professional understanding and performance. And I really want to make other people fall in love with it. I want them to see that they have the answer within them. And to look at their profession in a positive way and to look at their kids, at every child they see, not as a problem, but as an opportunity to learn. Because that's the way I view it. And if I can get other people on board, thinking that way, then we're going to have a powerful school. (Interview, October 14, 2004)

When leadership makes public a vision that the school will be a community where all are united by the need to learn and this need is made public, then we have ideal conditions for improvement of the school and individuals within this community (Barth, 2001). However, did the community retain a similar position with regard to learning? Ann put forward her stance regarding reflection:

> My personal growth or professional development is important. I feel that I am very self–reflective, and I know this is a great avenue to really focus on something specific in my teaching day or my teaching practice and improve upon it and then I can work on other areas, of course. (Interview, October 14, 2004, p. 2)
>
> The sense that this inquiry was an opportunity to meet with others and develop both self and the immediate learning community was present from the initial contact. Our work served to motivate, as Pam pointed out, being able to use this research, use this information and put it to good use in my classroom and in my practice. Like you said the other day, we do reflect, and it's just acting on the reflections, thinking about it, contemplating. Doing a few action-oriented items from my reflections, but I think this is really going to get me into it more. (Interview, October 13, 2004, p. 2).

Connecting, collaborating, and communicating are powerful variables that motivate and infuse action research with vigor. This empowerment leads to change and can be a source of democratic and dynamic energy (Woolhouse, 2005). At times my role was that of a critical friend responsible for new ideas, resources, feedback, and questioning. Pam continued to develop professionally and characterized her spirit later in our project by explaining,

> I get excited about everything, so let's see. I'm getting excited noticing that from the month of October to the month of February, tracking anecdotally the progress of my students, that there is a change in their attitudes towards math. I am, by asking them "Are you enjoying math? Why?", I am getting feedback from them telling [me] that yes they are enjoying it, and I would say out of 30 students, 28, it is always up there in the high 20s, are enjoying math, more than they ever did before. (Interview, March 3, 2005, p. 1)

Our efforts created a level of liveliness that otherwise may not have occurred. Our efforts were a means of improving problem-solving skills and impacting professional development vicariously through classroom-based changes. Increasing openness and confidence was another result.

## CONCLUSION

This action research effort energized participants and heightened professional identity via the embedded nature of our interactions (Altricter, 2005). Emergent themes such as change, professional development, and targeting enabled participants to improve praxis and literacy results as classroom scores increased as did everyday reports of success as recorded by teacher-researchers. For instance, Ann captured a summative view by suggesting,

> The fluency of their reading right now is exciting me because they are showing stronger comprehension, they're self checking a little better when they read to me orally, which

> means they are paying attention to what, they're taking in what they're reading and they know they have to go back. That's the most excited, how strong they are becoming, oral readers and comprehending what has been read. (Interview, March 3, 2005, p. 2)

We viewed the classroom level assessment outcomes together and concluded that self and peer evaluation results improved on average between 6% to 9% for each class while project completion increased from 82% to 100% overall. This fact alone was responsible for enhancing outcomes, and we believed it was critical. Another dividend included larger portfolios that were more complex in contrast to previous submissions. Classroom presentations and classroom tests showed enhanced results with all students moving into level three, which is the provincial standard, and a few more than last term realized level four. Reports of these outcomes were captured on video during concluding interviews and motivated classroom teachers to move forward. All participants believed that our project had boosted classroom energy, focus, and outcomes.

Participants were empowered as they collected their data and made research decisions while both the principal and university researchers were supportive and guided when necessary. Participants assessed students and deemed that improvements were made, and this was noted in many processes and products developed in the classroom. Our action research effort affirmed the professionalism of teaching and created an open dialogue that fostered progress.

## FUTURE PLANS

All participants indicated that they wanted to continue to improve and extend this research effort into the following school year. This would involve new students. The insights, skills, and growth achieved this year would support each participant's renewed efforts to realize continued improvement for students and participants.

## REFERENCES

Action Research International Journal (1997) Action research defined. *Action Research International Journal, 5* (2), p 32–34. Retrieved November 24, 2003 from http://www.triangle.co.uk/ear/.

Altrichter, H. (2005). The role of the professional community in action research. *Educational Action Research, 13* (1), 11–25.

Baily, C. A. (2007). *A guide to qualitative field research* (2nd ed.). Thousand Oaks, CA: Pine Forge Press.

Barth, R. (2001). *Learning by heart*. San Francisco: Jossey-Bass.

Bruner, J. (1990). *Acts of meaning*. Cambridge, MA: Harvard University Press.

Carr, W. & Kemmis, S. (1986). *Becoming critical: Education, knowledge and action research*. London: Falmer Press.

Connelly, F. M. & Clandinin, D. J. (1988). *Teachers as curriculum planners: Narratives of experience*. Toronto, ON: OISE Press.

Corey, S. M. (1949). Action research, fundamentalresearch and educational practices. *Teachers College Record, 50*, 509–514.

Deshler, D. & Ewert, M. (May 25, 1995). *Participatory action research: Traditions and major assumptions. PARnet* Retrieved October 15, 1999, from http://www.parnet.org/tools/tools.

Feldman, A. (1999). The role of conversation in collaborative action research. *Educational Action Research, 7*(1), pp. 125–144.

Flores, E. & Granger, S. (1996). The role of thecollaborator in action research. In S. E. Noffke & R. B. Stevenson (Eds.), *Educational action research: Becoming practically critical*. New York, NY: Teachers College Press.

Hopkins, D. (1993). *A teacher's guide to classroom research* (2nd ed.). Philadelphia, PA: Open University Press.

Kemmis, S. & McTaggart, R. (Eds.). (1988a). *The action research planner*. Victoria, Geelong: Deakin University Press.

King, N. R. (1988). Recontextualizing the curriculum. *Theory into Practice, 25*(1), 36–40.

McNiff, J. (1988). *Action research: Principles and practice*. Basingstoke: Macmillan Education.

McNiff, J., Lomax, P., & Whitehead, J. (1996). *You and your action research project*. New York, NY: Routledge.

McTaggart, R. (1996). Issues for participatory action researchers. In D. Nightingale & J. Cromby (Eds.), *Social constructionist psychology* (1999). Buckingham: Open University Press. Retrieved November 18, 2003, from http://www.psy.dmu.ac.uk/michael/qual_reflexi vity.htm.

Mitchell, C. & Sackney, L. (2001). Building capacity for a learning c ommunity. *Canadian Journal of Educational Administration and Policy* (CJEAP), 19.

Noffke, S. E. (1995). Action research and democratic schooling: Problematics and potentials. In S. E. Noffke & R. B. Stevenson (Eds.), *Educational action research: Becomingpractically critical*. New York, NY: Teachers College Press.

Noffke, S. E. (1997). Professional, personal, and political dimensions of action research. In M.Apple (Ed.), *Review of research in education*, 22, (pp. 305–343). Washington, D.C.: American Educational Research Association.

Ontario Ministry of Education. (2003). *Early reading strategy: The report of the expert panel on early reading in Ontario*. Toronto, ON: Author.

Pearson Education. (2003). *Developmental reading assessment, grades 4–8, second edition*. Don Mills, ON: Author.

Sagor, R. (1992). *How to conduct collaborative action research*. Alexandria, VA: Association for Supervision and Curriculum Development.

Schon, D. (1983). *The reflective practitioner*. London: Maurice Temple.

Sergiovanni, T. L. (2000). *The lifeword of leadership: Creating culture, community, andpersonal meaning in our schools*. San Francisco: Jossey-Bass.

Stenhouse, L. (1975). *An introduction to curriculum research and development*. London: Heinmann.

Stevenson, R. B., Noffke, S. E., Flores, E., & Granger, S. (1995). Teaching action research: A case study. In S. E. Noffke & R. B. Stevenson (Eds.), *Educational action research: Becoming practically critical*. New York, NY: Teachers College Press.

Stringer, E. T. (1996). *Action research: A handbook for practitioners*. Thousand Oaks, CA: Sage Publications, Inc.

Stringer, E. (2004). *Action research in education*. Upper Saddle River, NJ : Merrill/ Prentice Hall.

Woolhouse, M. (2005). You can't do it on yourown: gardening as an anthology for personallearning from collaborative action research group. *Educational Action Research, 13*(1), 27–41.

Zuber-Skerritt, O. (Ed.). (1996). *New directions in action research*. London, UK: Falmer Press.

**APPENDIX A**

Question 1: Your motivation in the first interview was——. Has your motivation changed; and what learning has occurred?
Question 2: For Principal–Tell us, how you are facilitating this project? Are there any obstacles?
Question 3: For Teachers–Describe where are you now with your Action Research project? What is helping you with this study? What appears to be a barrier to your progress?
Question 4: Tell us, have you refined or changed your question?
Question 5: Describe what you are feeling towards this study. Are you excited or not?
Question 6: Describe any confusion?
Question 7: Have your views of literacy changed? If so–how so?
Question 8: You were employing certain strategies at the previous interview. Have you been refining them? Using new ones?
Question 9: What data are you collecting or planning to collect?
Question 10: Any other comments?

## THINK ABOUT IT 18.7

After reading the sample action research article, answer the following questions:

1. What was the problem or focus of the research?
2. What was learned from reviewing the literature?
3. How was the research question framed?
4. What change was made or action taken?
5. What types of data were collected?
6. How were the data analyzed?
7. What interpretations were made?
8. Based on the concluding statements, what action do you think was proposed next?

**Answers**

1. To enhance literacy outcomes in elementary classrooms.
2. Self-reflection and action research as a group activity can improve instruction.
3. Can a group of teachers engaged in action research improve literacy?
4. Teachers and the principal participated in the Practical Action Research Project to improve student literacy and change praxes. This involved formal and informal meetings of teachers and collection and review of data combined with reflection and sharing.
5. Student work, observations, and interviews.
6. Primarily case study analysis using qualitative techniques.
7. Findings were reported in themes: goals, change, reflexivity, and professional development. Participants had a common need to improve and grow. All participated in self-questioning behavior. Improved problem-solving and classroom-based changes occurred. Student scores and performance increased.
8. The Practical Action Research Project would be continued and perhaps expanded to include others.

## SUMMARY

Action research in the field of education has recently experienced a resurgence. Focused on solving practical problems in everyday settings, action research does not aim to generalize. The three main characteristics of action research are that it is locally focused, conducted by the practitioner, and results in changes in practice. It most commonly uses qualitative research approaches.

The action research process is cyclical, involving reflection, planning, acting, and observing. In education, problems investigated using action research typically focus on student learning, curriculum, instructional strategies, and professional development. Triangulation of data is important in action research. The types of data collected can be categorized as experiencing, enquiring, or examining. Rigor and ethical treatment are important in action research as in other research.

Data analysis typically relies on qualitative coding processes and focuses on description and sense making. Interpretation can be enhanced through the use of visuals, such as concept maps. Once interpretation is completed, the most important step is developing an action plan and then taking action based on the findings.

## KEY CONCEPTS

action plan
action research
authentic student work
axial coding
brainstorming
classroom action research
coding
collaborative action research
concept mapping
credibility
critical action research
data analysis
data interpretation
description
enquiring
examining
experiencing
explanation
extension questions
field notes
grand tour questions
informed consent
journaling
literature review
Moses effect
open coding
parental consent
participatory action research
practical action research
practitioner research
problem analysis
professional development schools
prompt questions
reflection
rigor
school-based study groups
selective coding
sense making
teacher inquiry
teacher research
triangulation

## EXERCISES

1. What are the three key components in the definition of action research?
2. List at least five ways that action research may differ from other approaches to research.
3. Describe a problem or area of focus in education from your own experience that might be investigated using action research.
4. Draw a pictorial representation of the variables or elements you think contribute to the problem.
5. What are the three categories of data described? Provide at least two examples of data sources that would be in each category.
6. Name three different types of coding that might be used in action research.
7. Some authors in the field claim that only one criterion of validity is necessary in action research. What is that criterion?
8. How important is generalizability in action research?
9. Why is action research seeing a resurgence in the field?

**10.** Listed here are some actual action research problems that have been investigated and written about. How do these problems fit into the four categories discussed, using "S" for student learning, "C" for curriculum, "I" for instruction, or "T" for teacher professional development?

**a.** Mr. Daniels was worried about taking over a class midyear from a teacher whose style was much more laid back, whereas his style was much more structured.

**b.** Ms. Grace wondered what strategies she could use to make smoother transitions between classroom activities so as to reduce wasted time.

**c.** Mrs. Hall was concerned about the lack of reinforcement opportunities students in her French class had available once they went home at the end of the school day.

**d.** Mr. Robbins wondered whether WebCT (an online course development tool) could provide an improved alternative for students to share their reflections on student teaching.

**e.** Ms. Stevens wonders whether the new district-adopted character education curriculum is biased.

**f.** Principal Talbot is concerned that so few Hispanic/Latino parents attend school functions.

## ANSWERS

**1.** The research is situated in a local context and focused on a local issue. The research is conducted by and for the practitioner. The research results in an action or a change implemented by the practitioner in the context.

**2.** Responses will vary.

**3.** Responses will vary.

**4.** Responses will vary.

**5.** Experiencing, enquiring, and examining. Examples in each may vary. For experiencing, typical responses might be observation logs or field notes, journaling, and shadow studies. For enquiring, typical responses might be interviews, focus groups, and surveys. For examining, typical responses might be student work samples, records, and test scores.

**6.** Open, axial, and selective coding

**7.** Whether or not the solution developed from the study actually solves the problem

**8.** Generalizability is not the goal in action research. The primary goal is to solve a real problem in the practitioner's context and to do something to improve it.

**9.** Movements to professionalize teaching and calls for accountability and data-driven decision making at both the state and national levels.

**10.** a, T; b, T; c, S; d, I; e, C; f, S

## REFERENCES


Dana, N. F., & Yendol-Silva, D. (2003). *The reflective educator's guide to classroom research: Learning to teach and teaching to learn through practitioner inquiry*. Thousand Oaks, CA: Corwin Press.

Hendricks, C. (2009). *Improving schools through action research: A comprehensive guide for educators*. Upper Saddle River, NJ: Pearson.

Johnson, A. P. (2008). *A short guide to action research* (3rd ed.). Boston: Pearson.

Mertler, C. A. (2009). *Action research: Teachers as researchers in the classroom* (2nd ed.). Thousand Oaks, CA: Sage.

Mills, G. E. (2003). *Action research: A guide for the teacher researcher* (2nd ed.). Upper Saddle River, NJ: Merrill Prentice Hall.

Newmann, F. M., & Wehlage, G. G. (1993). Five standards of authentic instruction. *Educational Leadership,* 50, 8–12.

Sagor, R. (2000). *Guiding school improvement with action research*. Alexandria, VA: Association for Supervision and Curriculum Development.

Stringer, E. (2008). *Action research in education* (2nd ed.). Upper Saddle River, NJ: Pearson/Merrill Prentice Hall.

Thomas, R. M. (2005). *Teachers doing research: An introductory guidebook*. Boston: Pearson Education.

CHAPTER 19

# Mixed Methods Research*

That which works, works.

Aldous Huxley

## INSTRUCTIONAL OBJECTIVES

After studying this chapter, the student will be able to:

1. Define key terms in mixed methods research.
2. Recognize the philosophical bases of mixed methods research.
3. Understand the purposes for conducting mixed methods research.
4. Differentiate between different mixed methods designs.
5. Explain the reasons why a mixed method study might be conducted.
6. Understand the limitations of mixed methods research.

We previously discussed both quantitative and qualitative approaches to research. Now we discuss an expanding area of interest for researchers—combining quantitative and qualitative approaches in mixed methods research. Social scientists conduct mixed methods research primarily because they believe multiple approaches may provide better information to understand a particular phenomenon under investigation.

*This chapter was written with the assistance of Dr. Thomas Christ, College of Education, University of Hawaii at Manoa.

# DEFINING MIXED METHODS RESEARCH

Greene (2008) asked the question, "Is mixed methods social inquiry a distinctive methodology?" The answer today seems to be "yes." Mixed methods research combines quantitative and qualitative research methods in different ways, with each approach adding something to the understanding of the phenomenon. If mixing methods offers a better understanding of the research problem than a single method design, then it is worth considering. Practitioners, evaluators, policymakers, and funding agencies now recognize that combining multiple forms of data within a single study is not only legitimate but also, at times, preferable.

The popularity of mixed methods has grown exponentially during the past 30 years, as demonstrated by the increase in the number of published specialized books from a few in the early 1990s to dozens currently published internationally. Journals devoted to publishing mixed methods research, such as the *Journal of Mixed Methods Research* and the *International Journal of Multiple Research Approaches*, have added to the knowledge base. The National Science Foundation issued guidelines for mixed methods evaluations in 1997, followed in 1999 by the National Institute of Health. In 2006, the American Educational Research Association formed a special interest group devoted to mixed methods research with more than 200 members from the United States and abroad. The first International Mixed Methods Conference was conducted at Cambridge University in 2004. In 2007, Johnson, Onwuegbuzie, and Turner published an article that addressed the issue of defining mixed methods research.

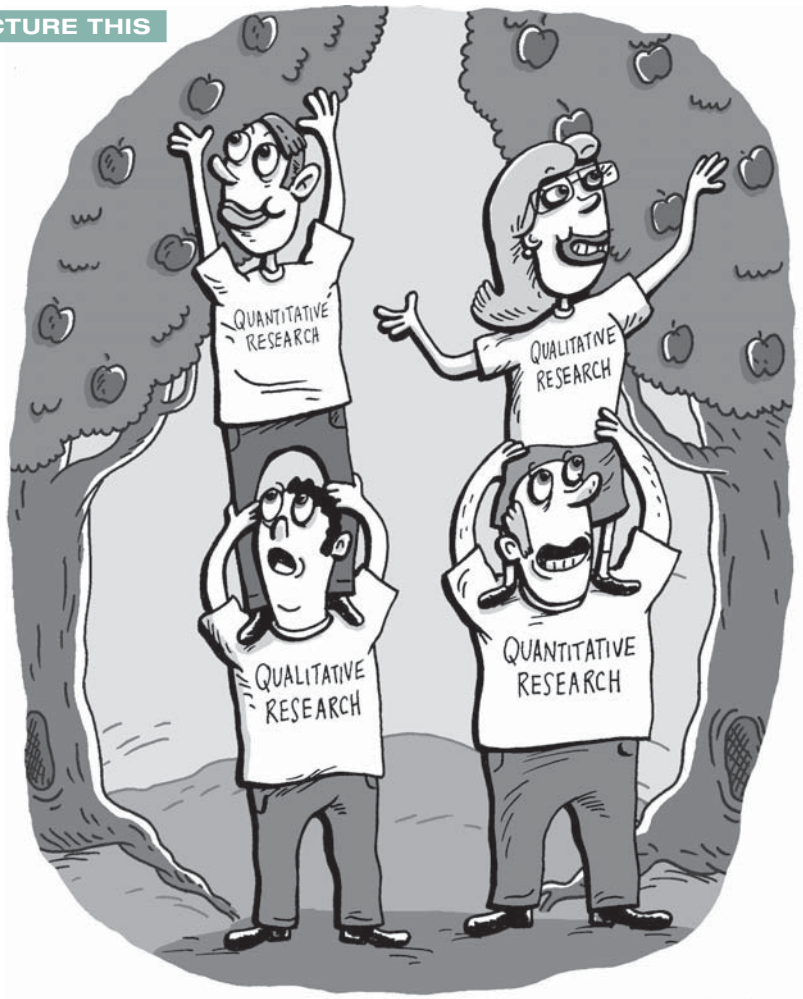
As with other research approaches, mixed methods research is not without controversy. There are concerns about the time and expertise necessary to combine quantitative and qualitative research within one study, as well as the resources and funding necessary to conduct larger scale, more complicated projects. Some argue that often the only way to properly conduct a mixed methods study is to create a team in which some members are skilled at conducting qualitative research and others quantitative research. This still does not alleviate the challenges of time or resources required.

## THE "THIRD" WAVE

To begin the discussion of mixed methods research, we must start with a consideration of **paradigms.** Most researchers are familiar with the "paradigm wars," in which paradigm "purists" argued that quantitative and qualitative methods were incompatible because these two paradigms are opposed in conception and practice. The mixed methods approach rejects the paradigm purists stance in favor of a **pragmatic approach.** Rather than choosing between methods considered in the past to be paradigmatically incompatible, pragmatism focuses on "what works" to answer the research question. Pragmatism shifts energies away from philosophical underpinnings and focuses on actions. This pragmatic mixed approach is considered the "third paradigm" or the "third wave" in research.

The goal of mixed methods research is not to replace qualitative or quantitative approaches but, rather, to combine both approaches in creative ways that utilize the strengths of each within a single study. By mixing methods in ways that minimize weaknesses or ensure that the weaknesses of one approach do not overlap significantly with the weaknesses of another, the study is strengthened.

**PICTURE THIS**




Joe Rocco

Table 19.1 summarizes the emphases in mixed method research, which can be compared with information from previous chapters on the distinctions between qualitative and quantitative research.

## CLASSIFYING MIXED METHODS

Defining mixed methods research is not an easy task because there are a multitude of design variations, paradigmatic combinations, and sources of quantitative or qualitative data. In the research field, there is continuing discussion regarding clarification of mixed methods research in relationship to monomethod research, multimethod research, and mixed model research.

In a **monomethod research** design, one method, either qualitative or quantitative, with corresponding data collection, analysis, and accompanying procedures, is used to answer the research question. **Multimethod research** employs different types of data collecting methods—for example, both survey

**Table 19.1** Emphases in Mixed Method Research

| | |
|---|---|
| Reasoning | Both deductive and inductive reasoning are used. |
| View of human behavior | Human behavior is considered somewhat predictable. |
| Focus | The focus is multiple, not restricted, but not entirely open. |
| Nature of observation | Behavior is studied in controlled and uncontrolled environments; multiple contexts or conditions. |
| Nature of reality | A commonsense pragmatic view of the world is typical. |
| Form of data | Multiple forms of data are used, both quantitative and qualitative. |
| Nature of variables | Variables can be a mixture of words, numbers, images, etc. |
| Data analysis | Both quantitative and qualitative approaches to data analysis are used. |
| Results | Corroborated findings may or may not generalize to other situations. |
| Form of reporting | Eclectic methods of reporting are used. |

and archival data. Multimethod research occurs when the research questions are investigated by using two different data collection procedures (e.g., observations and focus groups) or by combining two research methods (critical theory, grounded theory, or case study) from the same research tradition (qualitative or quantitative).

Tashakkori and Teddlie (2003) discuss **mixed model research,** in which mixing of qualitative and quantitative approaches occurs in all stages of a study (formulation of the research questions, data collection procedures and research method, interpretation of the results, and final inferences) or across stages of a study (e.g., qualitative questions and quantitative data).

**Mixed methods research** combines qualitative and quantitative approaches in a single or multiphased study. The mixing process may occur in any or all phases of the research, including the methodology, the logistics of the design, the specified methods to be conducted, the procedures or data collection, and the analysis. However, mixed methods research is more than quantitative and qualitative research mixed: It incorporates and embraces blends of paradigms, philosophical assumptions, and theoretical perspectives directly driven by the purpose of the study and the intended audience. As Yin (2006) indicates, unless there is a purposeful and logical blending of the qualitative and quantitative methods within the design, the result is instead two distinctly separate studies.

## PURPOSES FOR CONDUCTING MIXED METHODS RESEARCH

Authors in the field have identified various reasons or functions for conducting mixed method research and these differing purposes impact the designs of studies. The concept of purpose gets to the heart of why a mixed methods approach is chosen: whether to seek corroboration of findings, to elaborate or clarify findings, to further develop interpretations, to investigate contradictions, or to expand the breadth or depth of a study. The five general purposes discussed here were identified by Greene, Caracelli, and Graham (1989) and have been reiterated by others.

**Triangulation** seeks to examine the convergence of evidence from different methods that study the same phenomenon or to corroborate findings from one method by examining the findings using a different method. The value of triangulation is

that it allows the researcher the opportunity to examine whether findings converge, are inconsistent, or contradict. The purpose of triangulation is to collect, analyze, and merge results to better understand a research problem. Triangulation allows researchers to collect separate forms of data at separate times or simultaneously. A researcher might conduct focus groups and also gather quantitative ratings using a survey form to assess program participants' perceptions. Or, for example, teacher survey responses might indicate that particular classroom practices are routinely used. Classroom observations might provide information that triangulates to corroborate (or contradict) the reported behaviors.

**Complementarity** involves seeking elaboration, illustration, enhancement, or clarification of findings from one method using results from the other. Different approaches are used to measure different facets of a single phenomenon. For example, one might conduct qualitative interviews with parents to determine influences on their perceptions of the quality of education their child is receiving in the school combined with a quantitative questionnaire to determine how they perceive their child's school compares in quality ratings with other schools.

**Development** uses the results from one approach to develop or inform the other approach. The study can begin with qualitative data analyzed in the first phase as a way to establish constructs to be measured quantitatively in the second phase. Alternatively, quantitative survey results from an initial phase of a study might be used for purposeful selection of participants for a second qualitative phase.

**Initiation** is aimed at discovering paradoxes or contradictions in findings that might lead to reframing a theory. This approach is used to add breadth and depth to the inquiry. The focus is on intentionally analyzing quantitative and qualitative data that are inconsistent or contradictory. Such a focus might lead to fresh or new insights into a phenomenon. For example, finding through qualitative data that parents held negative perceptions of their child's school experiences while they simultaneously rated the school as "good" on a quantitative survey could initiate a mixed methods design to find an explanation.

**Expansion** is focused on expanding the breadth and range of the inquiry by using different methods for different components of the study. For example, a researcher might be interested in examining the effectiveness of a professional development program in raising student achievement. Qualitative observations might be used to determine whether and how well the teachers are using the professional development approaches in the classroom, which is sometimes referred to as implementation fidelity. Quantitative test scores could be used as a quantitative measure of program effectiveness.

## MIXED METHODS DESIGNS

In considering the design of a mixed methods study, the researcher must look at a number of factors and ask him- or herself a number of questions: Will the study involve one or more methods of data collection? Will the study have one phase, two phases, or multiple phases? Will the data collection occur sequentially or concurrently? Will data conversion occur? Will the study be mixed in the initial stages only, across stages, in the end stages only, or in some other combination? Does the qualitative or quantitative component have priority or are they of equal

importance? Teddlie and Tashakkori (2006) discuss several key areas that the researcher should clearly articulate:

1. The number and type of data collection approaches that will be used
2. The number of phases or strands in the study
3. The type of implementation process to be used
4. The level or stage of integration of the approaches
5. The priority of the methodological approaches

How the researcher answers these questions determines the most appropriate mixed methods design to be used. Six mixed methods designs are described here.

In **concurrent designs,** both qualitative and quantitative data are collected separately but at approximately the same time. Analyses are conducted separately and interpretations are made for each set of data. Results from one set of data are not used to build on during analysis. Following separate collection, analysis, and interpretation phases, the researcher integrates the inferences. **Concurrent triangulation** occurs when quantitative and qualitative data are collected and analyzed separately but at the same time, with the findings converging in the conclusions in order to answer an overarching research question. The challenge associated with using a concurrent design is that it requires a significant level of expertise on the part of the researcher to analyze both the quantitative and the qualitative data.

**Parallel designs** are those in which data are collected and analyzed separately, similar to concurrent designs. However, in concurrent designs, inferences are made in a more integrated manner, whereas in parallel designs each data set leads to its own set of inferences. Indeed, sometimes the results may be reported in two separate write-ups in the same report. Some authors have referred to this design as a **quasi-mixed method design.**

In **sequential designs,** data that are collected and examined in one stage inform the data collected in the next phase. Data analysis begins before all data are collected and the analysis may influence choices made in conducting the next phase of the study. Each separate phase may shape the conceptual and methodological approaches used in following phases. Different forms of data are collected in sequence at different phases in a study. For example, a quantitative survey collected and analyzed in the first phase may inform the second qualitative phase of the study, or qualitative observations conducted in the first phase may inform development of a quantitative survey in the second phase. The sequential ordering provides information necessary to conduct a more thorough study.

**Fully mixed designs** or fully integrated designs involve mixing of the qualitative and quantitative approaches in an interactive way throughout the study. At each stage, one approach may influence the implementation of the other. There are multiple points of integration throughout the study, from data collection to data analysis and interpretation and inferences.

**Conversion designs** involve transforming data, which means that data collected in one form (e.g., numbers or text) are converted to a different form and then analyzed. Qualitative data may be converted to quantitative data or vice versa. Data are analyzed from both perspectives and inferences are made based

on both sets of analyses. For example, qualitative data might be **quantitized** by counting the number of times a particular word is used or the number of times a particular theme is identified. Quantitative data from a survey might be used to create a written profile of a group. **Data transformation** is a term used in mixed methods research to describe when a researcher begins with qualitative data, such as interviews or secondary source documents, and through analysis transforms the words into numerical values for comparative or statistical analysis. The transformed data are then compared with data from a comparable method (quantified qualitative data compared with statistical results from a separate phase, or qualitized descriptive statistics from a survey compared to coded themes that emerge from analysis of interviews). Conversion designs require that at least one of the data sources be transformed into another type of data. Some researchers argue that it is inappropriate to quantify qualitative data because it considerably reduces the value of the data as a tool for understanding a phenomenon.

In an **embedded design,** one form of data supports a second form of data within a single study. For example, if the purpose for the research project is to inform administration or change policy, quantitative and qualitative data may be required to convince those with the power to make changes that the results are credible. The rationale for an embedded design is that a single data set is not sufficient to answer different questions, and each type of question requires different types of data. The most common form of embedded mixed methods research is when quantitative data are embedded within a qualitative case study and the quantitative data are supportive of the major qualitative findings. Alternatively, qualitative data can be supportive of statistical results by addressing questions that are unanswerable using experimental or correlation research. The hallmark of an embedded design is that the different data sets are mixed at the design level rather than in the conclusions so that one of the data sources is relegated to playing a supplemental role to the prominent data source. One challenge that must be dealt with when conducting embedded research is that the purpose of the study may not be clear to the participants or research team members involved in the project.

## NOTATION SYSTEM

The most widely accepted **notation system** used in mixed methods designs, first introduced by Morse (1991), uses a plus sign (+) to indicate that the data collection and analysis of methods occur at the same time. An arrow (→) indicates that data collection and analysis occur in sequence. Morse also indicates that the weight or importance of the methods within the study should be denoted by using uppercase letters for prominence and lowercase letters to indicate less dominant methods. There are no specific rules that determine appropriate proportions of qualitative and quantitative research in a mixed methods study. Some researchers use parentheses to indicate methods that are embedded within other methods. The notation system rules are shown here

Weighting priority
QUAL + QUAN (both are equally important)
QUAN + qual (quantitative approach is dominant)
QUAL + quan (qualitative approach is dominant)

Sequence
QUAN → Qual (quantitative collection or analysis occurs first followed by qualitative collection/analysis)
QUAL → Quan (qualitative collection or analysis occurs first followed by quantitative collection/analysis)

Embedded
QUAL (quan)
QUAN (qual)

Visually representing a mixed methods design is a useful tool for helping the researcher to determine and show logic in the relationship between the components. The visual interactive model allows for a clear representation of how the purpose, conceptual framework, methodology, methods, and issues including validity/credibility interact with the central overarching research question.

## DATA ANALYSIS IN MIXED METHODS RESEARCH

In general, because mixed methods research is a combining of qualitative and quantitative methods, the basic analysis techniques used in those approaches hold true in mixed methods research as well. The primary difference is in integrating the data. The basic analysis techniques described in previous chapters would be used to analyze the various types of data collected in a mixed methods study.

However, Onwuegbuzie and Teddlie (2003) present a seven-stage conceptualization of **mixed methods data analysis** that is useful to consider, particularly in conversion designs. Stage 1, **data reduction,** involves analyzing the qualitative data via theme analysis or thematic coding while also analyzing the quantitative data via descriptive statistics, factor analysis, etc. Stage 2, **data display,** involves using tables and graphs to display the quantitative data and using other forms, such as matrices, rubrics, and lists, to describe the qualitative data. Stage 3 is **data transformation,** in which qualitative data (words) may be transformed into quantitative data (numbers) and/or quantitative data (numbers) may be transformed into qualitative data (narrative). Stage 4 is **data correlation,** which involves comparing the data from the different analyses (quantitizing and qualitizing compared to the originals). The authors use the term *correlation* in a different sense than typically considered in quantitative research. Here, the researcher is comparing, for example, the original qualitative data and the quantitized qualitative data to determine whether the two sets seem to reflect similar findings. Stage 5 is **data consolidation,** in which both sets of data are combined to create a new set of data or variables. Stage 6, **data comparison,** involves comparing data from the qualitative and quantitative data sources. Finally, stage 7 is **data integration,** in which the data and interpretations are integrated into either a coherent whole or reported in two separate sets (qualitative and quantitative) of coherent wholes. Figure 19.1 and Table 19.2 present qualitative data (McKenzie, 2008) that were quantitized in order to show how the data transformation stage might work.

None of the students or staff members were asked questions pertaining to financial aid, however, the topic came up serveral times. Those who were interviewed commented on how the SSS program helped them navigate the financial maze, teaching them the language of financial aid and providing them with information, support and encouragement needed to successfully get through college.

*I was motivated and ready to dive into the college experience; however, I had very limited financial resources. I did not know how to apply for financial aid, and I did not know if I would qualify for assistance. I saw an advertisement for the Student Support Services program, which stated that they provided financial aid assistance. Therefore, I applied to the program and was accepted. A representative helped me to complete the Free Application for Federal Student Aid (FAFSA). We later discovered that because I was classified as a low-income student that I qualified for a full Pell Grant, a state award, and a few other awards. I also qualified for a supplemental award that was disbursed through the Student Support Services program. I received enough money to pay all of my educational expenses. They were able to help me to interpret my student aid report and assist me with the application process on an annual basis. The great thing about the program was that it helped me to cope with the college environment, and they helped me obtain the financial resources needed in order to stay in school. (Elias Brooks)*

*I had been taking college courses for quite awhile. Consequently, I had depleted much of my savings and was in serious need of financial assistance. An SSS counselor told me that they would review my application to determine if I qualified for a supplemental award. It was determined that I was eligible. The extra funds helped me to purchase books, pay my tuition, and provide me with bus fare to and from school. College students have a lot to worry about, but the program really helped students to worry less about finances. I received reminders about financial aid application due dates, scholarships, and workshops. They basically eliminated all of the excuses (money issues) that I had for dropping out of college. The closer I came to graduating, the more information they provided about graduate school fellowships and scholarships. Many students joined the program for various reasons. Some people needed the academic assistance (tutoring, study skills, test taking, etc.), others needed the guidance offered by SSS staff. I joined because of the opportunity to receive financial assistance, during my undergraduate and graduate years. (Smokey Mann)*

*They took my hand, as if I were a little child, and helped me to achieve my financial and academic pursuits. The counselors informed me that they provided financial aid assistance and help with applying for the financial aid. They asked me to bring in my tax returns, W2s, and other important financial aid paperwork. They helped me to obtain enough financial aid to pay for all of my classes, books and fees, and housing costs. (Frank Speaks)*

*Completing the FAFSA was a lot like filing your tax returns. It was a process that requied numerous sets of eyes. I would complete the form and take it to my SSS advisor prior to going to the financial aid office. They would catch my errors and help keep my application from being set on the side because of such. My SSS advisor has close working relationship with the people in financial aid; therefore, he made sure that my application was processed on time and that I received enough free financial aid to remain in school and eventually graduate. However, they didn't just stop there. Student Support Service representatives counseled me and other students about loan debt and loan default. The SSS counselors and/or workship presenters would encourage us to wait to borrow money once we were in graduate school. (Elias Brooks).*

**Figure 19.1** Quotes from a Qualitative Study on College Graduates' Perspectives on Student Support Services Programs

*Source:* Quotes reproduced with permission from McKenzie, J. (2008). *The role of student support services programs in motivating students to persist in postsecondary institutions*. Unpublished doctoral dissertation, Northern Illinois University, DeKalb.

**Table 19.2** Quantitized Data: Student Support Services Student Perceptions Related to Financial Aid

| Comments About | Student A | Student B | Student C | Total |
|---|---|---|---|---|
| Limited resources available for college | X | X | | 2 |
| Limited knowledge about financial aid | X | | | 1 |
| Need for assistance with completing forms and getting through the process | X | X | X | 3 |
| Financial aid helped stay in college/achieve dream | X | X | X | 3 |
| Importance of personal support provided by staff | X | X | X | 3 |

## RIGOR IN MIXED DESIGNS

Teddlie and Tashakkori (2006) suggest two ways to examining the **rigor** of mixed methods studies. One approach is to consider **design quality,** or the standards used to evaluate the **methodological rigor** of the study. Design quality can be affected by within-design consistency (Is the design consistent with the questions?), design suitability (Are the methods appropriate for the questions asked?), design fidelity (Are the procedures implemented with quality?), and analytic adequacy (Are the data techniques used appropriately?).

The second approach is to examine **interpretive rigor.** Is there interpretive agreement or interpretive distinctiveness across approaches? Is there interpretive consistency, where inferences follow findings in terms of type, intensity, and scope? Is there theoretical consistency such that inferences are consistent with known theories? Is there integrative efficacy such that meta-inferences adequately incorporate inferences that stem from both qualitative and quantitative phases of the study?

Teddlie and Tashakkori (2006) argue that both forms of rigor should be combined in an integrative model of quality. Design quality focuses on standards used to evaluate methodological rigor, whereas interpretive rigor uses standards to evaluate the validity of the conclusions.

## STRENGTHS AND WEAKNESSES OF MIXED METHODS RESEARCH

As with any research approach, there are both strengths and weaknesses in conducting mixed methods research. Johnson and Onwuegbuzie (2004, p. 21) provide a good chart that describes the strengths and weaknesses of mixed methods research. On the positive side, the use of words, narrative, and images can be used to add meaning to numbers, and using numbers can add precision to the qualitative data. Mixed methods research can take advantage of the combined strengths of qualitative and quantitative approaches and can use the strengths of one method to overcome the weaknesses of another. A broader range of research questions can be examined because the researcher is not confined to a single method. Mixed methods research can provide stronger evidence for a conclusion through corroboration of findings. The researcher may have insights that could have been missed with only a single method. The combination may produce more complete understandings of phenomenon or more complete knowledge to inform theory or practice.

However, there are also a number of weaknesses. It is difficult for a single researcher to carry out both quantitative and qualitative research. It is difficult to have equal skill sets in both methods, and especially if data are collected simultaneously, there is a great time commitment. The researcher must be able to understand the complexities of both approaches so as to make wise decisions about how they can appropriately be mixed. Conducting a mixed methods study is likely to be more expensive than using a single approach. Quantitizing and qualitizing data can have its own problems. Also, interpreting conflicting results may be difficult. Additional resources for those interested in learning more about mixed methods research include Creswell and Plano Clark (2007), Greene (2007), Mertens (2004), and Tashakkori and Teddlie (1998, 2003).

### THINK ABOUT IT 19.1

Think about a research project or topic you might undertake. List at least four research questions you might ask. For each question, identify a method for answering the question and indicate the strengths and weaknesses of that approach. Now determine how you might combine at least two of the approaches to strengthen the study while minimizing the weaknesses of the different approaches.

## AN EXAMPLE: MIXED METHODS SEQUENTIAL DESIGN

**Purpose of the Study**

Funding is one variable that has an effect upon the quantity of disability support services, but quantity is not necessarily synonymous with quality. Determining other variables that improve services is critical in times of declining funds. The intent of this longitudinal mixed methods study was to examine what constitutes exemplary postsecondary supports in North America and how one institution met the challenges of a significant budget reduction. Analysis of quantitative and qualitative data generated from 1999 to 2005 provided the opportunity to better understand how services could be modified as a way of maintaining a level of support necessary for student success.

This longitudinal exploratory mixed methods study of exemplary disability support services was conducted in three distinct phases. The first phase of research began in 1999 with a national survey of 1500 disability support coordinators repeated in 2001. Phase I was framed using a postpositivist philosophical conception where the survey was analyzed to determine if significant changes occurred over time. Exploratory factor analysis revealed trends and constructs, but many questions that emerged as a result of the analysis could not be answered using traditional quantitative research methods. The survey findings were helpful in two ways. First, the results provided information necessary for purposeful selection of three exemplary postsecondary institutions, and second, the results guided the construction of key semistructured interview questions used to collect data for the qualitative cross case analysis at the three sites.

Phase II of the research study was influenced by constructivist and critical paradigms and consisted of a qualitative cross case analysis. Three sites were purposefully selected using

**Exhibit 19.1** Provides an example of a mixed methods sequential design.

criterion that included information from the national survey presented to experts in the field of disability studies in order to solicit nominations of applicable exemplary sites. Once the sites were selected, information from the literature and the survey findings were used to generate a set of semistructured interview questions that reflected the intent of the research questions. Interviews were conducted over several months, transcribed, and analyzed for each case individually and then together in the cross case analysis. Open coding technique as first described by Glaser (1978) was used to develop the one set of "nodes" used in the three cases. This first step allowed the opportunity to locate and define core categories directly from the data. Next, core categories for each case were analyzed to determine "conceptual connectors," which highlighted how categories are related to their properties (Glaser, 1992, p. 38). This process followed the constant comparison format first described by Lincoln and Guba (1985) which uses a form of inductive logic to develop emerging themes and categories from a mass of narrative data. Modified version of Glaser's comparative coding technique (1992) helped guide the final cross case analysis of the three sites. This modified version focused primarily upon the theoretical constructs, not earlier open codes, to generate overarching theoretical conclusions used to highlight similarities and anomalies in services between sites.

Phase III of this analysis followed a pragmatic philosophical approach to research, studying a critical, single site faced with a significant budget reduction. The third phase highlights the action orientation of pragmatism embracing the value of methodological flexibility in a longitudinal study as the follow-up was warranted when funding at the site was drastically cut. Thus, if postpositivism heavily influenced this phase, the research questions should have been predetermined (Yin, 2003) as is common in typical quasi and true experimental designs. In essence, the third phase embracing a postpositivist paradigm would have been treated separately. This highlights how pragmatism allows for flexibility where researchers are not forced to conform to one methodological paradigm or predetermined overarching research questions. Compelling circumstances in the form of a significant budget reduction prompted action and a change to the study design. By adding a longitudinal phase and modifying two of the four research questions, the opportunity to explore how the organization offset damaging effects of decreased funds strengthened the entire study and was much more meaningful for the intended audience of policymakers and disability support coordinators. Thus, circumstances of the condition under examination and flexibility in analysis allowed the study to expand and develop the critical theme of leadership that emerged during the cross case analysis. By extending and reformulating the study of one site over time, leadership qualities were further examined from four different levels within the organization (coordinator, supervisor, support staff, and students), adding credibility and depth to the results (Patton, 1999).

**Mixed Methods Design Implications**

The sequential design provided the opportunity to determine how leadership characteristics and staff cohesion had a direct effect upon the provision of disability support services. This finding would not have emerged if the study was limited to survey research or concurrent collection of quantitative or qualitative data. The sequential multilevel design in the final phase of analysis helped validate how the coordinator's positive tone had influence upon the entire operation of the support service office.

Declining financial support was another critical topic that could only be studied through a longitudinal design. Concurrent data collection techniques cannot reveal trends or be used to explore hypothesis about change. Through a longitudinal examination, surprising results were revealed. Not all aspects of the budget reduction were negative as initially hypothesized. Staff indicated that many of the services were made more efficient and improved despite the loss of personnel. The budget reduction did require significant modifications to the format

**Exhibit 19.1** *Continued*

and content of services, but as the modifications were purposeful and executed with care, the quality of services did not degrade.

In summary, change occurred as a result of a significant budget reduction. Some changes were devastating to highly coveted and valuable supports while other changes resulted in unforeseen benefits including staff cohesion and increased efficiency. Much of what was learned in this study could not have been revealed unless the quantitative and qualitative phases were exploratory and sequential. By utilizing mixed methods and a longitudinal design in a single study, many of the critical themes that emerged would not have been evident. This study highlights how a pragmatic approach to research is critical. In this case, the research questions, not ideologies or the researchers' preference for quantitative or qualitative analysis, drove the study. Both methods served their purposes. The quantitative survey phase helped establish a nationally representative description of the types of services provided to students with disabilities in postsecondary education. Exploratory factor analysis also established several constructs that were useful throughout all phases of the study while revealing parameters used in the purposeful selection of exemplary sites for further analysis. The qualitative cross case analysis revealed similarities and anomalies in each of the three purposefully selected exemplary sites. This phase of the analysis was particularly useful for gaining a better understanding of how policy is related to funding and how leadership qualities influence staff cohesion and effectiveness. Finally, the longitudinal phase was the only way to reveal how services changed over time. In particular, the sequential collection of data allowed the analysis of a particularly relevant subject, reactions to a reduction in funding. This sequential phase, which emerged as a result of a legislative mandate, appears to be most meaningful for those directing and supervising support service as they face the challenge of providing more students mandated services with reduced financial support.

Although the cross case analysis was particularly helpful in describing similarities and anomalies at exemplary sites, the concurrent collection of data could not reveal how services and supports evolved over time. As example, the cross case analysis indicated that many of the services provided were quite similar despite completely different locations, policy, and legislation at the three exemplary sites, but funding did have an effect upon the quantity of supports made available to the students. The final phase revealed just what happens when a budget reduction occurs. Contrary to expectations, when changes were executed by dedicated individuals following a well thought out plan, the services did not degrade as initially expected.

The final analysis in this study examined the findings from all three phases collectively. Several themes were evident when the results were merged. One example is that exemplary services can be attributed to distinct goals and drive established by dedicated coordinators who exhibited an extraordinary degree of motivation to provide the best possible service for students. These leadership qualities did not stop with the coordinators as they extended directly to the supervisors, the specialists, and the support staff. Although competency was expected in this study that focused upon exemplary support services, most revealing was that leadership qualities appeared infectious in the organizations crossing boundaries and extending all the way to the students.

**Exhibit 19.1** *Continued*

*Source:* Christ, T. (2007). A recursive approach to mixed methods research in a longitudinal study of postsecondary education disability support services. *Journal of Mixed Methods Research, 1*(3), 226–241.

Figure 19.2 (Christ, 2008) provides a visual representation of his longitudinal mixed methods design of an exploratory three-phase study conducted to determine what constitutes exemplary disability support services in North America and how funding affects those services.

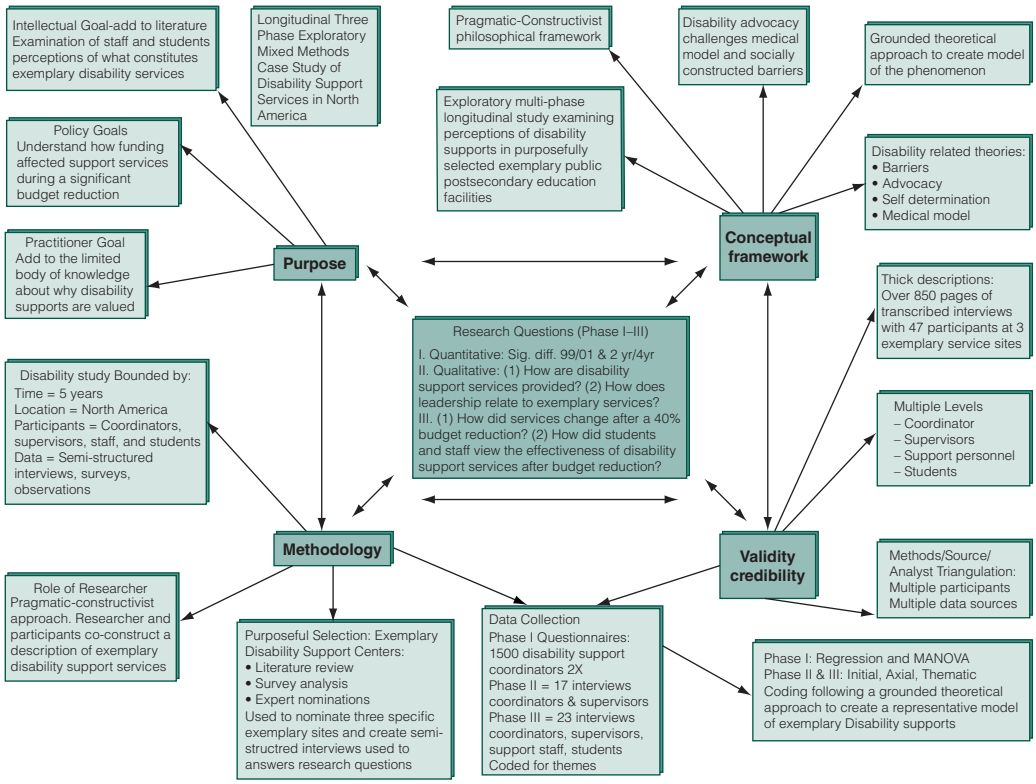



**Figure 19.2** Exploratory Longitudinal Mixed Methods Analysis of Disability Support Services

*Source:* Christ, T. (2008, April). *Mixed methods diagram: Recursive grounded technique in longitudinal multistage mixed methods study of disability supports: Emerging research questions*. Paper presented at the 2008 annual meeting of the American Educational Research Association, New York.

## SUMMARY

Mixed methods research is gaining recognition as a pragmatic approach to investigating phenomena that takes advantage of the strengths of both quantitative and qualitative research and seeks to minimize the weaknesses of each. Proponents assert that by using mixed methods the researcher can acquire a better understanding than he or she can using a single method. Mixed methods research requires combining qualitative and quantitative approaches, but that combination can take a variety of forms. The sequencing and the weighting of the quantitative and qualitative approaches differ in the various mixed methods designs. Integration of data interpretations can be difficult.

## KEY CONCEPTS

complementarity
concurrent designs
concurrent triangulation
conversion designs
data comparison
data consolidation
data correlation
data display
data integration
data reduction
data transformation
design quality
development
embedded designs
expansion
fully mixed designs
fully mixed methods
initiation
interpretive rigor
methodological rigor
mixed methods data analysis
mixed methods research
mixed model research
monomethod research

multimethod research
notation system
parallel designs
paradigm
pragmatic approach
quantitization
quasi-mixed method design
rigor
sequential designs
triangulation

## EXERCISES

1. When does mixing of methods occur in a mixed methods study?
2. In what order does the mixing occur?
3. At what level does the mixing occur?
4. In what proportions should the mixing occur?
5. Does the type of data gathered dictate the type of analysis done in the mixed study?
6. For each of the following descriptions, indicate the type of mixed method design used:
   a. A quantitative review of elementary school discipline records indicates that males are significantly more likely to be sent to the principal's office for behavior problems. The researcher decides to observe classrooms to describe the types of behaviors that result in being sent to the office. Based on qualitative observations, the researcher believes that teachers' reactions to the same behavior exhibited by males and females tend to be different. A quantitative measure of gender bias response is then given to teachers.
   b. A researcher conducts qualitative interviews with teachers to assess how well they are implementing a new instructional strategy that the school has adopted. During the same week, another researcher who is part of the team conducts random observational assessments using a quantitative tool to determine the frequency of use of the strategy. Both sets of data are used to answer the research question about teacher changes in instruction.
7. Match the purpose of the study to the appropriate term:
   a. In an attempt to develop an understanding of the influence of early experiences on the development of leadership styles, a researcher uses results from a measure of leadership style to purposely select five principals to interview in-depth a bout their experiences growing up and in their early careers.
   b. A researcher examines an apparent contradiction between new teacher self-reports of high personal use of technology and low levels of technology use in teaching by combining collection of survey data and analysis of lesson plans with focus group interviews and observations.
   c. A researcher attempts to corroborate teachers' reports of integrating technology into courses by examining lesson plans, conducting classroom observations, and interviewing students.
8. List two strengths and two weaknesses of using mixed methods research.

## ANSWERS

1. Mixing may occur at any point from the purpose statement to the data collection, data analysis, interpretation, and integration of findings.
2. It can be done sequentially or interactively using one set of information to inform decisions about the other or it can be done simultaneously or in parallel where information is gathered and analyzed at the same time.
3. Mixing can occur within or across levels that include individuals, groups, or larger organizations.
4. Qualitative and quantitative methods can be equal or one or the other may be dominant. There are no set rules concerning proportionality.
5. No, qualitative data may be quantitized (turned into numbers) or quantitative data may be qualitized (turned into text).
6. a. Sequential design
   b. Concurrent design
7. a. Development
   b. Initiation
   c. Triangulation
8. Answers will vary.

## REFERENCES


Christ, T. (2007). A recursive approach to mixed methods research in a longitudinal study of postsecondary education disability support services. *Journal of Mixed Methods Research, 1*(3), 226–241.

Christ, T. (2008, April). *Mixed methods diagram: Recursive grounded technique in longitudinal multistage mixed methods study of disability supports: Emerging research questions*. Paper presented at the 2008 annual meeting of the American Educational Research Association, New York.

Creswell, J. (2008). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*. Upper Saddle River, NJ: Prentice Hall.

Creswell, J. & Plano Clark, V. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.

Glaser, B. (1978). *Theoretical sensitivity: Advances in the methodology of grounded theory*. Mill Valley, CA: Sociology Press.

Glaser, B. (1992). *The basics of grounded theory analysis: Emergence vs. forcing*. Mill Valley, CA: Sociology Press.

Greene, J. (2007). *Mixed methods in social inquiry*. Hoboken, NJ: Wiley.

Greene, J. (2008). Is mixed methods social inquiry a distinctive methodology? *Journal of Mixed Methods Research, 2*(1), 7–22.

Greene, J., Caracelli, V., & Graham, W. (1989). Towards a conceptual framework for mixed-method evaluation designs. *Education Evaluation and Policy Analysis, 11*(3), 255–274.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14–26.

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods. *Journal of Mixed Methods Research, 1*(1), 112–133.

Lincoln, Y., & Guba, E. (2005). Paradigmatic controversies, contradictions, and emerging confluences. In N. K. Denzin & Y. S. Lincoln (Eds.) *The Sage handbook of qualitative research* (pp. 191–214). Thousand Oaks, CA: Sage.

McKenzie, J. (2008). *The role of student support services programs in motivating students to persist in postsecondary institutions*. Unpublished doctoral dissertation, Northern Illinois University, DeKalb.

Mertens, D. (2004). *Research and evaluation in education and psychology: Integrating diversity in qualitative, quantitative and mixed methods* (2nd ed.). Thousand Oaks, CA: Sage.

Morse, J. M. (1991). Approaches to qualitative quantitative methodological triangulation. *Nursing Research, 40*(2), 120–123.

Onwuegbuzie, A., & Johnson, B. (2006). The validity issue in mixed research. *Research in the Schools, 13*(1), 48–63.

Onwuegbuzie, A., & Teddlie, C. (2003). A framework for analyzing data in mixed methods research. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 351–383). Thousand Oaks, CA: Sage.

Patton, M. (1999). Enhancing the quality and credibility of qualitative analysis. *Health Services Research, 35*(5), 1189–1208.

Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches*. Thousand Oaks, CA: Sage.

Tashakkori, A., & Teddlie, C. (Eds.). (2003). *Handbook of mixed methods in social & behavioral research*. Thousand Oaks, CA: Sage.

Teddlie, C., & Tashakkori, A. (2006). A general typology of research designs featuring mixed methods. *Research in the Schools, 13*(1), 12–28.

Yin, R. (2003). *Case study research: Design and methods* (3rd ed.). Thousand Oaks, CA: Sage.

Yin, R. (2006, April). *Mixed methods research: Are methods genuinely integrated or merely parallel?* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

CHAPTER 20

# Guidelines for Writing Research Proposals

It is very important to plan ahead.

## INSTRUCTIONAL OBJECTIVES

After studying this chapter, the student will be able to:

1. Describe the components of a quantitative research proposal.
2. Describe the components of a qualitative research proposal.
3. Identify common weaknesses in quantitative research proposals.
4. Critique quantitative research proposals.
5. Critique qualitative research proposals.
6. Select statistics appropriate for analyzing the data collected.
7. Describe ethical and legal considerations in research.
8. Describe the role of an institutional review board.
9. Write a research proposal that meets the guidelines.

In most cases, researchers will need to present their projects in organized written form at two stages: (1) the initial stage, which requires preparation of a research proposal to be presented to a professor or to a dissertation committee, and (2) the final stage, a finished report of the results of the research. This chapter deals with the initial stage; Chapter 21 discusses the final stage.

In research, as in sampler making, planning ahead is essential.

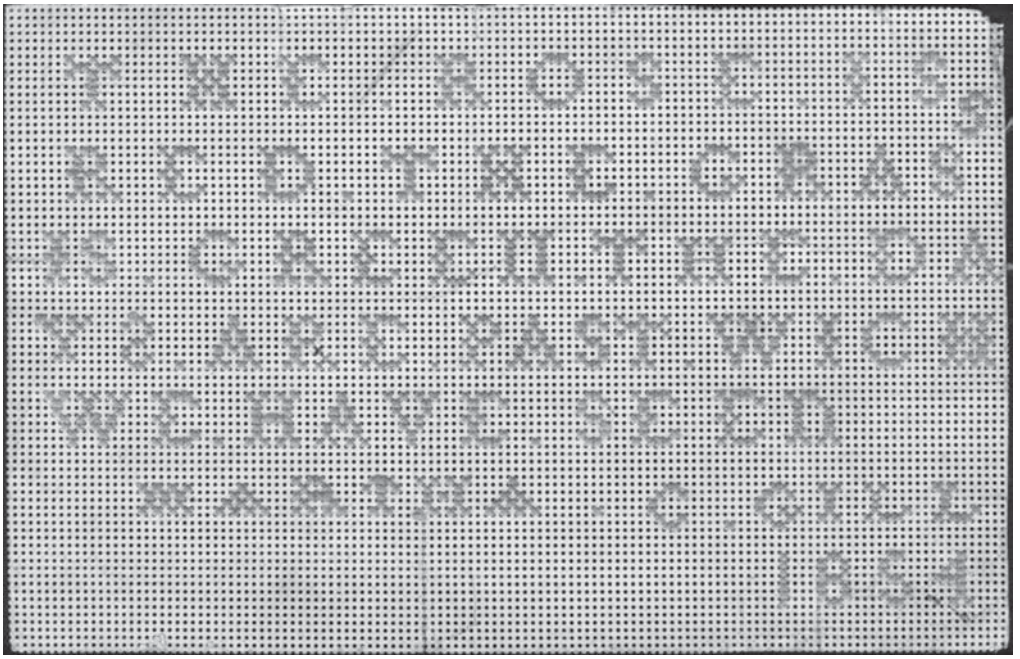


**Figure 20.1** Nineteenth-Century Sampler Discovered by One of the Authors in a Box of Books Purchased at a Farm Auction

## WRITING A RESEARCH PROPOSAL

A research proposal is a written plan for a project that will be submitted to others for evaluation. Writing the research proposal can be the most crucial and exciting step in the research process. At this stage, the whole project crystallizes into concrete form. The researcher's inspirations and insights are translated into step-by-step plans for discovering new knowledge. In a well-organized and well-written proposal, researchers demonstrate that they know what they are seeking and that they will know how to successfully complete the planned project. The readers will evaluate the worth of the proposed study and may make suggestions for improving it. The general format of a proposal is much the same whether it is a relatively informal outline offered by a student to satisfy the requirements of a research course, a formal thesis or dissertation proposal presented to a committee, or a funding request to a foundation or governmental agency.

## QUANTITATIVE RESEARCH PROPOSAL

The following outline shows the typical components of a quantitative research proposal:

1. Introduction
   a. Statement of the problem
   b. Review of the literature
   c. Statement of the hypothesis(es)
   d. Significance of the study

2. Methods
   a. Participants and methods of selection
   b. Instruments
   c. Procedures
3. Analysis of data
   a. Data organization
   b. Statistical procedures
4. Protection of human participants
5. Time schedule and budget
   a. Time schedule
   b. Budget
6. References

Although you need not follow this outline rigidly, it should provide a useful guide for writing any proposal because all aspects listed here must be considered. You should be aware, however, that your university department may require a specific format.

## INTRODUCTION

A crucial part of a research proposal for quantitative research is the introduction to the proposed study. In the very first paragraph, state the research problem clearly and unambiguously, avoiding emotionally charged words or phrases. Then link the problem to the body of information available in the field and establish the importance of and the need for carrying out the proposed research. The introduction should convince the readers that you have reviewed the related research and have a thorough understanding of the problem. Regardless of how tightly you formulate the design of the study and how well you select the statistical procedures, unless the introduction is written carefully and intelligently, other parts of the proposal will probably not receive serious consideration. It is not unusual for proposals to be turned down solely on the basis of a poor introduction, without much consideration being given to the proposed methodology and statistical design. The introduction provides the framework for the rest of the proposal. So prepare this section with care, caution, and the aim of promoting the reader's interest in the problem.

The introduction to a research proposal should include (1) a statement of the problem, (2) a review of the literature, (3) a statement of the hypothesis(es), and (4) the rationale for the study.

### Statement of the Problem

State the problem clearly and directly very early in the introduction, ideally at the beginning of the first paragraph. The problem statement should imply a question about the relationship between specified variables. The statement can take interrogative form or can simply state, "The purpose of the study is to explore the relationship between . . . and . . . ." You need not operationalize the variables at this point, only mention them in conceptual form. This section of the introduction should also include a brief description of the background of the problem and the potential significance of the study, although it is imperative to avoid the

temptation to sell the importance of the topic before stating it. Two common errors to watch for are (1) beginning the introduction with an elaborate presentation of the background of a problem before you have clearly stated the problem and (2) concentrating on a justification for the study at this point, burying the problem statement in the discussion or bringing it in only vaguely near the end. Another common error is to assume that the reader knows as much about the content of the question as you do. State the problem so that it can be understood by someone who is generally sophisticated but who is relatively uninformed concerning your specific problem.

At an appropriate point in this section, define, in the way they will be used in the study, any terms that may not be familiar to the reader or to which you are ascribing special meanings. The specific limitations of the scope of the study and a foreshadowing of the hypothesis should close the section.

### Review of the Literature

The literature review presents what is currently known about the problem under consideration and thus sets the scholarly context for the question or hypotheses of the proposed study. The review shows how the proposed research fits in with the existing body of knowledge. Enthusiastic beginning researchers often imagine their proposed study is unique and that no related research is available. This is very rarely the case. A thorough search almost always turns up several research papers related to at least some aspects of a proposed study. Even if there is no research in the field, there is usually literature of a theoretical or speculative nature that should be included as part of the background of a study. Of course, you should include theories and research results contrary to your stated hypothesis as well as those in agreement with it. You should include the findings from some of the landmark studies in the field as well as the most recent research. This will show your readers that your question has been investigated by others over a period of years and is a problem worthy of and in need of further investigation.

Include not only literature dealing with research on the variables of the proposed study but also discussions of the theoretical background of the problem. The literature review allows you to show your committee or other readers that you really know the scholarship that has been done on the question. The literature cited need not be exhaustive but should contain the most relevant and the most rigorous research on your question.

Try to integrate the studies so that a complete picture of the background of the problem emerges. You need to point our how the literature presented relates to the proposed study and the similarities and discrepancies among the findings. A pitfall to avoid in discussing related literature is to present a series of abstracts, one per paragraph. This offers the audience tedious reading and misses the opportunity for laying meaningful groundwork for the study. Not all related studies need be discussed in detail. In reviewing several similar studies, you can describe the most important one and then simply state that the results were confirmed in similar studies that are mentioned but not described in detail.

The related literature section should conclude with a discussion of the findings and their implications. Here, you share the insights gained from your review of the literature and can point out gaps in what is known about the topic, thus leading directly to the question you propose to investigate. The literature review can

be very time-consuming, but a thorough review during the proposal stage will save you time later when you write the final report. Only minor changes may be necessary.

### Statement of the Hypothesis(es)

The hypothesis should state concisely the expected relationship between the variables in the study. The research hypothesis is typically determined by the implications of the related research findings so far and the deductive logic of the underlying theory. Explain the link between theory and hypothesis so that the reader will be able to understand the rationale for the hypothesis.

Some research, like surveys, may simply state the problem in question form. For example, How do the parents in the Monroe County School District feel about a proposed new tech high school? But most quantitative research states a hypothesis. It should probably be stated in operational form; however, if this is not possible, it can be stated in conceptual terms followed by any operational definitions of the variables. Operational definitions of all variables must be included in any research proposal.

### Significance of the Study

In this section, you need to explain why the problem is an important one to study. Try to answer the following question: What difference will this study make to anyone? This section may be handled in two stages: implications and applications.

*Implications* Because the aim of research is to increase knowledge, the author of the proposal should show how his or her particular study will do this by discussing how the results will contribute to theory and knowledge in the specific area to which the research question is related and to what extent these results will be useful in solving problems and answering questions in the general field. Finally, you show how the results of the study will provide grounds for further research in the area.

In addition, you may explain how your own experience and expertise, coupled with the facilities and goals of the institution where the study is being carried out, place you in a favorable position to solve the problem in question.

*Applications* You should try to convince readers of the potential application of the findings to educational practice. To find the extent to which the study has application to educational practice, you may ask the following questions: Will the results of my study change anything in my field? and Would my results help teachers, school counselors, principals, or other educators to improve their work? This aspect carries considerable weight in attracting research funds for performing the study. Many foundations evaluate research proposals on the basis of potential application to practice.

## METHODS

In the methods section of the proposal, the writer details *how* the study will be carried out so that the research question will be answered or the hypothesized relationships will be observed, if in fact these relationships exist. This section should also convince the reader that what you plan to do is the best procedure for investigating

the question. Note that this section is written using future tense because nothing has yet been done. If the methodology you describe in the proposal is actually the same as the method you eventually use in conducting the study, then you have a major section of the final report already written. It is only necessary to change the future tense of the proposal to the past tense used in the final report.

The methods section includes a description of the population of interest, proposed sampling procedures, methods of data collection, research design, and procedures. A convenient way of presenting the research methodology is to categorize all information regarding the design as (1) participants, (2) instruments, or (3) procedures, as appropriate.

### Participants

The first step in identifying the participants in a study is to describe the population of interest: Is the study concerned with college freshmen, dyslexic 6-year-olds, principals of elementary schools, and so forth? Then the writer/researcher describes the procedure for drawing the sample from the population. If random selection is not possible, explain why you have chosen a particular procedure for sample selection and how well the sample to be used will resemble the population of interest. You need to include the number of participants and a description of their relevant characteristics, such as age, gender, academic level, ethnicity, and ability level. A careful description of the subjects can help the reader of the proposal to determine if, in the reader's view, the results of the study can be generalized from your sample to the appropriate population.

The following example describes the participants in a proposal:

> The children in this experiment all attend a public elementary school in the southeastern United States. Three intact classrooms will participate. There will be 17 children (11 boys and 6 girls) in kindergarten, 24 in second grade (13 boys and 11 girls), and 21 in fourth grade (10 boys and 11 girls). The children represent a variety of social, economic, and cultural backgrounds: Approximately one-third are African Americans, a few are Asian Americans, and the majority are European Americans.

### Instruments

The goal of a research project is to investigate relationships between constructs. However, because constructs are usually impossible to measure directly, you must select or develop indicators that will approximate them as well as possible. These indicators are the instruments you will use for data collection. If an instrument is already established, the proposal should include its name and reported evidence of its reliability and validity for the intended purpose. In cases in which the instruments are to be developed by the researcher, it is necessary to outline the procedure to be followed in developing them. This outline should include the steps that will be taken to obtain validity and reliability data on these instruments. If describing the reliability and validity procedures results in so much detail that it interrupts the continuity of the proposal, it is preferable to include this material in an appendix rather than in the text.

A description of the instrument might include, for example, the following:

> Students' prerecess and postrecess classroom behavior and their outdoor recess behavior will be observed daily for 2 months. A total of four observers, blind to the

purpose of the study, will be used: Three will be regular observers and one will be a rotating observer whose primary job is to be a reliability judge. Reliability will be measured with a κ coefficient.

### Procedures

A careful description of the procedures of a study is a basic requirement of a quantitative research proposal. In this section, the author describes the way in which the study will be carried out. By designing the study explicitly as an operation to permit the observation of the hypothesized relationships, you lay the foundation for the study.

Include in this section all the information necessary for replication of the study. In an experimental research proposal, for example, the author should describe the design by listing the groups and how subjects will be assigned to groups and by giving a step-by-step description of the manipulations planned for each. These steps should be completely designated in operational form. You should also mention the possibility of confounding variables and specify how you propose to control for these variables. For instance, you might control for student and teacher differences by randomly assigning students and teachers to the control and experimental groups and making the time spent, physical setting, and facilities equivalent. If the proposal involves a survey, you need to describe all the steps involved in the development of the instrument, its distribution, follow-up procedures, training of the interviewers (if used), and so on.

You should also mention any known limitations of the proposed methodology. The limitations are any factors that may negatively impact the results of the study or the generalizability of the results. Limitations may include size of the sample, selection of the participants, the setting for the study, use of a pretest, length of the study, and so on. But you should be careful about pointing out too many weaknesses. Bryant (2004) writes, "Don't downplay the potential importance of your study. One wants to be cognizant of weaknesses in design and method, but I think it is a strategic mistake to speak too loudly of weaknesses in concept or method until after the dissertation proposal has been approved" (p. 59). Assuming the study goes forward, one can always discuss limitations in the final report.

Documents such as teaching or reading materials planned for an experiment need not be included in the main text of the proposal because discussion of such details usually interrupts continuity of the proposal. We recommend that the author place these documents in an appendix, describing them briefly but clearly in the procedures section. However, it is essential to explain in this section any differences in the presentation of these materials to the different groups involved in the experiment.

After the section on procedures has been drafted, read it to verify that you have described all steps necessary to answer every question and test every hypothesis. You can check the completeness of this section with the following question: Given the time and resources needed, could the reader carry out this research by following the steps as described? If you can answer this question in the affirmative, this section is complete.

## PROTECTION OF HUMAN SUBJECTS

This section should contain a brief summary of the procedures you will follow to protect the rights of the participants in your study. Any research that involves

human subjects must be reviewed by your school's institutional review board. The board may not review the entire proposal in detail, but it will be interested in this section in order to evaluate your plan for protecting the human subjects. In addition, the board may require you to complete certain forms relevant to the proposed investigation. We discuss ethical issues in more detail later in this chapter.

## DATA ANALYSIS

The next part of the research proposal describes the methods of handling and presenting data and outlines the statistical procedures to be used. Group this information in sections covering data organization and statistical procedures.

### Data Organization

In this section, you describe your plan for organizing and presenting the results of the investigation. This is best done by reference to questions or hypotheses of the study. List each hypothesis and then indicate the data that will be collected to address each one.

Planning in advance for the organization and presentation of data enables a researcher to determine whether the information being collected is relevant to the research questions. Those who bypass this step often find they have wasted considerable time and money in collecting irrelevant pieces of information.

At this stage, you should list the bits of information that will be available for each participant and be able to picture how you will organize and present the data. Tables, figures, and charts are essential means for organizing and summarizing a whole set of data.

### Statistical Procedures

The design of the study determines the statistical analysis to be employed, not vice versa. In other words, the researcher decides what design will permit observation of the hypothesized relationships and then selects the statistical analysis that fits the questions asked and the nature of the data involved. In this section, one should explain what data will be considered, what statistical procedures will be used and why, and what results will be important in confirming the hypothesis.

The most commonly used statistical procedures were described in previous chapters. They are summarized for convenience in Tables 20.1 (descriptive statistics) and 20.2 (inferential statistics). The reader is also referred to Rowntree (2004) for a discussion of statistical procedures.

Table 20.1 is designed to help identify the indexes that may be used to describe in summary form the data of a study. The appropriate statistical procedure is determined partly by the type of measurement scale characterizing the dependent variable. Therefore, the rows in the table are identified as interval, ordinal, and nominal. Columns 1 through 3 list the various purposes that descriptive statistics may serve. The following are the most common uses of these statistics:

1. To provide an index to describe a group or the difference between groups (measures of central tendency)
2. To provide an index to describe the variability of a group or differences in the variability of groups (measures of variability)
3. To locate an individual in a group (indexes of individual location)

**Table 20.1** Descriptive Statistics

| Type of Scale of Dependent Variable | Purpose of the Statistic: (1) Central Tendency | | (2) Variability | | (3) Location | |
|---|---|---|---|---|---|---|
| | One Group | More Than One Group | One Group | More Than One Group | One Group | More Than One Group |
| Interval | Mean | Difference between means | Standard deviation of variance | Difference between standard deviation of variances | *z* score, or other standard score | Difference between an individual's score in more than one distribution |
| Ordinal | Median | Difference between medians | Quartile deviation[a] | Difference between quartile deviations[a] | Percentile rank | Difference between an individual's percentile rank in more than one distribution |
| Nominal | Mode | Difference between modes | Range | Difference between ranges | Label or categorization | Label or categorization |

| Type of Scale of Dependent Variable | Purpose of the Statistic: (4) Correlation | | (5) Subsets | | (6) Interaction | |
|---|---|---|---|---|---|---|
| | One Group | More Than One Group | One Group | More Than One Group | One Group | More Than One Group |
| Interval | Pearson *r* | Difference in Pearson *r* for some variables in two groups | | | Difference between observed cell means and expected cell means in factorial ANOVA (observed interaction) | Differences in observed interaction among groups |
| Ordinal | Spearman's rho or Kendall's tau[a] or $W$[a] | Difference in Spearman rho, Kendall's tau, or $W$ | | | | |
| Nominal | Point biserial correlation,[a] biserial correlation,[a] or phi coefficient | Difference in correlations for same variables in two groups | Proportion or percentage | Differences in proportions or percentages | Difference between observed cell frequencies and expected cell frequencies | Differences in observed interaction among groups |

[a]This statistic is not described in this text but may be found in any number of statistics texts.

4. To provide an index to describe the relationship of variables within a population (measures of correlation)
5. To describe how a set is divided into subsets
6. To describe the interaction among two or more variables in relation to a criterion (measures of interaction)

**Table 20.2** Inferential Statistics

| Type of Scale of Dependent Variable | Purpose of the Statistic: (1) Central Location | | (2) Variability | | (3) Individual Location | |
|---|---|---|---|---|---|---|
| | One Group | More Than One Group | One Group | More Than One Group | One Group | More Than One Group |
| Interval | Standard error of the mean | *t* Test or one-way ANOVA | | Bartlett's test[a] or *t* test for homogeneity of variance;[a] *F*-max[a] statistics | Standard error of measurement[a] | Standard error of difference scores[a] |
| Ordinal | Standard error of median[a] | Wilcoxon and Mann–Whitney tests,[a] Kruskal–Wallis one-way ANOVA,[a] or Friedman's test[a] | | | | |
| Nominal | | | | | | |

| Type of Scale of Dependent Variable | Purpose of the Statistic: (4) Correlation | | (5) Subsets | | (6) Interaction | |
|---|---|---|---|---|---|---|
| | One Group | More Than One Group | One Group | More Than One Group | One Group | More Than One Group |
| Interval | *t* Test for Fisher's *z* transformation or *F* test for linearity[a] | *t* Test for Fisher's *z* transformation[a] | | | *F* test for multifactor ANOVA | *F* test for multifactor ANOVA |
| Ordinal | Test for Spearman's rho or Kendall's tau[a] | Test for Kendall's *W* | | | | |
| Nominal | Chi-square test for goodness of fit | Cochran's *Q*[a] | Chi-square or binomial test[a] | Chi-square test of independence | Log linear analysis[a] | Chi-square test for log linear analysis[a] |

[a]This statistic is not described in this text but may be found in any number of statistics texts.

The required cell can be located by identifying the row and column heading appropriate to your study. Each cell is divided, and the section to use is determined by whether the study is concerned with one group or with more than one group. (Recall that you may choose a procedure for a lower scale of measurement but not the reverse; for example, you may use a median or a mode to describe interval data but may not use a mean to describe ordinal or nominal data.) In determining what type of scale to use in expressing the data, consider the advantages of each of three scales. Interval data provide more information than ordinal data, and ordinal data provide more information than nominal data. In making inferences, statistical tests of interval data are more "powerful" than tests of ordinal data; that is, you have a greater chance of rejecting a null hypothesis when you use interval measures than when you use ordinal measures. In the same manner, ordinal tests are more powerful than nominal measures. Therefore, when a

choice is possible, one prefers interval data to ordinal and ordinal data to nominal data. For example, if you have interval data for the dependent variable and want an index to describe the difference between groups, Table 20.1 identifies the difference between two means as an appropriate statistic. (You could use the difference between medians, but this would be less powerful than the difference between means.)

If the study is inferential in nature, the researcher will proceed to test the statistical significance of the index selected. Table 20.2 lists appropriate statistics for this purpose. You will also want to calculate a measure of effect size.

Remember that a statistical procedure is selected on the basis of its appropriateness for answering the question involved in the study. Nothing is gained by using a complicated procedure when a simple one will do just as well. Statistics are to serve research, not to dominate it. We often tell our students, "Imagine you have gotten a fantastic grant for doing your study, but you are run over by a truck the very next day. Could a colleague pick up your proposal, actually conduct the study, and analyze the data?" If you can honestly answer this question in the affirmative, the data analysis part of the proposal is complete.

## TIME SCHEDULE AND BUDGET

A proposal should conclude with a presentation of time schedule and budget.

### Time Schedule

The researcher must plan with regard to the feasibility of completing the study within the time limitations. You need to identify the sequence of stages necessary to complete the proposed research and to estimate the time needed to complete each stage. Be generous in allocating time for each step because unanticipated events can cause delays. Research almost always takes more time than we think. Dividing the project into manageable stages and then assigning a date to complete each stage helps systematize the project, enables you to estimate the effort that will be required to complete the entire project, and provides opportunities for periodically evaluating the progress of the project.

### Budget

Reviewing the previous sections of the proposal, the researcher now needs to determine the resources necessary to carry out the proposed research and make careful estimates of the costs of such things as supplies, mailing, travel expenses, duplicating, personnel, and consultant fees. If you are requesting external funding, most institutions have someone, such as a contract officer, to help plan the budget and fit it into the requirements of the anticipated funding agency.

## REFERENCES

In this section, the researcher must list the references that were cited in the literature review as well as any other sources cited in the development of the proposal. The reference list includes the full publication information for each entry, arranged in alphabetical order according to the senior author's last name. In the field of education, the American Psychological Association (APA) format is the most popular for listing references in journals, dissertations, and other research

publications (APA, 2001, pp. 215–281). Beginning researchers should become familiar with APA style.

## CRITIQUING THE PROPOSAL

After completing the draft of a proposal, the author/researcher should read it again carefully with a critical eye. It is also profitable to have colleagues read the proposal. Often, someone else can identify weaknesses or omissions that are not evident to the author. Some of the common weaknesses found in proposals are as follows:

Introduction

1. The problem is not clearly and unambiguously stated. The reader cannot be certain what the researcher plans to do.
2. The problem is too broad for a single study.
3. The problem lacks practical or theoretical significance.
4. The study has not been linked to the existing body of knowledge in the field.

Methods

5. The proposed design is not appropriate for answering the research question.
6. There is no discussion of extraneous variables and how they might be controlled.
7. The proposal does not provide operational definitions of the variables.
8. The proposal does not adequately discuss the validity and reliability of measuring instruments.
9. The proposal is not sufficiently detailed to allow for replication.

Analysis of Data

10. The statistical procedures are not appropriate for analyzing the data.

## IMPORTANCE OF COMPLETING THE PROPOSAL BEFORE COLLECTING DATA

In research, a clear, well-stated, comprehensive proposal indicates that the prospective researcher is actually ready to set the study in motion. It shows that the researcher knows what to do, why to do it, and how to do it. A prospective researcher who cannot produce a complete and coherent proposal is clearly not yet ready to proceed to the data collecting stage of the project. Novice researchers are often inclined to say, “Let me collect my data now and decide what to do with it later.”

Simultaneously collecting data and writing the proposal may seem to be a time-saving procedure, but that is seldom the case. Countless work-hours and thousands of dollars have been wasted in just that way. Until the proposal is formulated, you cannot be sure exactly what data you will need or what will be the best way to handle this information in light of the study’s purpose. A well-written proposal will be a great benefit when you write the final report. If the introduction, review of literature, and methods sections are comprehensive and well organized in the proposal, then they will be ready for the final product.

# QUALITATIVE RESEARCH PROPOSAL

Although a qualitative research proposal contains many of the same components as a quantitative proposal, there are some significant differences. Because the design of qualitative research is emergent, the proposal is less structured and more flexible. Whereas a qualitative proposal begins with a broad picture of the intended study, the contour and details of the study, including the guiding questions and procedures, may change during the research process. A qualitative study responds to the context and setting of the study so that a proposal may not in the end be entirely aligned with what actually happens.

The style of writing in a qualitative proposal is more personal and less formal than it is in a quantitative proposal. Quantitative proposals tend to read more like a science report, whereas qualitative proposals have a more narrative style and may read more like a story. The writer may refer to himself or herself using the first person, "I," which would not be acceptable in quantitative proposals. A qualitative proposal should reflect the format and personalized style that will be used in the final product.

A well-written qualitative research proposal helps others understand what is under study, why the study is important, how the inquiry will be conducted, and how the results will contribute to the body of knowledge. A brief outline of recommended components in a qualitative proposal follows.

1. Introduction
   a. Purpose of the study
   b. Situating the self
   c. Initial guiding questions
   d. Review of relevant literature/discourse
2. Research procedure
   a. Site and selection of the sample
   b. Description of the setting
   c. Role of the researcher
   d. Data collection methods
   e. Ethical issues
3. Data analysis/presentation
   a. Data analysis strategies
   b. Validity and dependability issues
4. Importance/significance of the study
5. Time schedule and budget
6. References
7. Appendices

## INTRODUCTION

### Purpose

The introduction provides a clear statement of the purpose of the study. Provide sufficient background information so that the reader can understand the purpose of the study and can place it in context. One might identify the research purpose or focus for a qualitative study from everyday practice and experience or from

current social or political issues. Although not as common as in quantitative research, the focus may be derived from theory with an intent to elaborate or modify the theory. The research purpose or focus of the study should be realistic, not too vast and not too narrow.

### Situating the Self

As you discuss the purpose and background of the research, you should assess your own emotional involvement in the topic. Although you should be interested, perhaps even passionate, about your topic, you must not be so emotionally involved that your own biases prevent you from adequate interpretation of the findings. One way to ensure that you consider this aspect is to include a section in the introduction that situates yourself in the study. Why are you interested in this topic? What personal background and knowledge do you bring to the study? Some scholars recommend an autobiographical section that addresses the potential for bias.

### Initial Research Questions

Once the overall focus of the study has been defined, you are ready to state the central research question and the subquestions to be addressed. These questions lay out the conceptual structure of the inquiry. Qualitative research questions are typically about process or understanding—what happens, why or how something happens, how those involved understand the experience, and what it means to them. These questions will help identify what to observe or what to ask in interviews and direct the analysis, similar to how hypotheses in quantitative research direct the methodology. As the study unfolds, however, these initial questions may be modified. In this case, you may need to meet again with your professor and/or committee to discuss any changes and to get their approval for the changes.

### Related Literature/Discourse

The literature review helps you identify what is already known, how it relates to your question, how your study might contribute to greater understanding of the topic, and the potential theoretical frameworks that might inform the study. Reviewing the literature helps prevent duplication of research and helps avoid making mistakes in research design that others have made.

There are two key differences in the literature review in quantitative and qualitative studies. First, in a qualitative proposal, it is not assumed that the literature review is complete in the proposal. Reviewing literature in a qualitative study is an ongoing process and involves an interactive process throughout the study. As the study evolves, new literature may need to be explored. Also, literature may be embedded throughout the final document to support findings and is not restricted to a particular section of the report. Second, different kinds of information are considered appropriate to be included in the qualitative proposal; hence the call by some authors to call this section a "review of discourse" rather than a review of literature. For example, it may be appropriate to trace the historical evolution of the study topic, to trace the conceptual threads or themes within discourse on the topic, to map various schools of thought or ideological positions or disciplinary perspectives related to the topic, or to describe the positions of various stakeholder groups. You may also review personal experience and knowledge that are relevant to the inquiry.

## RESEARCH PROCEDURE

This section describes the overall qualitative approach to be used in the study. Will it be a case study, an ethnographic study, or one of the other qualitative approaches? You may need to include a rationale for the selection of the specific research genre. One should then explain how each of the guiding questions will be addressed—what information will be used, how the information will be gathered, from what sources, and so forth.

### Site and Selection of the Sample

Qualitative proposals should contain a complete description of the site you will choose and the rationale for choosing that site. This section also provides information about the sample, the size, and the method for selecting the sample. Typically, samples in qualitative research are much smaller than those in quantitative research. The sample size might be one, as in a case study, or it might involve a larger group, such as the entire faculty of a high school or all the children in a particular elementary school. If you use purposive sampling, you should describe the selection criteria and the essential attributes participants must have, and you should explain why these criteria are important in the study. Although you may not know the exact number of individuals, cases, or documents you will eventually include in the study, you should approximate and explain how adjustments may be made as the study is under way.

### Role of the Researcher

In this section, the researcher describes the procedure for gaining entry to the chosen site and how a relationship of trust will be established with the participants. The researcher will specify his or her role as an observer, a participant observer, or some combination of the two.

### Data Collection Methods

One must describe the methods that will be used to collect the data to answer each research question. Indicate whether you will use observation, interviews, or documents. If you plan to use observations, tell what kind will be conducted and where they will be done. Will semistructured or open interview processes be used? You might provide examples of questions that you will ask. If you plan to analyze documents, tell what documents you will use and how they will be collected. You will want to specify any audiovisual materials you will use, such as recordings, videotapes, and photographs. The quality of qualitative research is greatly improved by using multiple methods of collecting data. Recall that multiple methods permit triangulation, which provides important evidence of the credibility of the study.

You also need to discuss how you will handle ethical issues such as obtaining informed consent and maintaining confidentiality. Issues related to the ethical treatment of human subjects are discussed later in this chapter.

## DATA ANALYSIS

Typically, the qualitative researcher will collect a massive amount of data from the interviews, observations, and field notes. You will need to tell how you will

organize and analyze these data. This section should also address issues related to the study's validity and dependability and discuss the strategies that will be used to enhance credibility.

## SIGNIFICANCE OF THE STUDY

You will need to convince the reader that your proposed study is important. Tell what can be learned as a result of the study and why that is worth knowing. State the implications of the findings for educational practice and/or theory. We recommend, however, that you exercise caution in discussing the potential contributions of your study to the organized body of knowledge in education. Remember that the findings may be unique to the single group or specific context of your study and may not necessarily transfer to other contexts or groups.

## TIME SCHEDULE AND BUDGET

As was true for the quantitative research proposal, you will need to estimate the time and costs involved in conducting the study. Qualitative research requires a great deal of time in the field and at least an equal amount of time for data analysis.

## CRITIQUING THE QUALITATIVE PROPOSAL

After you have completed writing the proposal, go through it again and see if you have addressed the following questions:

Introduction

1. Is the purpose of the study clearly stated?
2. Is there a description of the writer's background and motivation that might affect the collection and interpretation of data?
3. Is there a clearly stated central or guiding research question(s)?
4. Does the review of literature provide context for the study?

Research Procedure

5. Is there an explanation of how each of the guiding research questions will be addressed?
6. Is there a rationale for choosing a particular site and a description of the steps to be taken in gaining access to the site?
7. Is there information about the sample and the method for selecting the sample?
8. Is there sufficient information about the data collection methods? Are multiple methods of collecting data planned?
9. Have ethical issues been addressed?

Analysis of Data

10. Is there an adequate description of how the data will be organized and interpreted?
11. Is there a discussion of the significance of the study?

## ETHICAL AND LEGAL CONSIDERATIONS

Strict adherence to **ethical standards** in planning and conducting both qualitative and quantitative research is most important. Researchers have obligations both to their subjects and to their profession. The proposal should state how informed consent, confidentiality, and other ethical issues will be handled. Organizations such as the American Educational Research Association (AERA) and the APA have formal ethical guidelines for conducting research. AERA's code of ethics incorporates a set of standards designed specifically to guide the work of researchers in education. The ethical standards for AERA, adopted in 1992 and currently under revision, state that as educational researchers,

> we should strive to maintain the integrity of our research, of our research community, and of all those with whom we have professional relations. We should pledge ourselves to do this by maintaining our own competence and that of people we induct into the field, by continually evaluating our research for its ethical and scientific adequacy, and by conducting our internal and external relations according to the highest ethical standards." (p. 23)

### OBLIGATION TO SUBJECTS

When employing human subjects in research, you must respect their rights, dignity, privacy, and sensitivities. Eleven aspects of researchers' obligations to participants are identified in the AERA standards (1992, p. 24):

1. Participants, or their guardians, in a research study have the right to be informed about the likely risks involved in the research and of potential consequences for participants and to give their **informed consent** before participating in research. Educational researchers should communicate the aims of the investigation as well as possible to informants and participants (and their guardians), and appropriate representatives of institutions, and keep them updated about any significant changes in the research program.
2. Informants and participants normally have a right to **confidentiality,** which ensures that the source of information will not be disclosed without the express permission of the informant. This right should be respected when no clear understanding to the contrary has been reached. Researchers are responsible for taking appropriate precautions to protect the confidentiality of both participants and data to the full extent provided by law.
3. **Honesty** should characterize the relationship between researchers and participants and appropriate institutional representatives. Deception is discouraged; it should be used only when clearly necessary for scientific studies and should then be minimized. After the study the researchers must explain to the participants and institutional representatives the reasons for the deception.
4. Educational researchers should be sensitive to any locally established institutional policies or guidelines for conducting research.
5. Participants have the right to withdraw from the study at any time.
6. Educational researchers should exercise caution to ensure that there is no exploitation for personal gain of research populations or of institutional

settings of research. Educational researchers should not use their influence over subordinates, students, or others to compel them to participate in research.

7. Researchers have a responsibility to be mindful of cultural, religious, gender, and other significant differences within the research population in the planning, conduct, and reporting of their research.
8. Researchers should carefully consider and minimize the use of research techniques that might have negative social consequences—for example, experimental interventions that might deprive students of important parts of the standard curriculum.
9. Educational researchers should be sensitive to the integrity of ongoing institutional activities and should alert appropriate institutional representatives of possible disturbances in such activities that may result from the conduct of the research.
10. Educational researchers should communicate their findings and the practical significance of their research in clear, straightforward, and appropriate language to relevant research populations, institutional representatives, and other stakeholders.
11. Informants and participants have a right to remain anonymous. This right should be respected when no clear understanding to the contrary has been reached. Researchers are responsible for taking appropriate precautions to protect the confidentiality of both participants and data. Those being studied should be made aware of the capacities of the various data-gathering technologies to be used in the investigation so that they can make an informed decision about their participation. It should also be made clear to informants and participants that despite every effort made to preserve it, **anonymity** may be compromised. Secondary researchers should respect and maintain the anonymity established by primary researchers.

## OBLIGATION TO THE PROFESSION

The researcher is also responsible to the consumers of research. Many research studies, in education as well as in other fields, are published in journals, monographs, books, and other media and are referred to and used by professionals in the field. You are morally obligated to plan a study in such a way that the findings obtained would not result in offering misleading information. Even more, you are obligated to report exactly and honestly what the findings were. Reporting research in a misleading way is a serious abuse of a researcher's responsibility to the profession. It is dismayingly easy to conduct research that becomes unintentionally deceptive. In a survey, a researcher may inadvertently phrase questions in such a way as to elicit the answers sought. In conducting research, someone may, without noticing, arrange things so that the experimental group has advantages that the control group does not have that are not part of the independent variable. Before beginning the research, it is recommended that a trusted colleague go through the research plan to look for potential sources of misleading outcomes.

If the statistical analyses fail to support, or contradict, one's hypothesis, it is very tempting to suppose that a different statistical analysis would have "come

out right." It is important to select the most appropriate statistical analysis before conducting a study and then derive conclusions from those statistics. It is also important to report the results of *all* statistical analyses, not just those that are statistically significant. For example, if you test 50 hypotheses and find that 2 are statistically significant, reporting only those 2 implies that they were unlikely to be a function of chance—when actually 2 statistically significant results in 50 are very likely to be a function of chance.

With computers it is very easy to search through the data for subsets of the sample in which one's hypotheses are confirmed despite retained null hypotheses for the total sample. You may, for example, discover what you seek among sixth-grade girls with Spanish surnames. It is appropriate to plan statistical contrasts between treatments, between sexes, between grade levels, and between ethnic groups before conducting a study. However, the results of unplanned post hoc analyses of subgroups can lead to misleading conclusions.

## LEGAL OBLIGATIONS

Researchers' obligations to human subjects are not only incorporated into the professional ethics statements of professional organizations such as AERA and the APA but also have been incorporated into federal regulations. Prior to 1974, there were several infamous studies in which researchers placed subjects at definite risk in order to conduct their study. One of the most heinous (critiqued by Jones, 1982) was the Tuskegee syphilis experiment, begun in 1932 in Tuskegee, Alabama, to study the progression of the disease without treatment. Later, when penicillin was discovered, subjects were not treated because they were part of the control group in a comparative study of the treatment. Stanley Milgram's (1963) research on reaction to authority deceived participants into administering "electric shocks" of up to 450 volts to accomplices who acted as though they were in pain. The participants later experienced stress and anxiety because they thought they had deliberately caused pain to others. Laud Humphrey's (1970) study of homosexuality was another controversial study because of its deception and gross invasion of privacy.

Studies such as these prompted the government to develop a unified set of regulations for research. In 1974, the federal government passed legislation that applies to all research performed by or under the general supervision of institutions receiving federal research funds. The National Research Act of 1974 requires that to ensure protection of participants, all proposed research involving human subjects be reviewed and approved by an authorized board before the study begins. The Act resulted in the establishment of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research to develop a basic set of ethical guidelines for conducting research.

This commission produced the **Belmont Report** (1979), which identified three essential ethical principles that must be followed in all research with human subjects: (1) protection of the human subjects from harm (physical or mental), (2) respect for subjects' right to know the nature and purpose of the study and their right to give or withhold consent to participate (the right of informed consent), and (3) respect for subjects' privacy. One stipulation of the law is that colleges and universities engaging in research with human subjects have an **institutional review board** (IRB) that must approve each proposal and certify

that the research will be conducted in accordance with the law. An IRB consists of a group of scholars from various departments who are mandated to review proposals to ensure that they comply with ethical guidelines. These boards, also referred to as human subjects review boards, require researchers to complete a lengthy form that provides a summary of the purpose and procedures of the proposed project and specific information about possible risks to subjects. Exhibit 20.1 lists the questions that the researcher must answer in a Summary Safeguard Statement.

If the members of the review board are satisfied that the research will not place subjects at risk, they sign the approval form. Failure to follow all the procedures almost guarantees that the research will not be approved.

### Protecting Subjects from Harm

For research that is funded by the Department of Health and Human Services and that involves human subjects outside the exempted categories, the federal regulations require a determination of whether the subjects will be placed "at risk" by the research procedures. If the researcher can demonstrate that the physical, mental, and social risks involved in the proposed project are no greater than those encountered "in daily life or during the performance of routine physical or psychological examinations or tests," the subjects are not considered "at risk." If the subjects are "at risk," then the researcher must be able to argue that (1) the value of the knowledge that is likely to be gained exceeds the potential harm; (2) all subjects will be fully informed of the risks in the procedures, and the voluntary and written consent of each subject will be obtained; and (3) appropriate medical or other support services will be available to subjects who participate in the research. If these conditions are met, the IRB will usually approve the proposed work. Typically, there is no appeal procedure from an IRB.

### Obtaining Informed Consent

Another stipulation of the National Research Act is that the researcher must obtain informed consent from research participants. Consent should be (1) voluntary, (2) informed, and (3) given by a competent individual. Obtaining consent is more than simply having a potential research participant sign a form. An informed consent statement has two purposes: (1) to enable potential research subjects to make an informed choice as to their participation in the study and (2) to document their decision to participate. To make an informed choice, potential participants must understand the purpose of the study, the procedures, the risks, and benefits of the project, as well as the obligations of both the participants and the researcher. Informed consent is documented by the use of a written consent form approved by the institution's IRB and signed by the subject or the subject's legally authorized representative.

In some cases, it may be argued that prior knowledge of the purposes of the study would bias the results. Participants may change their normal behavior or they may respond in the way they think the researchers expect. Thus, researchers sometimes do not tell participants the *true* purpose of the study but instead use **deception.** There are two types of deception: (1) passive deception, in which subjects are not told untruths but are simply not given sufficient information about the purpose of the study, and (2) active deception, in which

**Bloomington Campus Committee for the Protection of Human Subjects**
**Summary Safeguard Statement**

Project Title (if you wish to use a different title in the consent statement than is listed on page 3, explain here).

_____

_____

A. Briefly describe, in lay terms, the general nature and purpose of the proposed research, and where the study will take place. If student research, indicate whether for a course, thesis, dissertation, or independent research. If the study is only for a course, please review the Student Research Policy to ascertain if this project requires HSC review.

B. Describe how subjects will be recruited, how many (or estimate) subjects will be involved in the research, and how much time will be required of them. List specific eligibility requirements for subjects (or describe screening procedures), including those criteria that would exclude otherwise acceptable subjects. If your study uses only male or female subjects, explain why. For NIH-funded research only, address the inclusion of women, minorities, and children in the research. Disclose any relationship between researcher and subjects—such as teacher/student, superintendent/principal/teacher, and employer/employee.

C. Check appropriate box for type of vulnerable subject population involved when investigation specifically studies:
☐ minors (under age 18), ☐ fetuses, ☐ pregnant women, ☐ persons with mental disabilities, ☐ prisoners, ☐ persons with physical disabilities, ☐ economically or educational disadvantaged, ☐ other vulnerable population.
If any of the above are used, state the necessity for doing so. Please indicate the approximate age range of the minors to be involved.

D. List all procedures to be used on human subjects or describe what subjects will do. If done during regular class time, explain what nonparticipants will do. If you are taping, explain that here. *Asterisk* those you consider experimental. For those asterisked

**Exhibit 20.1** Indiana University, Bloomington Campus Committee for the Protection of Human Subjects Summary Safeguard Statement, 2007

subjects are deliberately given misinformation about the true purpose of the study. The true purpose is completely different from what the researcher tells the subjects. Ethical guidelines state that the deception must be justified in terms of benefits that outweigh the risks. It is still important to maintain the voluntary nature of participation even if the purposes are not fully or accurately presented prior to the procedures. In cases in which it is important to use deception, the researcher must debrief the participants by giving a full explanation after the completion of the study. Coercion of subjects and fraudulent explanations of purpose are prohibited by virtually every professional code of ethical standards.

The obtaining of informed consent from people who are not considered capable of representing themselves is problematic. If their competence limits their ability to give consent, researchers should obtain these individuals' *assent* to participate

procedures, describe the usual method(s), if any, that were considered and why they were not used.

E. State the potential risks—for example, physical, psychological, financial, social, legal, or other—connected with the proposed procedures. Briefly describe how risks to subjects are reasonable in relation to anticipated benefits. Describe procedures for protecting against, or minimizing, potential risks. Assess their likely effectiveness. If you are using an electrical device that is attached directly to subjects, explain how the subjects will be protected from shock.

F. Describe methods for preserving confidentiality. How will data be recorded and stored, with or without identifiers? If identifiers are used, describe the type: names, job titles, number code, etc. How long are identifiers kept? If coding system is used, is there a link back to the subject's ID? If yes, where is the code list stored in relation to data and when is the code list destroyed? How will reports be written, in aggregate terms or will individual responses be described? Will subjects be identified in reports? Describe disposition of tapes/films at the end of the study. If tapes are to be kept, indicate for how long and describe future uses of tapes.

G. What, if any, benefit is to be gained by the subject? In the event of monetary gain, include all payment arrangements (amount of payment and the proposed method of disbursement), including reimbursement of expenses. If class credit will be given, list the amount and the value as it relates to the total points needed for an A. List alternative ways to earn the same amount of credit. If merchandise or a service is given, indicate the value. Explain the amount of partial payment/class credit if the subject withdraws prior to completion of the study.

H. What information may accrue to science or society in general as a result of this work?

I. Coinvestigators, Cooperating Departments, Cooperating Institutions. If there are multiple investigators, please indicate only one person on the Documentation of Review and Approval (page 3) as the principal investigator; others should be designated as coinvestigators here. Coinvestigators . . . should sign here, pledging to conform to the sentences on page 3. If you are working with another institution, please include a letter of cooperation from that institution. Please provide the person's name and e-mail address.

**Exhibit 20.1** *Continued*

and the consent for their participation from their guardians. Researchers facing this problem are advised to consult with the chair of their IRB or with the legal counsel of their institution or organization. The Office for Protection from Research Risks in the Department of Health and Human Services is usually willing to consult on these matters and may be able to provide current information on legal rulings relevant to various groups.

Exhibit 20.2 presents a sample informed consent statement showing the kind of information the researcher would provide to potential participants. The individuals would read and sign, if they are willing to participate.

### Right to Privacy

Respect for **privacy** is at the heart of the conduct of ethical research with human participants. A piece of federal legislation called the Family Educational Rights and Privacy Act of 1974 (also called the Buckley Amendment) was designed to

**Informed Consent Statement**

*Indiana University-Bloomington*

You are invited to participate in a research study. The purpose of this study is _____

_____.

**Information**

Describe all procedures, preferably in chronological order, which will be employed in the study. Point out any that are considered experimental and explain technical and medical terminology.

State the amount of time required of the subject per session and for the total duration of the study.

*If applicable to your study, describe:*

Alternative procedures or courses of treatment (when experimental procedures are being used).

The number of subjects that will be participating in the research.

Information concerning taping or filming.

A disclaimer for the use of deception.

**Risks**

List the foreseeable risks or discomforts, if any, of each of the procedures to be used in the study, and any measures that will be used to minimize the risks.

**EMERGENCY MEDICAL TREATMENT** (if applicable)

In the unlikely event of physical injury resulting from your participation in this research, emergency medical treatment will be provided at no cost to you. Be certain that you immediately notify the researcher if you are injured. If you require additional medical treatment, you will be responsible for the cost. No other compensation will be provided if you are injured in this research.

**Benefits**

List the benefits you anticipate will be achieved from this research to the subjects, others, or the body of knowledge.

**Confidentiality**

Describe the extent, if any, to which confidentiality of records identifying the subject will be maintained. OR, explain when and how confidentiality will be broken.

**Exhibit 20.2** Sample Informed Consent Statement

protect the privacy of students' educational records. It specifies that any data that identifies a student may not be made available unless written permission is obtained from the student (if of age) or a parent/legal guardian. The consent specifies what data may be accessed, for what purpose, and by whom. Participants

**Compensation**

For participating in this study you will receive _____. Other ways to earn the same amount of credit are _____. If you withdraw from the study prior to its completion, you will receive _____.

**Contact**

If you have questions at any time about the study or the procedures (or you experience adverse effects as a result of participating in this study), you may contact the researcher.

Name Address Phone number e-mail

If you feel you have not been treated according to the descriptions in this form, or your rights as a participant in research have been violated during the course of this project, you may contact the office for the Indiana University Bloomington Human Subjects Committee.

Address Phone number e-mail

**Participation**

Your participation in this study is voluntary; you may refuse to participate without penalty. If you decide to participate, you may withdraw from the study at any time without penalty and without loss of benefits to which you are otherwise entitled. If you withdraw from the study before data collection is completed, your data will be returned to you or destroyed.

**Consent**

I have read this form and received a copy of it. I have had all my questions answered to my satisfaction. I agree to take part in this study.

Subject's signature _____ Date _____

I agree to allow my child, _____, to take part in this study.

Parent's signature _____ Date _____

Witness signature _____ Date _____ (required if form is read to subject)

Consent form date (date you construct or revise the form)

**Exhibit 20.2** *Continued*

*Source:* Form provided by the Indiana University Bloomington Campus Institutional Review Board.

in research must be protected from the risk that information they give during a study could be released to outsiders where it might have damaging consequences. Two aspects of the privacy issue are anonymity and confidentiality. *Anonymity* refers to the process of protecting the identity of specific individuals.

No identification is attached to the data obtained; not even the researcher knows who contributed the data. *Confidentiality* refers to the process of keeping the information obtained from an individual during a study secret and private. There is broad professional agreement that all subjects have an inherent right to privacy. If the researcher does not need to collect the individual's name and other identifying information, it is recommended that the information not be collected. If it is necessary to collect the data for follow-up or other purposes, then it is the researcher's responsibility to provide secure storage for that information and to control access to it. In general, it is recommended that only the principal researcher and those staff members who must know individual data for work-related purposes have access to them. If students or colleagues are to be given access to the data or if the data are to be placed in an archive, then it is the responsibility of the principal researcher to remove all personal identifying information from those files.

For research funded by the Department of Education, the investigator must provide assurance for the secure storage and maintenance of all names and other identifying information. If the research project collects any personally incriminating or socially damaging information, this obligation is even more important. Failure to keep information of this type secure could lead to sanctions under the federal regulations (disqualification from further grant support) or to civil suits for personal damages.

Current regulations limit the federal regulations to those projects funded by grants from the Department of Health and Human Services and provide several broad **categorical exemptions** from the regulations. The five major categorical exemptions as set forth in the federal document *Protection of Human Subjects* (Public Welfare, 1993) are as follows:

> (1) Research conducted in established or commonly accepted educational settings, involving normal educational practices, such as (i) research on regular and special education instructional strategies or (ii) research on the effectiveness of or the comparison among instructional techniques, curricula, or classroom management methods. (2) Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures, or observation of public behavior, unless: (i) Information obtained is recorded in such a manner that human subjects can be identified, directly or through identifiers linked to the subjects; and (ii) any disclosure of the human subjects' responses outside the research could reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation. (3) Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures, or observation of public behavior that is not exempt under paragraph (b)(2) of this section, if: (i) The human subjects are elected or appointed public officials or candidates for public office; or (ii) federal statute(s) require(s) without exception that the confidentiality of the personally identifiable information will be maintained throughout the research and thereafter. (4) Research involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner

> that subjects cannot be identified, directly or through identifiers linked to the subjects. (5) Research and demonstration projects which are conducted by or subject to the approval of department or agency heads, and which are designed to study, evaluate, or otherwise examine: (i) Public benefit or service programs; (ii) procedures for obtaining benefits or services under those programs; (iii) possible changes in or alternatives to those programs or procedures; or (iv) possible changes in methods or levels of payment for benefits or services under those programs. (pp. 293–294)

Although these regulations exempt much of educational research from direct federal regulations, the IRBs remain in place, and institutions may elect to continue regulations that are more restrictive in character or broader in scope than the federal regulations. There has been some informal pressure for institutions to continue the broader regulations. The researcher should plan to consult his or her IRB to determine institutional rules and policies.

## SUMMARY

A research proposal presents the exact nature of the problem to be investigated and a detailed description of the methods to be used. It is the researcher's step-by-step plan for discovering new knowledge. It is at this stage that the researcher's inspiration and insights crystallize into concrete form. The following categories of information should be included in a quantitative research proposal.

First, a clear statement of the problem, accompanied by unambiguous definitions of terms, should be made early in the proposal.

Second, a review of pertinent literature should follow. A good review of literature shows what is so far known about the problem and lays the foundation for stating hypotheses regarding relationships between variables under consideration. In addition, this part should be written with the aim of providing a foundation for interpreting results.

Third, in the discussion of methods that follows the introductory section, methods for subject selection, methods of data collection, observational procedures, and measurement techniques should all be described with sufficient detail so that a reader could carry out the research by following the proposed steps exactly as the original writer of the proposal would.

Fourth, the next part of the research proposal describes the procedures to be used for data presentation, such as tables, figures, and charts, and introduces the statistical techniques that will be used for data analysis.

Fifth, a discussion of the potential importance of the study should follow. Here, the researcher should attempt to show how the findings will increase knowledge and what the results will mean to theory and research in the field of interest. A discussion of the applications of the findings to practice is helpful to readers who wish to assess the importance of the proposed research.

Sixth, the final section of the proposal contains the time schedule and estimated budget of the study. This information is useful to readers in making an overall evaluation of the proposal.

Proposals for qualitative research do not follow a fixed format, but they should provide general information about the purpose of the study, a rationale, a description of subjects and setting, the methodology and data collection procedures, and importance.

A matter of considerable weight in planning research is the observation of ethical standards. Participants must have the right of informed consent, they must be protected from harm, and their privacy must be respected.

## KEY CONCEPTS

anonymity
Belmont Report
categorical exemptions
confidentiality
deception
ethical standards
honesty
informed consent
institutional review board
privacy

## EXERCISES

1. What are the basic components of a quantitative research proposal?
2. How does a qualitative research proposal differ from a quantitative one?
3. At what point in the proposal should a clear statement of the problem be made?
4. Rewrite the following hypotheses, operationalizing all variables:
   a. Children who learn reading by the whole-language approach read better than those taught by a traditional approach.
   b. High school students who score above the top quartile of the XYZ Mechanical Aptitude Test make better mechanics.
   c. Scores on the math subtest of the SRA Achievement Test for bright seventh-grade students who have been instructed with the inductive math approach for 1 year will exceed scores of bright seventh-grade students who have been instructed with a textbook-centered approach.
5. What are some confounding variables that may affect differences of mean achievement scores between classes of the same grade level? How could you control for these variables in your proposed procedures?
6. What is the appropriate statistic for measuring correlation if the scales of the variables are nominal?
7. What measure of central tendency is appropriate for interval data?
8. For what types of data is ANOVA appropriate?
9. What is the function of an institutional review board?
10. How might some research results be affected by subjects' knowledge of the purpose of the research? How is the requirement of informed consent met in these circumstances?
11. What precautions should be taken to ensure confidentiality of responses and subjects' privacy in research projects?
12. List some common faults to avoid in preparing a quantitative research proposal.
13. Read about one of the infamous studies mentioned in this chapter (Milgram, Humphrey, Tuskegee, etc.) and list the ethical principles that were violated.
14. Which of the following proposed studies would *not* be exempt from federal regulations for research with human subjects?
   a. A study comparing two methods of teaching reading in the elementary schools of a district
   b. A survey of the attitudes of college freshmen toward the advising procedures at a university
   c. A history of the School of Education at Indiana State University
   d. A study of the effectiveness of Ritalin in the treatment of hyperactivity among elementary school children
15. Which of the following is not an ethical principle in research with humans?
   a. Keep the identity of individual participants and the results of their participation private.
   b. Obtain informed consent from all participants.
   c. Deceive participants about the nature and purpose of the research.
   d. Allow participants to withdraw from the study at any time.
   e. Explain the nature and value of the research to the participants.
16. In the middle of your research project, some participants decide they do not want to continue in the study. What should you do?
   a. Try to convince them to continue because of the importance of the study.
   b. Have them explain to you exactly why they do not want to continue.
   c. Agree to let them drop out of the study.
   d. Use what data you have already obtained from them.

**17.** Informed consent implies that researchers must inform participants about which of the following?
   **a.** The procedures to be followed in the study and any associated risks
   **b.** A detailed description of the purpose of the study
   **c.** The hypothesis(es) of the study
   **d.** The independent and dependent variables of the study

**18.** Distinguish between anonymity and confidentiality.

**19.** A school principal has given a graduate student permission to conduct a study in his school. Therefore,
   **a.** The student does not have to get approval from an institutional review board.
   **b.** The student will have to get consent from all the individuals taking part in the study.
   **c.** The student will need to inform and get the cooperation of teachers involved.
   **d.** Both b and c
   **e.** All of the above

## ANSWERS

**1.** Introduction, methodology, analysis of data, significance of the study, and time schedule and budget

**2.** A qualitative proposal is not as specific as a quantitative one because qualitative research evolves as the study progresses. The writing style is less formal and more personal than in a quantitative proposal.

**3.** In the first paragraph of the introduction

**4. a.** Third-graders who learned reading in first and second grade by the whole-language approach will score higher on the California Reading Test than third-graders who learned to read using a basal reading approach.
   **b.** Juniors and seniors who scored above the top quartile of the XYZ Mechanical Aptitude Test before becoming apprentices will be rated more highly by supervisors after 1 year in a mechanic's apprenticeship program than those scoring below the top quartile.
   **c.** Scores on the math subtest of the SRA Achievement Test of seventh-graders with IQs above 115 on the WISC who have been instructed with the inductive math approach for 1 year will exceed scores of similar students who have been instructed with the textbook-centered approach.

**5.** Different average ability levels, physical class environments, teachers, and types and amount of materials are some factors. One can control for some of these variables by pretesting for ability level, selecting classes with similar environments, and training teachers to certain levels of competence. Using large numbers of classes randomly assigned to conditions offers a different type of control.

**6.** Phi coefficient

**7.** The mean

**8.** Nominal-independent variables and interval-dependent variables

**9.** The institutional review board determines if the subjects in a proposed experiment are at risk under federal and institutional guidelines, determines if benefits outweigh the risks, and then grants or does not grant approval for the research to proceed.

**10.** Knowing that an unusual or experimental treatment is being used can influence subjects' psychological state and/or expectancy, which may detract from or otherwise influence the actual treatment effects. The subjects in such circumstances should be told that they will be informed of the purpose of the study when it is completed.

**11.** Subjects should not have to identify themselves unless necessary and should not be identified as individuals in the public report of the study unless they have given their consent.

**12.** Some possibilities include the following: The problem is trivial. The problem is not delimited. The objectives, hypotheses, or questions are too broadly stated. The procedures are lacking in detail.

A simple design is proposed for a complex problem. Relevant variables are not considered.

**13.** Answers will vary.
**14.** d
**15.** c
**16.** c
**17.** a
**18.** Anonymity ensures that an individual's name is not associated with the data obtained in a study; confidentiality refers to the practice of keeping information obtained from an individual private.
**19.** d

## REFERENCES

American Educational Research Association. (1992). Ethical standards of the AERA. *Educational Researcher, 21*(7), 23–26.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Bryant, M. T. (2004). *The portable dissertation advisor*. Thousand Oaks, CA: Corwin.

Humphreys, L. (1970). *The teahouse trade: Impersonal sex in public places*. Chicago: Aldine-Atherton.

Jones, J. H. (1982). *Bad blood: The Tuskegee syphilis experiment*. New York: Free Press.

Locke, L., Silverman, S., & Spirduso, W. (2007). *Proposals that work: A guide for planning dissertations and grant proposals*. Thousand Oaks, CA: Sage.

Marshall, C., & Rossman, G. (2006). *Designing qualitative research*. Thousand Oaks, CA: Sage.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology, 67*, 371–378.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research*. Washington, DC: U.S. Government Printing Office.

Public Welfare. (1993). *Protection of human subjects*, 45 C.F.R. 546.101.

Rowntree, D. (2004). *Statistics without tears: A primer for non-mathematicians*. Boston: Pearson.

CHAPTER 21

# Interpreting and Reporting Results of Quantitative Research

What you learn through research doesn't do the world any good unless you communicate it.

## INSTRUCTIONAL OBJECTIVES

After studying this chapter, the student will be able to:

1. List the principles for interpreting both anticipated and unanticipated results.
2. Distinguish between statistical and practical significance.
3. Explain the various circumstances that might result in the retention of the null hypothesis and discuss the implications of a retained null hypothesis.
4. Explain what the research findings mean in conceptual terms.
5. Write an appropriate title for a dissertation or research report.
6. Describe the nature of the content to be included in each section of a thesis or dissertation.
7. State the procedure to follow when preparing a research article for a professional journal.
8. State the procedure to follow when preparing a paper to be presented at a professional conference.
9. List points that should be included in a checklist for evaluating a quantitative research report.

Once the proposal has been accepted, the researcher can begin to collect and analyze the data. One point we must emphasize is that when hypothesis-testing research is set in motion, it is important that the study be carried out exactly as planned. This rule has ethical as well as practical implications.

To illustrate the ethical implications, suppose that Mr. Williams, a language teacher, has developed, with a great expenditure of time and effort, a system of teaching French that he believes is greatly superior to existing methods. To test the efficacy of this method, he establishes an experimental group taught by his method and a control group taught by

another method. He devises a series of weekly French achievement tests to serve as the dependent variable. Suppose he discovers in the first few weeks that the mean test scores for the two groups are almost identical. Having a major investment in his own method, he finds it difficult to believe that it is no better than the other, so he decides to sit in on the two classes to see what has "gone wrong." He discovers that the experimental group seems to show much greater knowledge and appreciation of French life and culture. Because he is determined to find a difference between scores, he decides to change his dependent variable to scores on tests on French life and culture.

Such a change would be unethical. Given two random groups, one can always find through chance alone a superiority in either group if one looks long enough. If the experimental group had not appeared superior on French life and culture, it might have been superior in verbal fluency, listening skills, on-task behavior, or some other variable. The language teacher must carry out the experiment as planned and not change the dependent variable. The investigator should report the evidence, suggesting a relationship between method and appreciation of French life and culture, but should make it clear to readers that this relationship was not hypothesized and therefore could easily have been a function of chance. It is unethical to abandon independent or dependent variables that do not seem to be "working out" or to add promising new ones.

Researchers are often tempted to add interesting new variables that crop up in their investigation. But adding new variables is unwise from a practical standpoint because such a tactic can confuse the results of a hypothesis-testing study and obscure the meaning of the results. More important, the theoretical base for interpreting these variables has not been laid. Thus, again the best advice to researchers is to leave new variables for later studies.

Of course, this caveat only applies to research in which predetermined hypotheses are being tested. In qualitative research, the hypotheses emerge as the study proceeds and are redefined when appropriate. Also, in descriptive research—such as surveys and trend analyses that do not begin with a hypothesis—there is no reason why a researcher should not add variables when appropriate.

## WRITING THE FINAL REPORT

Following data analysis, the researcher is ready to write the final report. A research project is of little value unless the findings can be communicated to others. The final report will tell what you did, what you found, and how your study is related to the body of knowledge in your area. Writing the final report is not as difficult a task as a beginning researcher might think. Much of the writing for the introduction, review of literature, and methodology has already been done in the initial proposal. Only minor revisions and a switch to the past tense should be needed on these three preliminary sections. The reader is referred to Johnson (2003) and Rudestam and Kjell (2007) for helpful suggestions on writing the report.

Because busy professionals will read the report, it should be as concise and as logically organized as possible. Anecdotes, stories of personal experiences, and argumentative discourse are out of place in a quantitative report. This does not mean the report must be dull and pedantic. If the researcher has approached the study with enthusiasm, this spirit tends to be conveyed between the lines.

Also, because the purpose of the report is to present the research rather than the personality of the author, keep the tone impersonal. In keeping with this, first-person pronouns are never used in quantitative reports, although they are acceptable in qualitative reports. Thus, you would *not* write, "I randomly assigned participants to the two treatment groups" but, rather, "Participants were randomly assigned to the two treatment groups." Despite a natural enthusiasm about the importance of the work, the author should not brag about it but should leave its evaluation to readers and to posterity.

A formal and uniform method of presenting research reports has evolved. Although at first glance these formalities may seem inhibiting, in practice they serve a useful purpose. It is important to have research reports arranged in such a way that readers know exactly where to find those specific parts they may be seeking. Otherwise, time is lost seeking relevant information. In addition, the use of an established format eliminates the need for devising one's own. As this topic is discussed, it will be seen that the established format follows logically the steps in a research project presented in previous chapters.

A research report may be presented as (1) a thesis or dissertation, (2) a journal article, or (3) a conference paper. A different reporting approach is used in each of these cases.

## THE THESIS OR DISSERTATION

Most universities have a preferred **style manual** that describes in detail the form the university requires. For students who are free to choose, several style manuals are listed at the end of this chapter. Once a manual has been chosen, the entire report should be styled according to its recommendations. You might also search your university library for copies of dissertations from your department and check the style and format.

The following outline lists the sequence and general components of a thesis or dissertation as described in most style manuals. Following the preliminary pages, the main body of the dissertation includes the introduction, literature review, methods, results, discussion, and conclusions and summary. Each of these sections is typically a chapter in the dissertation.

1. Preliminary pages
   a. Title page
   b. Abstract
   c. Acceptance page
   d. Acknowledgments or preface
   e. Table of contents
   f. List of tables
   g. List of figures

2. Text
   a. Introduction
      (1) The problem
      (2) The hypothesis(es)
      (3) Significance of the problem
      (4) Definitions of terms
   b. Review of literature
   c. Methods
      (1) Participants (subjects)
      (2) Procedures
      (3) Instruments
   d. Results
      (1) Presentation of data
      (2) Analysis of data
   e. Discussion
      (1) Interpretation of findings
      (2) Implications
      (3) Applications
   f. Conclusions and summary
      (1) Conclusions
      (2) Summary
3. Supplementary pages
   a. References
   b. Appendixes
   c. Vita (if required)

### Preliminary Pages

The preparation of the preliminary pages is largely a matter of following the rules of the style manual. However, one aspect of these pages that needs additional explanation at this point is the title of the study.

The title should describe, as briefly as possible, the specific nature of the study. A rule of thumb states that a title should have no more than 12 to 15 words. For example, consider (1) a study of culturally disadvantaged children that compares the reading readiness of those who have participated in a Project Head Start program with that of a matched group of children with no formal preschool experience, and (2) the title, "A Comparison of Reading Readiness Test Scores of Disadvantaged Children Who Have Attended Head Start Classes for Six Weeks or More with Similar Children with No Preschool Experience." Although this title does convey what the study is about, it is too long. Such phrases as "a comparison of," "a study of," and "an investigation into" are usually superfluous. Furthermore, most prospective readers will know that Project Head Start is a preschool experience designed for culturally disadvantaged children. However, to go to the other extreme by providing a title that is too brief or too vague to convey the nature of the study is a much more serious mistake. With vague or overly brief titles, a prospective reader must search out the article to determine what it is about. Titles such as "Head Start and Readiness" or "Reading among the Disadvantaged" illustrate this shortcoming. *The title should identify the major variables and the populations of interest.* The operational definitions of the major variables and the

description of the samples need not be included in the title. Because correct titling will ensure correct indexing, a useful strategy is for researchers first to decide under what key words they want their studies to be indexed, working from there to a concise title. In addition, the title should, if possible, begin not with an article ("a," "an," or "the") but with a key word. Any fanciful part should be relegated to subtitle because bibliographers often cut off the subtitle or alphabetize under the article ("the" or "a"), ensuring real confusion or loss to searchers. In our example, the important key words for indexing would be *reading readiness* and *Project Head Start*. Therefore, an appropriate title might be "Reading Readiness of Project Head Start and Non-Head Start Children." This title is reasonably brief, yet it gives the prospective reader a fairly precise indication of what the study is about.

Avoid at all costs emotion-laden titles, such as "We Must Expand the Head Start Program" or "Don't Let the Disadvantaged Become Poor Readers." The prospective reader will not expect research findings under such titles but, rather, armchair articles attempting to sell a point of view.

*Abstract* Most institutions require a separate abstract of the dissertation, which should include a precise statement of the problem and concise descriptions of the research methods, results, and conclusions. The abstract must be limited in length (typically 500 words or less). The abstract follows the title page.

## MAIN BODY OF THE DISSERTATION

We comment only briefly on the chapters making up the main body of the report because you will already have organized the ingredients of these sections in your proposal.

### Introduction

The introduction is a very important part of the dissertation; it will take up a good portion of the total length of your report. The introduction begins with a statement of the research question (problem). Give the background of the problem and state why you think your study will make a contribution to knowledge in the area. State the hypothesis and the reasoning that led to your expectation about the results of the study. Define any terms that might be unfamiliar to readers.

### Review of Literature

This chapter contains an extensive review of the literature related to your problem. Do not just list studies one after the other but, rather, synthesize their findings and point out agreements and disagreements among them. Also, show how they are related to your research problem.

### Methods

This chapter presents a detailed description of the methodology. It should be clearly written and should provide enough information that another researcher could read this section and replicate your study. In a quantitative study, this chapter typically has subsections with information on the participants, the research design, the variables and treatments, materials used to collect data, procedures, and the location of the study. In the section on participants, for example, tell the number of participants, how they were selected, and their major characteristics

such as age, gender, and race/ethnicity. In this chapter, one also needs to provide information on the validity and reliability of the measures used.

### Results

The results chapter follows methodology and presents the outcomes of the statistical analyses of the data. One generally reports descriptive statistics first followed by any inferential statistics. A recommended technique for presenting the results is to organize findings around the hypotheses; that is, the researcher restates the first hypothesis and presents findings concerning it, then repeats this procedure for each hypothesis in turn. You will report whether your results are statistically significant or not statistically significant. Remember that statistical significance does not indicate the size of the effect. You may need to calculate and report effect size, which indicates the strength of the relationship. Reporting effect size enables your readers to compare your findings with others' research using the same variables. You do not interpret or discuss the results at this point but, rather, merely present the findings. This chapter of the dissertation is thus relatively brief.

To facilitate comprehension, tables and figures are usually used to present the findings. They present the numbers and statistics more clearly and more concisely than is possible if the same information is presented only in text form. A *table* shows the quantitative data such as means and standard deviations organized in rows and columns. A *figure* shows the data in diagram or graphical form. The type of figure you use (bar graph, histogram, polygon, scatterplot, etc.) depends on the kind of data you have and what you want to show. Well-constructed tables and figures should "stand alone" so readers can understand them without having to refer to the text.

The first table in the report usually summarizes the descriptive statistics, such as means, standard deviations, correlations, percentages, and so on. Later tables present the results of applying inferential statistics and tests of significance to the data. Name the calculated statistic, give the degrees of freedom, and give the probability level at which the statistic was significant (or not significant). Use a summary table, for example, to present the results of an analysis of variance. Most style manuals, such as the *Publication Manual of the American Psychological Association* (American Psychological Association, 2001), provide examples of commonly used types of tables and figures and instructions for their construction. Nicol (1999, 2003) also provides suggestions for creating tables and figures. Computer software has greatly simplified the task of preparing tables and figures for research papers.

When writing the results section, the researcher refers to each table and each figure by number and comments only on the most important and interesting finding in each. Use present tense when pointing out the significant aspects of a table or figure. The following example was taken from Springer, Pugalee, and Algozzine (2007), who investigated the use of a computerized tool on improving the math skills of students who had failed the state competency test [Arizona Instrument to Measure Standards (AIMS)]:

> Table 2 shows means, standard deviations, and analysis of variance summary table for AIMS scaled scores. Significant differences ($t = -2.66$, $df = 26$, $p < .05$) were evident for students participating in the AIMS program ($M = 498.50$) and their peers in the control group ($M = 477.71$); the effect sizes for this comparison were large (.88–1.00). Significant occasion and group by occasion interaction (see Fig. 1) indicated outcomes favoring the treatment group. Pretest to posttest improvements

were significant for treatment group students ($t = -6.48$, $df = 13$, $p < .05$), but not for other students ($t = -1.39$, $df = 13$, $p > .05$). Of the 14 students who attended the AIMS class, eight (57 percent) passed the test whereas only two of the 14 (14 percent) control students passed the test. (p. 41)

## Discussion

In this chapter, you are ready for the discussion of the findings as they relate to the original question and hypothesis. This is the place to discuss the practical and/or theoretical implications of the findings. You might point out how your findings relate to previous research and suggest additional research that should be done. The following presents an excerpt from Springer et al.'s (2007) discussion of their overall results:

> In this study, we examined the effects of using a specialized remedial program grounded in evidence-based effective instructional practices (increased practice of critical skills, with continuous monitoring of progress and constant adjustment based on performance). Overall, students who participated in the experimental treatment demonstrated greater gains in overall performance in mathematics achievement as measured by their state's critical competency test required for high school graduation. The program provides an instructional management system that allows students to spend more time reviewing concepts and skills to enhance their performance resulting in positive growth in measured mathematics competence.
>
> The requirement to pass the AIMS math test is likely to create a crisis unless something is done to increase the percentage of students passing the test. The AIMS math skills improvement course significantly increases the number of students who pass the test after initially failing it. (p. 42)

A beginning researcher may find the discussion the most difficult but also the most rewarding to write. This chapter is difficult because there is no standard format for the content; you must use insight and original thinking to provide an explanation and interpretation of the results that you presented previously. Be ready to provide the answer to the original research question. First, discuss how the results support or fail to support the hypotheses of the study. In this *interpretation,* you must deal not only with expected results but also occasionally with unexpected or negative results. Here are some guidelines for interpreting various results.

*Interpreting Expected Results* Researchers are understandably pleased when the results of a study fit into the previously constructed framework and interpretation can proceed as expected. The study has "worked," and there is agreement between rationale and results. Only a few words of caution need apply in **interpreting expected results:**

1. *Do not make interpretations that go beyond the information.* This injunction may seem patently obvious, but researchers often get so excited when results are as expected that they draw conclusions that do not have a valid basis in the data. Even published research sometimes offers more interpretations than the data warrant.
2. *Do not forget the limitations of the study.* These limitations, of course, should have been previously identified in the study—limitations inherent in the

less-than-perfect reliability and validity of the instruments, limitations caused by the restriction in sampling, the internal validity problems, and so forth.

3. *Ethics require that the researcher report internal validity problems that could account for the results.* If, despite the researcher's best efforts, the nonexperimental variables were particularly benign for the experimental group and those for the control group were particularly malign, these conditions must be reported and taken into account in interpreting results. (For example, despite random assignment of teachers to groups, the experimental group may have mostly experienced teachers and the control group may have mostly inexperienced teachers.)
4. *Remember that statistical significance means only that for the appropriate degrees of freedom, the results are unlikely to be a function of chance.* **Practical and statistical significance** have very different meanings. Statistical significance does not mean that the results are significant in the generally accepted meaning of the word—that is, important, meaningful, or momentous. Do not assume that statistical significance guarantees momentous import to your findings.

Suppose that two equivalent groups have been subjected to two different systems of learning spelling during a 2-year period. Those using system A show a mean gain equivalent of 2.15 years of growth on standardized tests during the experiment, whereas those using system B show a gain of 2.20 in the same period. If the groups are large and/or if the differences within groups are small, the differences between the means would be statistically significant. However, a difference of half a month over a 2-year period is relatively meaningless in practical terms.

If system B is more expensive in terms of student time, teacher time, or materials, teachers would be unwise to adopt it simply because it produced statistically significantly greater gains than system A. If, in contrast, system B is less expensive, teachers would be inclined to favor it because its results are so similar to those of system A in practical terms.

The potential importance or meaningfulness of results must be established in the proposal before the study begins. A study is not important if it does not add meaningful information to the existing body of knowledge, no matter how statistically significant the results may be.

*Interpreting Negative Results* Researchers who find **negative results,** results opposite to those hypothesized, often develop sudden revelations concerning the shortcomings of their study. Their interpretation of results reads like a confession. The instruments were inadequate for measuring the variables involved, the sample was too small and was so unrepresentative that results cannot be validly generalized to a meaningful target population, and so on. Hindsight reveals internal validity problems that explain why the study did not come out as it "should have." Of course, any or all of these things could be true, and the shortcomings of any study should be reported no matter what the results. However, research is always a venture into the unknown, so there is no ultimate "should be." An investigator predicts the expected results of a study on the basis of theory, deduction, experiences, and the results of previous research. If these are so conclusive that there can be absolutely no doubt as to the results of this study, then the study is pointless in the first place.

When you undertake a study, you implicitly state that the outcome is a matter of conjecture, not a matter of certainty. When you complete your proposal, it is understood that you declare you will impartially seek to determine the true state of affairs with the best instruments and procedures available to use for that purpose. Therefore, you are obliged to accept and interpret your data no matter how the data stand. When the results contradict the theoretical rationale of the study, the discussion section of your report should include a reconsideration of the original theory in light of the findings. Researchers are often reluctant to present and interpret data that conflict with previous research or with well-established theory. However, it may be that their results are correct and previous results are wrong. The progress of the science of education will be retarded if investigators are reluctant to report findings that disagree with those reported in earlier studies. Contradictory results indicate that a question is not settled and may stimulate further research. Additional research or theory formation may eventually reconcile seemingly contradictory results. Theory is tentative and should not deter investigators from giving a straightforward interpretation of what was found.

Reconsideration of the theoretical base of a study belongs in the discussion section. One must not go back and rewrite the related literature and hypothesis sections of the report.

*Interpreting Results When the Null Hypothesis Is Retained* Because a null hypothesis may be retained for a variety of reasons, **interpreting a retained null hypothesis** can be particularly difficult. A retained null hypothesis may occur for the following reasons:

1. The null hypothesis is, in fact, true. There may be no relationship between variables. The experimental treatment may be no more effective than the control treatment.
2. The null hypothesis is false, but internal validity problems contaminated the investigation so badly that the actual relationship between variables could not be observed.
3. The null hypothesis is false, but the research design lacked the power to reject it.

Any of these states of affairs may be the case, but the investigator does not know which is true and therefore should not claim any one of them as the explanation for the results. One may suggest additional research, planned in such a way as to avoid the internal validity problems encountered, but still one must report a retained null hypothesis as lack of evidence and no more.

It is incorrect to present a retained null hypothesis as evidence of no relationship between variables. A retained null hypothesis must be interpreted as lack of evidence for either the truth or falsity of the hypothesis. A widely used toothpaste commercial stated that tests showed a particular toothpaste was unsurpassed in reducing tooth decay. Interpreting the term *unsurpassed* to mean "no significant difference," we can imagine a test in which a very small number of subjects were used and/or numerous internal validity problems were present. If a retained null hypothesis is the desired result of an experiment, it is remarkably easy to arrange for such an outcome.

Recall from Chapter 7 that the *power* of an experiment refers to the statistical ability to reject a null hypothesis when it is, in fact, false. This power is a function

of the size of the sample, the heterogeneity of subjects with reference to the dependent variable, the reliability of the measuring instruments used, and the nature of the statistical procedure used to test the hypothesis, as well as effect size. Researchers should take these factors into account when planning an experiment. The power of an experiment should be considered in planning the study. It must not be brought in at the end of a study to explain away lack of statistical significance. For example, one should not say, "The results would have been statistically significant if the sample had been larger."

With rare exceptions, the only legitimate interpretation of a retained null hypothesis is that "sufficient evidence for a conclusion has not been observed." If you are studying a small population and can do a complete census of that population, a retained null hypothesis can legitimately be interpreted as a lack of relationship between variables within that particular population. A retained null hypothesis also acquires credibility if you can show that the study was free of internal validity problems and had sufficient power to reject the null hypothesis with a relatively trivial effect size.

*Interpreting Unhypothesized (Serendipitous) Relationships* We previously emphasized that a researcher should not abandon a hypothesis during the conduct of a study in order to pursue more promising avenues that present themselves during the course of the study. This does not mean that you should ignore any **unhypothesized relationships** that you may observe in conducting a study. On the contrary, you should record and analyze them with the same rigor you employ in pursuing hypothesized relationships. Throughout the history of science, such **serendipitous discoveries** have often proved important. However, such findings should always be viewed with more suspicion than findings directly related to the hypothesis because there is a relatively great possibility that a spurious unhypothesized relationship will appear in a study. Such relationships should be reported, but they should be considered incidental to the main thrust of the investigation. Before they can be employed as the basis for conclusions, they should be made the subject of a later study specifically designed to investigate them.

*Implications/Application* It is very important that the researcher point out the specific *implications* of the research. The results may support or not support relevant theoretical positions. You may explain how the theory should be modified and may suggest further studies that would logically follow. Also, discuss how the results fit in with previous research findings. Lastly, give some attention to stating the possible *application* of the findings to educational practice. The sections on implications and applications of the results are often not sufficiently developed because the writer assumes these will be as obvious to the reader as they are to the investigators. In fact, in the conduct of the study the investigators probably gained insights into the problem that are deeper than those most of their readers can be assumed to have. Therefore, researchers' interpretations should be more meaningful than those that readers might make for themselves.

### Conclusions and Summary

The conclusions and summary sections together form the capstone of the report.

*Conclusions* In the conclusions section, the researcher tells what the research findings mean in conceptual terms. The conclusions indicated by the research findings should be limited to those that have direct support in the research findings. Researchers are often tempted to conclude too much. The hypotheses provide a convenient framework for stating conclusions; that is, indicate in this section whether or not the findings support the hypotheses. It is important to distinguish between *results* and *conclusions*. Results are direct observations summarized and integrated by the statistical analysis. A conclusion is an inference based on the results, expressed in terms of the study's hypothesis, such as one group's treatment being more effective than the other group's treatment. For example, a study might result in the observation that the mean spelling test scores of students taught spelling by method A are significantly higher than the mean of students taught by method B. The conclusion that method A is more effective than method B is not a direct result of the study but, rather, is an inference based on the results of the study.

In the Springer et al. (2007) study mentioned previously, the researchers concluded, "Our current study is important because the results extend findings of previous research by demonstrating the potential of the program to influence the performance of students who have previously failed a high school exit examination. We strongly recommend that high schools implement this approach" (p. 43).

*Summary* Because the summary will be more widely read than other sections of the report, its wording must be particularly clear and concise. The summary usually includes a brief restatement of the problem(s), the main features of the methods, and the most important findings. On completing a draft of this section, check it carefully to determine whether it gives a concise but reasonably complete description of the study and its findings. Also check to ascertain that no information has been introduced here that was not included in the appropriate preceding sections. It is a good idea to have a colleague read the conclusions section to determine if you are communicating as well as you intended.

## THINK ABOUT IT 21.1

A study was conducted to determine what high school data provided the best predictor of college freshman students' success as measured by grade point average (GPA). The study looked at total high school GPA; GPA in only "academic" courses such as English, math, science, history, foreign language, and the like; rank in high school class; and SAT scores. Interpret the following table that was included in the report of this study.

**Table 1** Correlation Matrix for Variables in a Prediction Study

| | Freshman GPA | Total HS GPA | Academic HS GPA | SAT |
|---|---|---|---|---|
| HS GPA (total) | .52 | | | |
| Academic GPA | .57 | .85 | | |
| SAT | .44 | .43 | .41 | |
| HS rank* | −.54 | −.68 | −.77 | −.37 |

*Highest rank is 1.

**Answer**

Table 1 shows that the single variable having the highest correlation with freshman college GPA is the high school GPA based on academic courses only (.57). The next highest correlation (−.54) is between rank in high school class and freshman GPA. The correlation between total high school GPA and college freshman grades (.52) is just slightly lower. SAT scores have a lower correlation with college GPA than do the other high school measures.

### Supplementary Pages

*References* The reference list must include all sources mentioned in the text. Most universities insist that *only* these be listed, but a few ask that pertinent references not specifically mentioned also be listed. The style manual previously selected will give complete details on the method of listing references. It is important to follow these rules rigorously and completely. In fact, it is good strategy to learn them before carrying out the search through the literature for the proposal. By listing each reference in the correct form as it is encountered, you can avoid the extra time involved in finding the references again in order to have them in complete form for the bibliography. List them on cards or enter them in your computer so you can file them in alphabetical order.

*Appendixes* The appendixes contain pertinent materials that are not important enough to include in the body of the report but may be of value to some readers. Such materials may include complete copies of locally devised tests or questionnaires together with the instructions and scoring keys for such instruments, item analysis data for measurements used, verbatim instructions to subjects, and tables that are very long or of only minor importance to the study.

*Vita* The authors of research reports are sometimes asked to include brief accounts of their training, experience, professional memberships, and previous contributions.

## THINK ABOUT IT 21.2

Critique each of the following titles of quantitative research articles taken from professional journals:

1. "Persuasion as a Dynamic, Multidimensional Process: An Investigation of Individual and Intraindividual Differences."
2. "Are We Creating Separate and Unequal Tracks of Teachers? The Effects of State Policy, Local Conditions, and Teacher Characteristics on New Teacher Socialization."
3. "Can Anything Good Come from Nazareth? Race, Class, and African American Schooling and Community in the Urban South and Midwest."
4. "The Alchemy of the Mathematics Curriculum: Inscription and the Fabrication of the Child."
5. "The Chilly Climate: Fact or Artifact?"

**Answers**

1. The title is too vague to convey clearly what the study is about. What individual and intraindividual differences are being investigated?
2. The title is a little long; the first part of the title seems loaded and could be eliminated. It would be better to start with "The Effects of…."
3. The first part of the title seems irrelevant and could be eliminated without sacrificing clarity.
4. The title is vague and a bit flowery. We may know that the study has to do with mathematics, but we do not know the variables being investigated.
5. The title is entirely too brief! One would have no idea what the study is about or what population was involved.

## THE JOURNAL ARTICLE

We suggest that graduate students submit reports of their research to professional journals for publication. The first task would be to decide which journal to use. There are specialized journals in reading, math, social studies, special education, etc. and also more generalized journals, such as the *Journal of Educational Research,* the *American Educational Research Journal,* and the *Journal of Educational Psychology.* Some journals publish quantitative research, some qualitative, and others accept both. A good first step is to look at your bibliography to determine which journal has published the greatest amount of work in your area of interest.

Information concerning the procedure for submission of manuscripts is usually on the inside of a journal's front cover or can be found on the journal's website. Many journals now accept electronic submissions. Most journals specify which style manual should be used (e.g., the *Publication Manual of the American Psychological Association,* fifth edition, 2001). If a manual is not specified, you can determine the preferred style, method of referencing, and so on from looking at the articles included in a recent issue of the journal. A research article submitted to a journal follows the same general outline as a dissertation but must be much shorter. Brevity is essential if you hope to get the article accepted for publication. A thesis or dissertation functions to demonstrate a student's competence and requires a full setting forth of the research question, related literature, methodology, results, and reflective elaboration. The journal article, in contrast, requires mainly communication of the author's contribution to knowledge. For the sake of economy of journal space and readers' time, the article must be concise. The journal article generally consists of an abstract (100- to 150-word summary of the study), introduction, methods, results, discussion, and list of references. The various sections contain basically the same content as the sections in the dissertation, except that only essential material is included. The results section is probably of greatest interest to the reader and thus represents a greater proportion of the article than it would in a dissertation. One might finish with recommendations for further research.

The main text is followed by a list of references actually cited in the journal article. You also provide any tables and figures on separate pages at the end of the article and indicate where they should be placed in the text.

A brief cover letter should accompany the manuscript. The editor usually sends the author a postcard acknowledging receipt of the manuscript and circulates copies of it among the appropriate members of the editorial board for review. From this point considerable time usually elapses before the author is informed whether the article has been accepted (6 weeks is probably typical). After an article is accepted, it is usually many months (6 to 12) before it appears in print. When a manuscript is rejected by a journal, the rejection notice is usually accompanied by a statement of the reasons for this rejection. A rejection by one journal does not necessarily mean that the article is unworthy of publication. A number of factors—such as competition for space, changes in editorial policy, or bias of reviewers—can influence the decision on publication. An article that has been rejected by one journal may be revised and submitted to another. It is wise to make use of the reviewers' comments in the revision. Many articles make the rounds of several journals before finding a home. It is not ethical, however, to submit an article to more than one journal at a time.

## THE PROFESSIONAL CONFERENCE PAPER

Because of the great lapse of time between submitting a journal article and its appearance in print, many researchers like to present their research at a professional conference before submitting it to a journal. It is often easier to get a paper accepted for a conference than it is to get it published in a journal. Your professor could give suggestions about upcoming conferences in your field. For example, the American Educational Research Association and the National Council on Measurement in Education have a joint conference each spring.

You will need to find out the requirements for submitting a presentation proposal. This information is usually included in the association's journal well in advance of the conference. You may submit your proposal either in hard copy or online. You will find that some organizations now only accept online proposals.

Papers presented ("read") at professional meetings are prepared in much the same manner as journal articles. They are not necessarily always reports of completed research but may be progress reports of ongoing projects. Although at some conferences papers are read, we believe it is much more effective to present the paper from notes rather than to read the content word for word. It is more interesting and the audience generally reacts more positively to a paper that is presented rather than read. If the audience has a copy of the paper in hand, they can read the paper silently more rapidly than the author can read it aloud; thus, there may be a gap in attention. It is common now for presenters to use PowerPoint presentations at conferences.

A conference paper is less formal than a journal article and can usually be more precisely geared to its audience. The audience can generally be expected to be familiar with details of related research and methods of measurement.

The paper is frequently organized as follows:

1. Title
2. Abstract
3. Direct statement of the hypothesis

4. Brief description of the procedures
5. Findings, conclusions, and implications

The time allowed for presenting a paper is usually quite brief, frequently less than 30 minutes. Therefore, the paper should focus on the most important aspects of one's study.

Many attendees will want the presented paper for later reference. Rather than taking hard copies of the paper to distribute as was done in the past, presenters now simply provide a web address for the attendees to check and download the paper if they wish. Some presenters even distribute a CD or DVD rather than a paper copy.

### POSTER SESSION PRESENTATIONS

Professional conventions often include poster sessions. Poster presenters are assigned a bulletin board (typically $4 \times 8$ feet) in a designated room and a designated time frame, typically 1 to 1½ hours. The author mounts the presentation on the bulletin board and stands by it during the designated time. Those interested in the presentation will inspect it and may ask the presenter about it. Only a brief review of the study can fit onto the bulletin board, so only the abstract and highlights can be presented in this format. The presenter should have available the details needed to answer any questions that may be asked. Presenters may have hard copies of the complete report for those who ask. Technology is changing poster presentations as well. Now some have laptop displays, distribute DVDs or CDs with their information, or provide cards with URLs to access.

## CHECKLIST FOR EVALUATING QUANTITATIVE RESEARCH REPORTS

The following checklist should be useful for evaluating your own quantitative research reports and the reports of others. It brings together many of the topics presented in this text.

*Title*

Is the title brief but informative?

Does the title make clear the population of interest and the major variables?

Have vague, ambiguous, and emotion-laden terms been avoided?

*Statement of the Problem*

Have the variables of interest been identified?

Has the theoretical foundation for the study been developed?

Is the problem clearly stated?

Is there a justification or rationale for the study?

*Related Literature*

Is the related literature relevant and sufficient?

Are the connections between the present study and the previous research and theory made clear?

*Hypotheses*

Are the hypotheses explicit?

Do the hypotheses follow logically from the statement of the problem?

*Participants*

Is the population of interest defined?

Is the method for selecting the sample explicit?

Does the sampling allow for generalization to the population of interest?

*Procedures*

Are the procedures described well enough to allow replication of the study?

Do the procedures include appropriate operational definitions of the independent and dependent variables?

Do the procedures provide sufficient control for internal validity?

Do the procedures provide sufficient control for external validity?

*Instruments*

Are the instruments adequately described?

Is information on the validity and reliability of the instruments provided?

Are the instruments appropriate operational definitions of the dependent variables?

*Analysis of the Data*

Are the descriptive statistics used appropriate for summarizing the data?

Are the inferential statistics used appropriate for testing the hypotheses?

Are the statistics appropriate for the level of measurement of the data?

*Results*

Are the results of all hypothesis tests presented?

Have effect sizes been reported?

Are the results clearly presented?

Are tables and figures used appropriately?

*Discussion*

Is there a clear interpretation of the results?

Does the author present implications of the study results?

Are the implications presented based on the results of the study, not on what the author hoped or expected to be true?

Are appropriate applications discussed?

Do the applications follow logically from the results of the study?

Are the connections between the results and theory and existing literature shown?

Are there suggestions for future research?

*Conclusions*

Are the conclusions clearly presented?

Do the conclusions follow logically from the results of the study?

Has the author avoided reaching conclusions that are not directly supported by the outcomes of the study?

*Summary*

Is the summary clear, concise, and sufficiently complete?

The reader is referred to Mitchell and Jolley (2006) for a helpful checklist for the evaluation of research articles.

## STYLE MANUALS

The following are widely used manuals detailing general form and style for theses and dissertations:

Amato, C. J. (2002). *The world's easiest guide to using the APA: A user-friendly manual for formatting research papers according to the APA style guide* (3rd ed.). Corona, CA: Stargazer.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

American Psychological Association. (2008). APA-StyleHelper 5.1 software. Conforms to the latest standards of the fifth edition of the *Publication Manual* of the APA. Please note: StyleHelper 5.1 is *not* compatible with Microsoft Word 2007.

American Psychological Association. Free online style tips at www.apastyle.org.

Gelfand, H. (2002). *Mastering APA style: Student workbook and training guide.* Washington, DC: American Psychological Association.

Gibaldi, J. (2006). *MLA handbook for writers of research papers.* New York: Modern Language Association.

Rudestam, K. E., & Kjell, E. (2007). *Surviving your dissertation: A comprehensive guide to content and process.* Thousand Oaks, CA: Sage.

Slade, C. (2007). *Form and style: Research papers, reports, theses.* Boston: Houghton Mifflin.

Strunk, W., Jr., White, E. B., & Kalman, M. (2007). *The elements of style.* New York: Penguin.

Turabian, K. (1996). *A manual for writers of term papers, theses, and dissertations* (6th ed.). Chicago: University of Chicago Press.

University of Chicago Press. (2007). *The Chicago manual of style.* Chicago: Author.

## SUMMARY

Researchers communicate the results of their investigations in written reports. They communicate what question was investigated, why and how it was investigated, and what was found. The research report may take the form of a thesis or dissertation, journal article, or presentation at a professional meeting.

The main text of a report of quantitative research includes introduction, methods, results, discussion, and conclusions/summary. Following the text is a list of references that indicates *all* the sources cited in the report and *only* those sources. The researcher follows standard guidelines for writing each of these sections. Probably the most difficult section to write is the discussion because here the researcher must explain the findings. Interpreting the results of a study is a straightforward task if, in the proposal, the researcher has laid a proper foundation for the research. The following cautions should be kept in mind:

1. Interpretation should be strictly based on the data derived from the study.

2. Internal and external validity problems and other limitations of the study should be considered.
3. Conclusions must be presented as probability statements rather than as facts.

Negative results deserve the same respect and interpretation as do positive results. A retained null hypothesis is interpreted as indication of insufficient evidence and no more. Unhypothesized results deserve attention as sources of future hypotheses.

Formal procedures have been developed for preparing dissertations, journal articles, and papers. Researchers need to familiarize themselves with the style manual required at a university or journal and follow it in the preparation of the final manuscript. The most popular format for educational research reports is the APA's *Publication Manual* (2001).

## KEY CONCEPTS

interpreting expected results
interpreting retained null hypothesis
negative results
practical and statistical significance
serendipitous discoveries
style manual
unhypothesized relationships

## EXERCISES

**1.** Indicate in which section of a journal article reporting quantitative research (introduction, methods, results, or discussion) the following statements would belong:
   **a.** "Participants were first-grade students in three schools located in the same rural district of a state in the Far West. Students were predominantly lower- to lower-middle-class Anglo Americans, with a few Hispanics."
   **b.** "The purpose of this research is to examine the effects of implementing two reading programs in comparable first-grade classes during one school year."
   **c.** "All first-graders were pretested on the Peabody Picture Vocabulary Test and the Woodcock Reading Mastery Test."
   **d.** "Students using program A tended to perform least well relative to program B students on Passage Comprehension, while surpassing them on Word Attack. One possible explanation lies in the students' previous experiences with the tasks required by the different reading comprehension testing formats."
   **e.** "Results of the ANOVA performed on the overall sample ($n = 184$) were significant, $F(4, 176) = 5.89$, $p < .001$."
   **f.** "There is a need to continue this study over time, replicate it elsewhere, and to experiment with combinations of the two reading programs and other strategies."

**2.** Indicate whether the format of the following statements written for a quantitative journal article is correct or incorrect. If incorrect, tell why and rewrite the statement using acceptable format.
   **a.** "I administered a pretest to both groups of subjects."
   **b.** "The means of the pretest and posttest for the two groups are summarized in the following table and the interaction is shown in the graph below."
   **c.** "The experimental group will be shown a series of films designed to change attitudes toward smoking."
   **d.** "Recently, a study at Indiana State evaluated the Reading Recovery Program in comparison to three alternative programs in four Indiana school districts."

**e.** "Groups of mothers were randomly assigned to one of four interview discussion groups ($n = 5$) held in a quiet setting within the school."

**3.** What is the difference between results and conclusions?

**4.** Explain the difference between statistical significance of the results and the significance of the study.

**5.** What states of affairs can lead to a retained null hypothesis in a quantitative study?

**6.** Decide whether each of the following titles is acceptable or unacceptable for quantitative research and give reasons for your choices:
- **a.** Grade-Point Average and Driver Education
- **b.** The Effects of Individualized Tutoring by Sixth-Grade Students Three Times a Week on Reading Performance of Below-Average Second-Grade Readers
- **c.** Children Should Be Taught through Discovery Learning!
- **d.** Relationship between Personality Characteristics and Attitudes toward Achievement of Good and Poor Readers

**7.** What are the differences in format for research reported in dissertation form, in journal form, and in a paper to be read at a conference?

**8.** Should one discuss research results that do not agree with one's hypothesis?

**9.** For the following situations, indicate whether you would report your research in a journal article or in a paper at a professional meeting:
- **a.** A progress report of an incomplete but ongoing research project
- **b.** Your goal is to communicate to the widest possible audience in the field.
- **c.** You wish to avoid the time lag in communicating results.

**10.** A researcher reported the following: "Achievement test scores for 36 subjects working under tangible incentives were mean = 72.0 and standard deviation = 10.9 on test 1 and mean = 76.0 with an SD = 9.7 on test 2. Combined average = 74.0, SD = 10.3. Scores for 34 subjects working under intangible incentives were mean = 68.0 with SD = 8.5 on test 1 and mean = 70.0 with SD = 10.2 on test 2. Combined average for intangible incentive group = 69.0 with SD = 9.3." Show how the researcher could use a table to present these findings so that they could be more readily comprehended by readers.

**11.** Select a journal to which you might want to submit an article and describe its requirements for submission.

**12.** In the results chapter of a dissertation, the writer presents
- **a.** Summary of data collected
- **b.** Statistical analyses
- **c.** Interpretation of findings
- **d.** All of the above
- **e.** Both a and b

**13.** In a dissertation, the abstract follows the
- **a.** Title page
- **b.** Appendix
- **c.** Summary/conclusions
- **d.** References

**14.** In which chapter of a dissertation would you point out any limitations of your research study's methodology?
- **a.** Introduction
- **b.** Methods
- **c.** Results
- **d.** Discussion
- **e.** Summary

**15.** Considering the true state of affairs, indicate whether the researcher was correct or incorrect in the following interpretations of the data:
- **a.** The data analysis showed that the results were statistically significant ($p < .01$). The researcher rejected the null hypothesis and accepted the alternative. (In reality, there is a genuine difference between the groups.)
- **b.** The results were statistically significant ($p < .001$). The researcher rejected the null hypothesis and accepted the alternative. (In reality, there is *no* difference.)
- **c.** The results were not statistically significant. The researcher retained the null hypothesis. (In reality, there is *no* difference between the groups.)
- **d.** The results were not statistically significant, and the researcher failed to reject the null hypothesis. (In reality, there is a difference between the groups.)

## ANSWERS

1. **a.** Methods
   **b.** Introduction
   **c.** Methods
   **d.** Discussion
   **e.** Findings
   **f.** Discussion
2. **a.** Incorrect, personal pronouns are not used in quantitative research reports: "A pretest was administered to both groups of subjects."
   **b.** Incorrect, tables and figures are always referred to by number: "The means of the pre- and posttests for the two groups are summarized in Table 1 and the interaction is shown in Figure 1."
   **c.** Incorrect, use past tense when writing the methods section: "The experimental group was shown a series of films designed to change attitudes toward smoking."
   **d.** Incorrect, no name or date is given for the reference to the research reported: "Smith and Green (1999) evaluated the Reading Recovery Program in comparison to three alternative programs in four Indiana school districts."
   **e.** Correct
3. A result is a direct observation. A conclusion is an inference based on results.
4. Statistical significance means only that the results are not likely to be a function of chance; the significance of the study is determined by the importance of the findings in regard to theory testing or practical implications.
5. A retained null hypothesis could result from the null hypothesis's actually being true in nature. It could also result from contamination by internal validity problems that obscure treatment effects, from lack of statistical power of the design used in the study, or from an inability of the instruments to measure accurately the effects of treatment on the dependent variable.
6. **a.** Unacceptable; no statement of relationship is given.
   **b.** Unacceptable; too wordy.
   **c.** Unacceptable; emotion-laden titles are not appropriate for research articles.
   **d.** Acceptable; meets criteria for title.
7. Dissertation form is the most formal and detailed in presentation; it follows the specifics of a particular style manual. The journal article is a more concise presentation, with a brief statement of the problem, related literature, and methodology; a greater proportion of the article is devoted to major results and a discussion of their significance. A paper that is to be read at a conference is the most informal; geared to its audience, it states the hypothesis, briefly describes the procedure, and emphasizes the most important findings.
8. Yes. Results contrary to one's expectations are as legitimate as any other results and should be interpreted as such.
9. **a.** Paper at a professional meeting
   **b.** Journal article
   **c.** Paper at a professional meeting
10. Answers will vary. The following is simply a suggestion"

| | Test | | |
|---|---|---|---|
| **Incentive** | **1** | **2** | **Combined** |
| Tangible ($n = 36$) | | | |
| M | 72.0 | 76.0 | 74.0 |
| SD | 10.9 | 9.7 | 10.3 |
| Intangible ($n = 34$) | | | |
| M | 68.0 | 70.0 | 69.0 |
| SD | 8.5 | 10.2 | 9.3 |

11. Answers will vary.
12. e
13. a
14. d
15. **a.** Correct decision (rejected a false null hypothesis)
    **b.** Type I error(rejected a true null hypothesis)
    **c.** Correct decision
    **d.** Type II error (retained a false null hypothesis)

## REFERENCES

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Johnson, A. P. (2003). *A short guide to academic writing*. Lanham, MD: University Press of America.

Mitchell, M., & Jolley, J. (2006). *Research design explained*. Belmont, CA: Thomson.

Nicol, A. (1999). *Presenting your findings: A practical guide for creating tables*. Washington, DC: American Psychological Association.

Nicol, A. (2003). *A practical guide for presenting figures, posters, and presentations*. Washington, DC: American Psychological Association.

Rudestam, K. E., & Kjell, E. (2007). *Surviving your dissertation: A comprehensive guide to content and process*. Thousand Oaks, CA: Sage.

Springer, S., Pugalee, D., & Algozzine, B. (2007). Improving mathematics skills of high school students. *The Clearing House, 81*(1), 37–44.

# Appendix

**Table A.1** Areas of the Normal Curve

| (1) $z$ | (2) Area between the Mean and $z$ | (3) Area beyond $z$ | (1) $z$ | (2) Area between the Mean and $z$ | (3) Area beyond $z$ |
|---|---|---|---|---|---|
| 0.00 | .0000 | .5000 | 0.35 | .1368 | .3632 |
| 0.01 | .0040 | .4960 | 0.36 | .1406 | .3594 |
| 0.02 | .0080 | .4920 | 0.37 | .1443 | .3557 |
| 0.03 | .0120 | .4880 | 0.38 | .1480 | .3520 |
| 0.04 | .0160 | .4840 | 0.39 | .1517 | .3483 |
| 0.05 | .0199 | .4801 | 0.40 | .1554 | .3446 |
| 0.06 | .0239 | .4761 | 0.41 | .1591 | .3409 |
| 0.07 | .0279 | .4721 | 0.42 | .1628 | .3372 |
| 0.08 | .0319 | .4681 | 0.43 | .1664 | .3336 |
| 0.09 | .0359 | .4641 | 0.44 | .1700 | .3300 |
| 0.10 | .0398 | .4602 | 0.45 | .1736 | .3264 |
| 0.11 | .0438 | .4562 | 0.46 | .1772 | .3228 |
| 0.12 | .0478 | .4522 | 0.47 | .1808 | .3192 |
| 0.13 | .0517 | .4483 | 0.48 | .1844 | .3156 |
| 0.14 | .0557 | .4443 | 0.49 | .1879 | .3121 |
| 0.15 | .0596 | .4404 | 0.50 | .1915 | .3085 |
| 0.16 | .0636 | .4364 | 0.51 | .1950 | .3050 |
| 0.17 | .0675 | .4325 | 0.52 | .1985 | .3015 |
| 0.18 | .0714 | .4286 | 0.53 | .2019 | .2981 |
| 0.19 | .0753 | .4247 | 0.54 | .2054 | .2946 |
| 0.20 | .0793 | .4207 | 0.55 | .2088 | .2912 |
| 0.21 | .0832 | .4168 | 0.56 | .2123 | .2877 |
| 0.22 | .0871 | .4129 | 0.57 | .2157 | .2843 |
| 0.23 | .0910 | .4090 | 0.58 | .2190 | .2810 |
| 0.24 | .0948 | .4052 | 0.59 | .2224 | .2776 |
| 0.25 | .0987 | .4013 | 0.60 | .2257 | .2743 |
| 0.26 | .1026 | .3974 | 0.61 | .2291 | .2709 |
| 0.27 | .1064 | .3936 | 0.62 | .2324 | .2676 |
| 0.28 | .1103 | .3897 | 0.63 | .2357 | .2643 |
| 0.29 | .1141 | .3859 | 0.64 | .2389 | .2611 |
| 0.30 | .1179 | .3821 | 0.65 | .2422 | .2578 |
| 0.31 | .1217 | .3783 | 0.66 | .2454 | .2546 |
| 0.32 | .1255 | .3745 | 0.67 | .2486 | .2514 |
| 0.33 | .1293 | .3707 | 0.68 | .2517 | .2483 |
| 0.34 | .1331 | .3669 | 0.69 | .2549 | .2451 |

**Table A.1** Areas of the Normal Curve (*continued*)

| (1) $z$ | (2) Area between the Mean and $z$ | (3) Area beyond $z$ | (1) $z$ | (2) Area between the Mean and $z$ | (3) Area beyond $z$ |
|---|---|---|---|---|---|
| 0.70 | .2580 | .2420 | 1.05 | .3531 | .1469 |
| 0.71 | .2611 | .2389 | 1.06 | .3554 | .1446 |
| 0.72 | .2642 | .2358 | 1.07 | .3577 | .1423 |
| 0.73 | .2673 | .2327 | 1.08 | .3599 | .1401 |
| 0.74 | .2704 | .2296 | 1.09 | .3621 | .1379 |
| 0.75 | .2734 | .2266 | 1.10 | .3643 | .1357 |
| 0.76 | .2764 | .2236 | 1.11 | .3665 | .1335 |
| 0.77 | .2794 | .2206 | 1.12 | .3686 | .1314 |
| 0.78 | .2823 | .2177 | 1.13 | .3708 | .1292 |
| 0.79 | .2852 | .2148 | 1.14 | .3729 | .1271 |
| 0.80 | .2881 | .2119 | 1.15 | .3749 | .1251 |
| 0.81 | .2910 | .2090 | 1.16 | .3770 | .1230 |
| 0.82 | .2939 | .2061 | 1.17 | .3790 | .1210 |
| 0.83 | .2967 | .2033 | 1.18 | .3810 | .1190 |
| 0.84 | .2995 | .2005 | 1.19 | .3830 | .1170 |
| 0.85 | .3023 | .1977 | 1.20 | .3849 | .1151 |
| 0.86 | .3051 | .1949 | 1.21 | .3869 | .1131 |
| 0.87 | .3078 | .1922 | 1.22 | .3888 | .1112 |
| 0.88 | .3106 | .1894 | 1.23 | .3907 | .1093 |
| 0.89 | .3133 | .1867 | 1.24 | .3925 | .1075 |
| 0.90 | .3159 | .1841 | 1.25 | .3944 | .1056 |
| 0.91 | .3186 | .1814 | 1.26 | .3962 | .1038 |
| 0.92 | .3212 | .1788 | 1.27 | .3980 | .1020 |
| 0.93 | .3238 | .1762 | 1.28 | .3997 | .1003 |
| 0.94 | .3264 | .1736 | 1.29 | .4015 | .0985 |
| 0.95 | .3289 | .1711 | 1.30 | .4032 | .0968 |
| 0.96 | .3315 | .1685 | 1.31 | .4049 | .0951 |
| 0.97 | .3340 | .1660 | 1.32 | .4066 | .0934 |
| 0.98 | .3365 | .1635 | 1.33 | .4082 | .0918 |
| 0.99 | .3389 | .1611 | 1.34 | .4099 | .0901 |
| 1.00 | .3413 | .1587 | 1.35 | .4115 | .0885 |
| 1.01 | .3438 | .1562 | 1.36 | .4131 | .0869 |
| 1.02 | .3461 | .1539 | 1.37 | .4147 | .0853 |
| 1.03 | .3485 | .1515 | 1.38 | .4162 | .0838 |
| 1.04 | .3508 | .1492 | 1.39 | .4177 | .0823 |

(*continued*)

**Table A.1** Areas of the Normal Curve (*continued*)

| (1) $z$ | (2) Area between the Mean and $z$ | (3) Area beyond $z$ | (1) $z$ | (2) Area between the Mean and $z$ | (3) Area beyond $z$ |
|---|---|---|---|---|---|
| 1.40 | .4192 | .0808 | 1.75 | .4599 | .0401 |
| 1.41 | .4207 | .0793 | 1.76 | .4608 | .0392 |
| 1.42 | .4222 | .0778 | 1.77 | .4616 | .0384 |
| 1.43 | .4236 | .0764 | 1.78 | .4625 | .0375 |
| 1.44 | .4251 | .0749 | 1.79 | .4633 | .0367 |
| 1.45 | .4265 | .0735 | 1.80 | .4641 | .0359 |
| 1.46 | .4279 | .0721 | 1.81 | .4649 | .0351 |
| 1.47 | .4292 | .0708 | 1.82 | .4656 | .0344 |
| 1.48 | .4306 | .0694 | 1.83 | .4664 | .0336 |
| 1.49 | .4319 | .0681 | 1.84 | .4671 | .0329 |
| 1.50 | .4332 | .0668 | 1.85 | .4678 | .0322 |
| 1.51 | .4345 | .0655 | 1.86 | .4686 | .0314 |
| 1.52 | .4357 | .0643 | 1.87 | .4693 | .0307 |
| 1.53 | .4370 | .0630 | 1.88 | .4699 | .0301 |
| 1.54 | .4382 | .0618 | 1.89 | .4706 | .0294 |
| 1.55 | .4394 | .0606 | 1.90 | .4713 | .0287 |
| 1.56 | .4406 | .0594 | 1.91 | .4719 | .0281 |
| 1.57 | .4418 | .0582 | 1.92 | .4726 | .0274 |
| 1.58 | .4429 | .0571 | 1.93 | .4732 | .0268 |
| 1.59 | .4441 | .0559 | 1.94 | .4738 | .0262 |
| 1.60 | .4452 | .0548 | 1.95 | .4744 | .0256 |
| 1.61 | .4463 | .0537 | 1.96 | .4750 | .0250 |
| 1.62 | .4474 | .0526 | 1.97 | .4756 | .0244 |
| 1.63 | .4484 | .0516 | 1.98 | .4761 | .0239 |
| 1.64 | .4495 | .0505 | 1.99 | .4767 | .0233 |
| 1.65 | .4505 | .0495 | 2.00 | .4772 | .0228 |
| 1.66 | .4515 | .0485 | 2.01 | .4778 | .0222 |
| 1.67 | .4525 | .0475 | 2.02 | .4783 | .0217 |
| 1.68 | .4535 | .0465 | 2.03 | .4788 | .0212 |
| 1.69 | .4545 | .0455 | 2.04 | .4793 | .0207 |
| 1.70 | .4554 | .0446 | 2.05 | .4798 | .0202 |
| 1.71 | .4564 | .0436 | 2.06 | .4803 | .0197 |
| 1.72 | .4573 | .0427 | 2.07 | .4808 | .0192 |
| 1.73 | .4582 | .0418 | 2.08 | .4812 | .0188 |
| 1.74 | .4591 | .0409 | 2.09 | .4817 | .0183 |

**Table A.1** Areas of the Normal Curve (*continued*)

| (1) $z$ | (2) Area between the Mean and $z$ | (3) Area beyond $z$ | (1) $z$ | (2) Area between the Mean and $z$ | (3) Area beyond $z$ |
|---|---|---|---|---|---|
| 2.10 | .4821 | .0179 | 2.45 | .4929 | .0071 |
| 2.11 | .4826 | .0174 | 2.46 | .4931 | .0069 |
| 2.12 | .4830 | .0170 | 2.47 | .4932 | .0068 |
| 2.13 | .4834 | .0166 | 2.48 | .4934 | .0066 |
| 2.14 | .4838 | .0162 | 2.49 | .4936 | .0064 |
| 2.15 | .4842 | .0158 | 2.50 | .4938 | .0062 |
| 2.16 | .4846 | .0154 | 2.51 | .4940 | .0060 |
| 2.17 | .4850 | .0150 | 2.52 | .4941 | .0059 |
| 2.18 | .4854 | .0146 | 2.53 | .4943 | .0057 |
| 2.19 | .4857 | .0143 | 2.54 | .4945 | .0055 |
| 2.20 | .4861 | .0139 | 2.55 | .4946 | .0054 |
| 2.21 | .4864 | .0136 | 2.56 | .4948 | .0052 |
| 2.22 | .4868 | .0132 | 2.57 | .4949 | .0051 |
| 2.23 | .4871 | .0129 | 2.58 | .4951 | .0049 |
| 2.24 | .4875 | .0125 | 2.59 | .4952 | .0048 |
| 2.25 | .4878 | .0122 | 2.60 | .4953 | .0047 |
| 2.26 | .4881 | .0119 | 2.61 | .4955 | .0045 |
| 2.27 | .4884 | .0116 | 2.62 | .4956 | .0044 |
| 2.28 | .4887 | .0113 | 2.63 | .4957 | .0043 |
| 2.29 | .4890 | .0110 | 2.64 | .4959 | .0041 |
| 2.30 | .4893 | .0107 | 2.65 | .4960 | .0040 |
| 2.31 | .4896 | .0104 | 2.66 | .4961 | .0039 |
| 2.32 | .4898 | .0102 | 2.67 | .4962 | .0038 |
| 2.33 | .4901 | .0099 | 2.68 | .4963 | .0037 |
| 2.34 | .4904 | .0096 | 2.69 | .4964 | .0036 |
| 2.35 | .4906 | .0094 | 2.70 | .4965 | .0035 |
| 2.36 | .4909 | .0091 | 2.71 | .4966 | .0034 |
| 2.37 | .4911 | .0089 | 2.72 | .4967 | .0033 |
| 2.38 | .4913 | .0087 | 2.73 | .4968 | .0032 |
| 2.39 | .4916 | .0084 | 2.74 | .4969 | .0031 |
| 2.40 | .4918 | .0082 | 2.75 | .4970 | .0030 |
| 2.41 | .4920 | .0080 | 2.76 | .4971 | .0029 |
| 2.42 | .4922 | .0078 | 2.77 | .4972 | .0028 |
| 2.43 | .4925 | .0075 | 2.78 | .4973 | .0027 |
| 2.44 | .4927 | .0073 | 2.79 | .4974 | .0026 |

(*continued*)

**Table A.1** Areas of the Normal Curve (*continued*)

| (1) $z$ | (2) Area between the Mean and $z$ | (3) Area beyond $z$ | (1) $z$ | (2) Area between the Mean and $z$ | (3) Area beyond $z$ |
|---|---|---|---|---|---|
| 2.80 | .4974 | .0026 | 3.10 | .4990 | .0010 |
| 2.81 | .4975 | .0025 | 3.11 | .4991 | .0009 |
| 2.82 | .4976 | .0023 | 3.12 | .4991 | .0009 |
| 2.83 | .4977 | .0024 | 3.13 | .4991 | .0009 |
| 2.84 | .4977 | .0023 | 3.14 | .4992 | .0008 |
| 2.85 | .4978 | .0022 | 3.15 | .4992 | .0008 |
| 2.86 | .4979 | .0021 | 3.16 | .4992 | .0008 |
| 2.87 | .4979 | .0021 | 3.17 | .4992 | .0008 |
| 2.88 | .4980 | .0020 | 3.18 | .4993 | .0007 |
| 2.89 | .4981 | .0019 | 3.19 | .4993 | .0007 |
| 2.90 | .4981 | .0019 | 3.20 | .4993 | .0007 |
| 2.91 | .4982 | .0018 | 3.21 | .4993 | .0007 |
| 2.92 | .4982 | .0018 | 3.22 | .4994 | .0006 |
| 2.93 | .4983 | .0017 | 3.23 | .4994 | .0006 |
| 2.94 | .4984 | .0016 | 3.24 | .4994 | .0006 |
| 2.95 | .4984 | .0016 | 3.30 | .4995 | .0005 |
| 2.96 | .4985 | .0015 | 3.40 | .4997 | .0003 |
| 2.97 | .4985 | .0015 | 3.50 | .4998 | .0002 |
| 2.98 | .4986 | .0014 | 3.60 | .4998 | .0002 |
| 2.99 | .4986 | .0014 | 3.70 | .4999 | .0001 |
| 3.00 | .4987 | .0013 | 3.90 | .49995 | .00005 |
| 3.01 | .4987 | .0013 | 4.00 | .49997 | .00003 |
| 3.20 | .4987 | .0013 | 4.50 | .4999966 | .0000034 |
| 3.03 | .4988 | .0012 | 5.00 | .4999997 | .0000003 |
| 3.04 | .4988 | .0012 | 5.50 | .499999981 | .000000019 |
| 3.05 | .4989 | .0011 | 6.00 | .499999999 | .000000001 |
| 3.06 | .4989 | .0011 | | | |
| 3.07 | .4989 | .0011 | | | |
| 3.08 | .4990 | .0010 | | | |
| 3.09 | .4990 | .0010 | | | |

**Table A.2** Table of *t* Values

| | Level of Significance for a Directional (One-Tailed) Test | | | | | |
|---|---|---|---|---|---|---|
| | .10 | .05 | .025 | .01 | .005 | .0005 |
| | Level of Significance for a Nondirectional (Two-Tailed) Test | | | | | |
| *df* | .20 | .10 | .05 | .02 | .01 | .001 |
| 1 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 | 636.619 |
| 2 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 31.598 |
| 3 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 12.941 |
| 4 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 8.610 |
| 5 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 6.859 |
| 6 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 5.959 |
| 7 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 5.405 |
| 8 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 5.041 |
| 9 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 4.781 |
| 10 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 4.587 |
| 11 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 4.437 |
| 12 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 4.318 |
| 13 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 | 4.221 |
| 14 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 4.140 |
| 15 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 4.073 |
| 16 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 4.015 |
| 17 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 | 3.965 |
| 18 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 | 3.922 |
| 19 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.883 |
| 20 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.850 |
| 21 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 | 3.819 |
| 22 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 | 3.792 |
| 23 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 | 3.767 |
| 24 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 | 3.745 |
| 25 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 | 3.725 |
| 26 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 | 3.707 |
| 27 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 | 3.690 |
| 28 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 | 3.674 |
| 29 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 | 3.659 |
| 30 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 | 3.646 |
| 40 | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 | 3.551 |
| 60 | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 | 3.460 |
| 120 | 1.289 | 1.658 | 1.980 | 2.358 | 2.617 | 3.373 |
| ∞ | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 | 3.291 |

*Source:* From *Statistical Tables for Biological, Agricultural, and Medical Research,* edited by R. A. Fisher and F. Yates. Pearson Education Limited. Copyright © 1938 by Oliver and Boyd. Reproduced with permission of the publisher.

**Table A.3** Critical Values of the Pearson Product Moment Correlation Coefficient

| | Level of Significance for a Directional (One-Tailed) Test | | | | |
|---|---|---|---|---|---|
| | .05 | .025 | .01 | .005 | .0005 |
| | Level of Significance for a Nondirectional (Two-Tailed) Test | | | | |
| $df = N - 2$ | .10 | .05 | .02 | .01 | .001 |
| 1 | .9877 | .9969 | .9995 | .9999 | 1.0000 |
| 2 | .9000 | .9500 | .9800 | .9900 | .9990 |
| 3 | .8054 | .8783 | .9343 | .9587 | .9912 |
| 4 | .7293 | .8114 | .8822 | .9172 | .9741 |
| 5 | .6694 | .7545 | .8329 | .8745 | .9507 |
| 6 | .6215 | .7067 | .7887 | .8343 | .9249 |
| 7 | .5822 | .6664 | .7498 | .7977 | .8982 |
| 8 | .5494 | .6319 | .7155 | .7646 | .8721 |
| 9 | .5214 | .6021 | .6851 | .7348 | .8471 |
| 10 | .4973 | .5760 | .6581 | .7079 | .8233 |
| 11 | .4762 | .5529 | .6339 | .6835 | .8010 |
| 12 | .4575 | .5324 | .6120 | .6614 | .7800 |
| 13 | .4409 | .5139 | .5923 | .6411 | .7603 |
| 14 | .4259 | .4973 | .5742 | .6226 | .7420 |
| 15 | .4124 | .4821 | .5577 | .6055 | .7246 |
| 16 | .4000 | .4683 | .5425 | .5897 | .7084 |
| 17 | .3887 | .4555 | .5285 | .5751 | .6932 |
| 18 | .3783 | .4438 | .5155 | .5614 | .6787 |
| 19 | .3687 | .4329 | .5034 | .5487 | .6652 |
| 20 | .3598 | .4227 | .4921 | .5368 | .6524 |
| 25 | .3233 | .3809 | .4451 | .4869 | .5974 |
| 30 | .2960 | .3494 | .4093 | .4487 | .5541 |
| 35 | .2746 | .3246 | .3810 | .4182 | .5189 |
| 40 | .2573 | .3044 | .3578 | .3932 | .4896 |
| 45 | .2428 | .2875 | .3384 | .3721 | .4648 |
| 50 | .2306 | .2732 | .3218 | .3541 | .4433 |
| 60 | .2108 | .2500 | .2948 | .3248 | .4078 |
| 70 | .1954 | .2319 | .2737 | .3017 | .3799 |
| 80 | .1829 | .2172 | .2565 | .2830 | .3568 |
| 90 | .1726 | .2050 | .2422 | .2673 | .3375 |
| 100 | .1638 | .1946 | .2301 | .2540 | .3211 |

*Source:* From *Statistical Tables for Biological, Agricultural, and Medical Research,* edited by R. A. Fisher and F. Yates. Pearson Education Limited. Copyright © 1938 by Oliver and Boyd. Reproduced with permission of the publisher.

**Table A.4** The 5 (Roman Type) and 1 (Boldface Type) Percent Points for the *F* Distribution (*continued*)

| Denominator | $N_1$ Degrees of Freedom for Numerator | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $df=N_1$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 14 | 16 | 20 | 24 | 30 | 40 | 50 | 75 | 100 | 200 | 500 | ∞ |
| 1 | 161 | 200 | 216 | 225 | 230 | 234 | 237 | 239 | 241 | 242 | 243 | 244 | 245 | 246 | 248 | 249 | 250 | 251 | 252 | 253 | 253 | 254 | 254 | 254 |
| | 4,052 | 4,999 | 5,403 | 5,625 | 5,764 | 5,859 | 5,928 | 5,981 | 6,022 | 6,056 | 6,082 | 6,106 | 6,142 | 6,169 | 6,208 | 6,234 | 6,258 | 6,286 | 6,302 | 6,323 | 6,334 | 6,352 | 6,361 | 6,366 |
| 2 | 18.51 | 19.00 | 19.16 | 19.25 | 19.30 | 19.33 | 19.36 | 19.37 | 19.38 | 19.39 | 19.40 | 19.41 | 19.42 | 19.43 | 19.44 | 19.45 | 19.46 | 19.47 | 19.47 | 19.48 | 19.49 | 19.49 | 19.50 | 19.50 |
| | 98.49 | 99.00 | 99.17 | 99.25 | 99.30 | 99.33 | 99.34 | 99.36 | 99.38 | 99.40 | 99.41 | 99.42 | 99.43 | 99.44 | 99.45 | 99.46 | 99.47 | 99.48 | 99.48 | 99.49 | 99.49 | 99.49 | 99.50 | 99.50 |
| 3 | 10.13 | 9.55 | 9.28 | 9.12 | 9.01 | 8.94 | 8.88 | 8.84 | 8.81 | 8.78 | 8.76 | 8.74 | 8.71 | 8.69 | 8.66 | 8.64 | 8.62 | 8.60 | 8.58 | 8.57 | 8.56 | 8.54 | 8.54 | 8.53 |
| | 34.12 | 30.82 | 29.46 | 28.71 | 28.24 | 27.91 | 27.67 | 27.49 | 27.34 | 27.23 | 27.13 | 27.05 | 26.92 | 26.83 | 26.69 | 26.60 | 26.50 | 26.41 | 26.35 | 26.27 | 26.23 | 26.18 | 26.14 | 26.12 |
| 4 | 7.71 | 6.94 | 6.59 | 6.39 | 6.26 | 6.16 | 6.09 | 6.04 | 6.00 | 5.96 | 5.93 | 5.91 | 5.87 | 5.84 | 5.80 | 5.77 | 5.74 | 5.71 | 5.70 | 5.68 | 5.66 | 5.65 | 5.64 | 5.63 |
| | 21.20 | 18.00 | 16.69 | 15.98 | 15.52 | 15.21 | 14.98 | 14.80 | 14.66 | 14.54 | 14.45 | 14.37 | 14.24 | 14.15 | 14.02 | 13.93 | 13.83 | 13.74 | 13.69 | 13.61 | 13.57 | 13.52 | 13.48 | 13.46 |
| 5 | 6.61 | 5.79 | 5.41 | 5.19 | 5.05 | 4.95 | 4.88 | 4.82 | 4.78 | 4.74 | 4.70 | 4.68 | 4.64 | 4.60 | 4.56 | 4.53 | 4.50 | 4.46 | 4.44 | 4.42 | 4.40 | 4.38 | 4.37 | 4.36 |
| | 16.26 | 13.27 | 12.06 | 11.39 | 10.97 | 10.67 | 10.45 | 10.27 | 10.15 | 10.05 | 9.96 | 9.89 | 9.77 | 9.68 | 9.55 | 9.47 | 9.38 | 9.29 | 9.24 | 9.17 | 9.13 | 9.07 | 9.04 | 9.02 |
| 6 | 5.99 | 5.14 | 4.76 | 4.53 | 4.39 | 4.28 | 4.21 | 4.15 | 4.10 | 4.06 | 4.03 | 4.00 | 3.96 | 3.92 | 3.87 | 3.84 | 3.81 | 3.77 | 3.75 | 3.72 | 3.71 | 3.69 | 3.68 | 3.67 |
| | 13.74 | 10.92 | 9.78 | 9.15 | 8.75 | 8.47 | 8.26 | 8.10 | 7.98 | 7.87 | 7.79 | 7.72 | 7.60 | 7.52 | 7.39 | 7.31 | 7.23 | 7.14 | 7.09 | 7.02 | 6.99 | 6.94 | 6.90 | 6.88 |
| 7 | 5.59 | 4.74 | 4.35 | 4.12 | 3.97 | 3.87 | 3.79 | 3.73 | 3.68 | 3.63 | 3.60 | 3.57 | 3.52 | 3.49 | 3.44 | 3.41 | 3.38 | 3.34 | 3.32 | 3.29 | 3.28 | 3.25 | 3.24 | 3.23 |
| | 12.25 | 9.55 | 8.45 | 7.85 | 7.46 | 7.19 | 7.00 | 6.84 | 6.71 | 6.62 | 6.54 | 6.47 | 6.35 | 6.27 | 6.15 | 6.07 | 5.98 | 5.90 | 5.85 | 5.78 | 5.75 | 5.70 | 5.67 | 5.65 |
| 8 | 5.32 | 4.46 | 4.07 | 3.84 | 3.69 | 3.58 | 3.50 | 3.44 | 3.39 | 3.34 | 3.31 | 3.28 | 3.23 | 3.20 | 3.15 | 3.12 | 3.08 | 3.05 | 3.03 | 3.00 | 2.98 | 2.96 | 2.94 | 2.93 |
| | 11.26 | 8.65 | 7.59 | 7.01 | 6.63 | 6.37 | 6.19 | 6.03 | 5.91 | 5.82 | 5.74 | 5.67 | 5.56 | 5.48 | 5.36 | 5.28 | 5.20 | 5.11 | 5.06 | 5.00 | 4.96 | 4.91 | 4.88 | 4.86 |
| 9 | 5.12 | 4.26 | 3.86 | 3.63 | 3.48 | 3.37 | 3.29 | 3.23 | 3.18 | 3.13 | 3.10 | 3.07 | 3.02 | 2.98 | 2.93 | 2.90 | 2.86 | 2.82 | 2.80 | 2.77 | 2.76 | 2.73 | 2.72 | 2.71 |
| | 10.56 | 8.02 | 6.99 | 6.42 | 6.06 | 5.80 | 5.62 | 5.47 | 5.35 | 5.26 | 5.18 | 5.11 | 5.00 | 4.92 | 4.80 | 4.73 | 4.64 | 4.56 | 4.51 | 4.45 | 4.41 | 4.36 | 4.33 | 4.31 |
| 10 | 4.96 | 4.10 | 3.71 | 3.48 | 3.33 | 3.22 | 3.14 | 3.07 | 3.02 | 2.97 | 2.94 | 2.91 | 2.86 | 2.82 | 2.77 | 2.74 | 2.70 | 2.67 | 2.64 | 2.61 | 2.59 | 2.56 | 2.55 | 2.54 |
| | 10.04 | 7.56 | 6.55 | 5.99 | 5.64 | 5.39 | 5.21 | 5.06 | 4.95 | 4.85 | 4.78 | 4.71 | 4.60 | 4.52 | 4.41 | 4.33 | 4.25 | 4.17 | 4.12 | 4.05 | 4.01 | 3.96 | 3.93 | 3.91 |
| 11 | 4.84 | 3.98 | 3.59 | 3.36 | 3.20 | 3.09 | 3.01 | 2.95 | 2.90 | 2.86 | 2.82 | 2.79 | 2.74 | 2.70 | 2.65 | 2.61 | 2.57 | 2.53 | 2.50 | 2.47 | 2.45 | 2.42 | 2.41 | 2.40 |
| | 9.65 | 7.20 | 6.22 | 5.67 | 5.32 | 5.07 | 4.88 | 4.74 | 4.63 | 4.54 | 4.46 | 4.40 | 4.29 | 4.21 | 4.10 | 4.02 | 3.94 | 3.86 | 3.80 | 3.74 | 3.70 | 3.66 | 3.62 | 3.60 |
| 12 | 4.75 | 3.88 | 3.49 | 3.26 | 3.11 | 3.00 | 2.92 | 2.85 | 2.80 | 2.76 | 2.72 | 2.69 | 2.64 | 2.60 | 2.54 | 2.50 | 2.46 | 2.42 | 2.40 | 2.36 | 2.35 | 2.32 | 2.31 | 2.30 |
| | 9.33 | 6.93 | 5.95 | 5.41 | 5.06 | 4.82 | 4.65 | 4.50 | 4.39 | 4.30 | 4.22 | 4.16 | 4.05 | 3.98 | 3.86 | 3.78 | 3.70 | 3.61 | 3.56 | 3.49 | 3.46 | 3.41 | 3.38 | 3.36 |
| 13 | 4.67 | 3.80 | 3.41 | 3.18 | 3.02 | 2.92 | 2.84 | 2.77 | 2.72 | 2.67 | 2.63 | 2.60 | 2.55 | 2.51 | 2.46 | 2.42 | 2.38 | 2.34 | 2.32 | 2.28 | 2.26 | 2.24 | 2.22 | 2.21 |
| | 9.07 | 6.70 | 5.74 | 5.20 | 4.86 | 4.62 | 4.44 | 4.30 | 4.19 | 4.10 | 4.02 | 3.96 | 3.85 | 3.78 | 3.67 | 3.59 | 3.51 | 3.42 | 3.37 | 3.30 | 3.27 | 3.21 | 3.18 | 3.16 |
| 14 | 4.60 | 3.74 | 3.34 | 3.11 | 2.96 | 2.85 | 2.77 | 2.70 | 2.65 | 2.60 | 2.56 | 2.53 | 2.48 | 2.44 | 2.39 | 2.35 | 2.31 | 2.27 | 2.24 | 2.21 | 2.19 | 2.16 | 2.14 | 2.13 |
| | 8.86 | 6.51 | 5.56 | 5.03 | 4.69 | 4.46 | 4.28 | 4.14 | 4.03 | 3.94 | 3.86 | 3.80 | 3.70 | 3.62 | 3.51 | 3.43 | 3.34 | 3.26 | 3.21 | 3.14 | 3.11 | 3.06 | 3.02 | 3.00 |
| 15 | 4.54 | 3.68 | 3.29 | 3.06 | 2.90 | 2.79 | 2.70 | 2.64 | 2.59 | 2.55 | 2.51 | 2.48 | 2.43 | 2.39 | 2.33 | 2.29 | 2.25 | 2.21 | 2.18 | 2.15 | 2.12 | 2.10 | 2.08 | 2.07 |
| | 8.68 | 6.36 | 5.42 | 4.89 | 4.56 | 4.32 | 4.14 | 4.00 | 3.89 | 3.80 | 3.73 | 3.67 | 3.56 | 3.48 | 3.36 | 3.29 | 3.20 | 3.12 | 3.07 | 3.00 | 2.97 | 2.92 | 2.89 | 2.87 |

**Table A.4** The 5 (Roman Type) and 1 (Boldface Type) Percent Points for the *F* Distribution (*continued*)

| Denominator | $N_1$ Degrees of Freedom for Numerator | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $df=N_1$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 14 | 16 | 20 | 24 | 30 | 40 | 50 | 75 | 100 | 200 | 500 | ∞ |
| 16 | 4.49 | 3.63 | 3.24 | 3.01 | 2.85 | 2.74 | 2.66 | 2.59 | 2.54 | 2.49 | 2.45 | 2.42 | 2.37 | 2.33 | 2.28 | 2.24 | 2.20 | 2.16 | 2.13 | 2.09 | 2.07 | 2.04 | 2.02 | 2.01 |
| | **8.53** | **6.23** | **5.29** | **4.77** | **4.44** | **4.20** | **4.03** | **3.89** | **3.78** | **3.69** | **3.61** | **3.55** | **3.45** | **3.37** | **3.25** | **3.18** | **3.10** | **3.01** | **2.96** | **2.89** | **2.86** | **2.80** | **2.77** | **2.75** |
| 17 | 4.45 | 3.59 | 3.20 | 2.96 | 2.81 | 2.70 | 2.62 | 2.55 | 2.50 | 2.45 | 2.41 | 2.38 | 2.33 | 2.29 | 2.23 | 2.19 | 2.15 | 2.11 | 2.08 | 2.04 | 2.02 | 1.99 | 1.97 | 1.96 |
| | **8.40** | **6.11** | **5.18** | **4.67** | **4.34** | **4.10** | **3.93** | **3.79** | **3.68** | **3.59** | **3.52** | **3.45** | **3.35** | **3.27** | **3.16** | **3.08** | **3.00** | **2.92** | **2.86** | **2.79** | **2.76** | **2.70** | **2.67** | **2.65** |
| 18 | 4.41 | 3.55 | 3.16 | 2.93 | 2.77 | 2.66 | 2.58 | 2.51 | 2.46 | 2.41 | 2.37 | 2.34 | 2.29 | 2.25 | 2.19 | 2.15 | 2.11 | 2.07 | 2.04 | 2.00 | 1.98 | 1.95 | 1.93 | 1.92 |
| | **8.28** | **6.01** | **5.09** | **4.58** | **4.25** | **4.01** | **3.85** | **3.71** | **3.60** | **3.51** | **3.44** | **3.37** | **3.27** | **3.19** | **3.07** | **3.00** | **2.91** | **2.83** | **2.78** | **2.71** | **2.68** | **2.62** | **2.59** | **2.57** |
| 19 | 4.38 | 3.52 | 3.13 | 2.90 | 2.74 | 2.63 | 2.55 | 2.48 | 2.43 | 2.38 | 2.34 | 2.31 | 2.26 | 2.21 | 2.15 | 2.11 | 2.07 | 2.02 | 2.00 | 1.96 | 1.94 | 1.91 | 1.90 | 1.88 |
| | **8.18** | **5.93** | **5.01** | **4.50** | **4.17** | **3.94** | **3.77** | **3.63** | **3.52** | **3.43** | **3.36** | **3.30** | **3.19** | **3.12** | **3.00** | **2.92** | **2.84** | **2.76** | **2.70** | **2.63** | **2.60** | **2.54** | **2.51** | **2.49** |
| 20 | 4.35 | 3.49 | 3.10 | 2.87 | 2.71 | 2.60 | 2.52 | 2.45 | 2.40 | 2.35 | 2.31 | 2.28 | 2.23 | 2.18 | 2.12 | 2.08 | 2.04 | 1.99 | 1.96 | 1.92 | 1.90 | 1.87 | 1.85 | 1.84 |
| | **8.10** | **5.85** | **4.94** | **4.43** | **4.10** | **3.87** | **3.71** | **3.56** | **3.45** | **3.37** | **3.30** | **3.23** | **3.13** | **3.05** | **2.94** | **2.86** | **2.77** | **2.69** | **2.63** | **2.56** | **2.53** | **2.47** | **2.44** | **2.42** |
| 21 | 4.32 | 3.47 | 3.07 | 2.84 | 2.68 | 2.57 | 2.49 | 2.42 | 2.37 | 2.32 | 2.28 | 2.25 | 2.20 | 2.15 | 2.09 | 2.05 | 2.00 | 1.96 | 1.93 | 1.89 | 1.87 | 1.84 | 1.82 | 1.81 |
| | **8.02** | **5.78** | **4.87** | **4.37** | **4.04** | **3.81** | **3.65** | **3.51** | **3.40** | **3.31** | **3.24** | **3.17** | **3.07** | **2.99** | **2.88** | **2.80** | **2.72** | **2.63** | **2.58** | **2.51** | **2.47** | **2.42** | **2.38** | **2.36** |
| 22 | 4.30 | 3.44 | 3.05 | 2.82 | 2.66 | 2.55 | 2.47 | 2.40 | 2.35 | 2.30 | 2.26 | 2.23 | 2.18 | 2.13 | 2.07 | 2.03 | 1.98 | 1.93 | 1.91 | 1.87 | 1.84 | 1.81 | 1.80 | 1.78 |
| | **7.94** | **5.72** | **4.82** | **4.31** | **3.99** | **3.76** | **3.59** | **3.45** | **3.35** | **3.26** | **3.18** | **3.12** | **3.02** | **2.94** | **2.83** | **2.75** | **2.67** | **2.58** | **2.53** | **2.46** | **2.42** | **2.37** | **2.33** | **2.31** |
| 23 | 4.28 | 3.42 | 3.03 | 2.80 | 2.64 | 2.53 | 2.45 | 2.38 | 2.32 | 2.28 | 2.24 | 2.20 | 2.14 | 2.10 | 2.04 | 2.00 | 1.96 | 1.91 | 1.88 | 1.84 | 1.82 | 1.79 | 1.77 | 1.76 |
| | **7.88** | **5.66** | **4.76** | **4.26** | **3.94** | **3.71** | **3.54** | **3.41** | **3.30** | **3.21** | **3.14** | **3.07** | **2.97** | **2.89** | **2.78** | **2.70** | **2.62** | **2.53** | **2.48** | **2.41** | **2.37** | **2.32** | **2.28** | **2.26** |
| 24 | 4.26 | 3.40 | 3.01 | 2.78 | 2.62 | 2.51 | 2.43 | 2.36 | 2.30 | 2.26 | 2.22 | 2.18 | 2.13 | 2.09 | 2.02 | 1.98 | 1.94 | 1.89 | 1.86 | 1.82 | 1.80 | 1.76 | 1.74 | 1.73 |
| | **7.82** | **5.61** | **4.72** | **4.22** | **3.90** | **3.67** | **3.50** | **3.36** | **3.25** | **3.17** | **3.09** | **3.03** | **2.93** | **2.85** | **2.74** | **2.66** | **2.58** | **2.49** | **2.44** | **2.36** | **2.33** | **2.27** | **2.23** | **2.21** |
| 25 | 4.24 | 3.38 | 2.99 | 2.76 | 2.60 | 2.49 | 2.41 | 2.34 | 2.28 | 2.24 | 2.20 | 2.16 | 2.11 | 2.06 | 2.00 | 1.96 | 1.92 | 1.87 | 1.84 | 1.80 | 1.77 | 1.74 | 1.72 | 1.71 |
| | **7.77** | **5.57** | **4.68** | **4.18** | **3.86** | **3.63** | **3.46** | **3.32** | **3.21** | **3.13** | **3.05** | **2.99** | **2.89** | **2.81** | **2.70** | **2.62** | **2.54** | **2.45** | **2.40** | **2.32** | **2.29** | **2.23** | **2.19** | **2.17** |
| 26 | 4.22 | 3.37 | 2.98 | 2.74 | 2.59 | 2.47 | 2.39 | 2.32 | 2.27 | 2.22 | 2.18 | 2.15 | 2.10 | 2.05 | 1.99 | 1.95 | 1.90 | 1.85 | 1.82 | 1.78 | 1.76 | 1.72 | 1.70 | 1.69 |
| | **7.72** | **5.53** | **4.64** | **4.14** | **3.82** | **3.59** | **3.42** | **3.29** | **3.17** | **3.09** | **3.02** | **2.96** | **2.86** | **2.77** | **2.66** | **2.58** | **2.50** | **2.41** | **2.36** | **2.28** | **2.25** | **2.19** | **2.15** | **2.13** |
| 27 | 4.21 | 3.35 | 2.96 | 2.73 | 2.57 | 2.46 | 2.37 | 2.30 | 2.25 | 2.20 | 2.16 | 2.13 | 2.08 | 2.03 | 1.97 | 1.93 | 1.88 | 1.84 | 1.80 | 1.76 | 1.74 | 1.71 | 1.68 | 1.67 |
| | **7.68** | **5.49** | **4.60** | **4.11** | **3.79** | **3.56** | **3.39** | **3.26** | **3.14** | **3.06** | **2.98** | **2.93** | **2.83** | **2.74** | **2.63** | **2.55** | **2.47** | **2.38** | **2.33** | **2.25** | **2.21** | **2.16** | **2.12** | **2.10** |
| 28 | 4.20 | 3.34 | 2.95 | 2.71 | 2.56 | 2.44 | 2.36 | 2.29 | 2.24 | 2.19 | 2.15 | 2.12 | 2.06 | 2.02 | 1.96 | 1.91 | 1.87 | 1.81 | 1.78 | 1.75 | 1.72 | 1.69 | 1.67 | 1.65 |
| | **7.64** | **5.45** | **4.57** | **4.07** | **3.76** | **3.53** | **3.36** | **3.23** | **3.11** | **3.03** | **2.95** | **2.90** | **2.80** | **2.71** | **2.60** | **2.52** | **2.44** | **2.35** | **2.30** | **2.22** | **2.18** | **2.13** | **2.09** | **2.06** |
| 29 | 4.18 | 3.33 | 2.93 | 2.70 | 2.54 | 2.43 | 2.35 | 2.28 | 2.22 | 2.18 | 2.14 | 2.10 | 2.05 | 2.00 | 1.94 | 1.90 | 1.85 | 1.80 | 1.77 | 1.73 | 1.71 | 1.68 | 1.65 | 1.64 |
| | **7.60** | **5.42** | **4.54** | **4.04** | **3.73** | **3.50** | **3.33** | **3.20** | **3.08** | **3.00** | **2.92** | **2.87** | **2.77** | **2.68** | **2.57** | **2.49** | **2.41** | **2.32** | **2.27** | **2.19** | **2.15** | **2.10** | **2.06** | **2.03** |
| 30 | 4.17 | 3.32 | 2.92 | 2.69 | 2.53 | 2.42 | 2.34 | 2.27 | 2.21 | 2.16 | 2.12 | 2.09 | 2.04 | 1.99 | 1.93 | 1.89 | 1.84 | 1.79 | 1.76 | 1.72 | 1.69 | 1.66 | 1.64 | 1.62 |
| | **7.56** | **5.39** | **4.51** | **4.02** | **3.70** | **3.47** | **3.30** | **3.17** | **3.06** | **2.98** | **2.90** | **2.84** | **2.74** | **2.66** | **2.55** | **2.47** | **2.38** | **2.29** | **2.24** | **2.16** | **2.13** | **2.07** | **2.03** | **2.01** |

**Table A.4** The 5 (Roman Type) and 1 (Boldface Type) Percent Points for the *F* Distribution (*continued*)

| Denominator | $N_1$ Degrees of Freedom for Numerator | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $df=N_1$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 14 | 16 | 20 | 24 | 30 | 40 | 50 | 75 | 100 | 200 | 500 | ∞ |
| 32 | 4.15 | 3.30 | 2.90 | 2.67 | 2.51 | 2.40 | 2.32 | 2.25 | 2.19 | 2.14 | 2.10 | 2.07 | 2.02 | 1.97 | 1.91 | 1.86 | 1.82 | 1.76 | 1.74 | 1.69 | 1.67 | 1.64 | 1.61 | 1.59 |
| | **7.50** | **5.34** | **4.46** | **3.97** | **3.66** | **3.42** | **3.25** | **3.12** | **3.01** | **2.94** | **2.86** | **2.80** | **2.70** | **2.62** | **2.51** | **2.42** | **2.34** | **2.25** | **2.20** | **2.12** | **2.08** | **2.02** | **1.98** | **1.96** |
| 34 | 4.13 | 3.28 | 2.88 | 2.65 | 2.49 | 2.38 | 2.30 | 2.23 | 2.17 | 2.12 | 2.08 | 2.05 | 2.00 | 1.95 | 1.89 | 1.84 | 1.80 | 1.74 | 1.71 | 1.67 | 1.64 | 1.61 | 1.59 | 1.57 |
| | **7.44** | **5.29** | **4.42** | **3.93** | **3.61** | **3.38** | **3.21** | **3.08** | **2.97** | **2.89** | **2.82** | **2.76** | **2.66** | **2.58** | **2.47** | **2.38** | **2.30** | **2.21** | **2.15** | **2.08** | **2.04** | **1.98** | **1.94** | **1.91** |
| 36 | 4.11 | 3.26 | 2.86 | 2.63 | 2.48 | 2.36 | 2.28 | 2.21 | 2.15 | 2.10 | 2.06 | 2.03 | 1.98 | 1.93 | 1.87 | 1.82 | 1.78 | 1.72 | 1.69 | 1.65 | 1.62 | 1.59 | 1.56 | 1.55 |
| | **7.39** | **5.25** | **4.38** | **3.89** | **3.58** | **3.35** | **3.18** | **3.04** | **2.94** | **2.86** | **2.78** | **2.72** | **2.62** | **2.54** | **2.43** | **2.35** | **2.26** | **2.17** | **2.12** | **2.04** | **2.00** | **1.94** | **1.90** | **1.87** |
| 38 | 4.10 | 3.25 | 2.85 | 2.62 | 2.46 | 2.35 | 2.26 | 2.19 | 2.14 | 2.09 | 2.05 | 2.02 | 1.96 | 1.92 | 1.85 | 1.80 | 1.76 | 1.71 | 1.67 | 1.63 | 1.60 | 1.57 | 1.54 | 1.53 |
| | **7.35** | **5.21** | **4.34** | **3.86** | **3.54** | **3.32** | **3.15** | **3.02** | **2.91** | **2.82** | **2.75** | **2.69** | **2.59** | **2.51** | **2.40** | **2.32** | **2.22** | **2.14** | **2.08** | **2.00** | **1.97** | **1.90** | **1.86** | **1.84** |
| 40 | 4.08 | 3.23 | 2.84 | 2.61 | 2.45 | 2.34 | 2.25 | 2.18 | 2.12 | 2.07 | 2.04 | 2.00 | 1.95 | 1.90 | 1.84 | 1.79 | 1.74 | 1.69 | 1.66 | 1.61 | 1.59 | 1.55 | 1.53 | 1.51 |
| | **7.31** | **5.18** | **4.31** | **3.83** | **3.51** | **3.29** | **3.12** | **2.99** | **2.88** | **2.80** | **2.73** | **2.66** | **2.56** | **2.49** | **2.37** | **2.29** | **2.20** | **2.11** | **2.05** | **1.97** | **1.94** | **1.88** | **1.84** | **1.81** |
| 42 | 4.07 | 3.22 | 2.83 | 2.59 | 2.44 | 2.32 | 2.24 | 2.17 | 2.11 | 2.06 | 2.02 | 1.99 | 1.94 | 1.89 | 1.82 | 1.78 | 1.73 | 1.68 | 1.64 | 1.60 | 1.57 | 1.54 | 1.51 | 1.49 |
| | **7.27** | **5.15** | **4.29** | **3.80** | **3.49** | **3.26** | **3.10** | **2.96** | **2.86** | **2.77** | **2.70** | **2.64** | **2.54** | **2.46** | **2.35** | **2.26** | **2.17** | **2.08** | **2.02** | **1.94** | **1.91** | **1.85** | **1.80** | **1.78** |
| 44 | 4.06 | 3.21 | 2.82 | 2.58 | 2.43 | 2.31 | 2.23 | 2.16 | 2.10 | 2.05 | 2.01 | 1.98 | 1.92 | 1.88 | 1.81 | 1.76 | 1.72 | 1.66 | 1.63 | 1.58 | 1.56 | 1.52 | 1.50 | 1.48 |
| | **7.24** | **5.12** | **4.26** | **3.78** | **3.46** | **3.24** | **3.07** | **2.94** | **2.84** | **2.75** | **2.68** | **2.62** | **2.52** | **2.44** | **2.32** | **2.24** | **2.15** | **2.06** | **2.00** | **1.92** | **1.88** | **1.82** | **1.78** | **1.75** |
| 46 | 4.05 | 3.20 | 2.81 | 2.57 | 2.42 | 2.30 | 2.22 | 2.14 | 2.09 | 2.04 | 2.00 | 1.97 | 1.91 | 1.87 | 1.80 | 1.75 | 1.71 | 1.65 | 1.62 | 1.57 | 1.54 | 1.51 | 1.48 | 1.46 |
| | **7.21** | **5.10** | **4.24** | **3.76** | **3.44** | **3.22** | **3.05** | **2.92** | **2.82** | **2.73** | **2.66** | **2.60** | **2.50** | **2.42** | **2.30** | **2.22** | **2.13** | **2.04** | **1.98** | **1.90** | **1.86** | **1.80** | **1.76** | **1.72** |
| 48 | 4.04 | 3.19 | 2.80 | 2.56 | 2.41 | 2.30 | 2.21 | 2.14 | 2.08 | 2.03 | 1.99 | 1.96 | 1.90 | 1.86 | 1.79 | 1.74 | 1.70 | 1.64 | 1.61 | 1.56 | 1.53 | 1.50 | 1.47 | 1.45 |
| | **7.19** | **5.08** | **4.22** | **3.74** | **3.42** | **3.20** | **3.04** | **2.90** | **2.80** | **2.71** | **2.64** | **2.58** | **2.48** | **2.40** | **2.28** | **2.20** | **2.11** | **2.02** | **1.96** | **1.88** | **1.84** | **1.78** | **1.73** | **1.70** |
| 50 | 4.03 | 3.18 | 2.79 | 2.56 | 2.40 | 2.29 | 2.20 | 2.13 | 2.07 | 2.02 | 1.98 | 1.95 | 1.90 | 1.85 | 1.78 | 1.74 | 1.69 | 1.63 | 1.60 | 1.55 | 1.52 | 1.48 | 1.46 | 1.44 |
| | **7.17** | **5.06** | **4.20** | **3.72** | **3.41** | **3.18** | **3.02** | **2.88** | **2.78** | **2.70** | **2.62** | **2.56** | **2.46** | **2.39** | **2.26** | **2.18** | **2.10** | **2.00** | **1.94** | **1.86** | **1.82** | **1.76** | **1.71** | **1.68** |
| 55 | 4.02 | 3.17 | 2.78 | 2.54 | 2.38 | 2.27 | 2.18 | 2.11 | 2.05 | 2.00 | 1.97 | 1.93 | 1.88 | 1.83 | 1.76 | 1.72 | 1.67 | 1.61 | 1.58 | 1.52 | 1.50 | 1.46 | 1.43 | 1.41 |
| | **7.12** | **5.01** | **4.16** | **3.68** | **3.37** | **3.15** | **2.98** | **2.85** | **2.75** | **2.66** | **2.59** | **2.53** | **2.43** | **2.35** | **2.23** | **2.15** | **2.06** | **1.96** | **1.90** | **1.82** | **1.78** | **1.71** | **1.66** | **1.64** |
| 60 | 4.00 | 3.15 | 2.76 | 2.52 | 2.37 | 2.25 | 2.17 | 2.10 | 2.04 | 1.99 | 1.95 | 1.92 | 1.86 | 1.81 | 1.75 | 1.70 | 1.65 | 1.59 | 1.56 | 1.50 | 1.48 | 1.44 | 1.41 | 1.39 |
| | **7.08** | **4.98** | **4.13** | **3.65** | **3.34** | **3.12** | **2.95** | **2.82** | **2.72** | **2.63** | **2.56** | **2.50** | **2.40** | **2.32** | **2.20** | **2.12** | **2.03** | **1.93** | **1.87** | **1.79** | **1.74** | **1.68** | **1.63** | **1.60** |
| 65 | 3.99 | 3.14 | 2.75 | 2.51 | 2.36 | 2.24 | 2.15 | 2.08 | 2.02 | 1.98 | 1.94 | 1.90 | 1.85 | 1.80 | 1.73 | 1.68 | 1.63 | 1.57 | 1.54 | 1.49 | 1.46 | 1.42 | 1.39 | 1.37 |
| | **7.04** | **4.95** | **4.10** | **3.62** | **3.31** | **3.09** | **2.93** | **2.79** | **2.70** | **2.61** | **2.54** | **2.47** | **2.37** | **2.30** | **2.18** | **2.09** | **2.00** | **1.90** | **1.84** | **1.76** | **1.71** | **1.64** | **1.60** | **1.56** |
| 70 | 3.98 | 3.13 | 2.74 | 2.50 | 2.35 | 2.23 | 2.14 | 2.07 | 2.01 | 1.97 | 1.93 | 1.89 | 1.84 | 1.79 | 1.72 | 1.67 | 1.62 | 1.56 | 1.53 | 1.47 | 1.45 | 1.40 | 1.37 | 1.35 |
| | **7.01** | **4.92** | **4.08** | **3.60** | **3.29** | **3.07** | **2.91** | **2.77** | **2.67** | **2.59** | **2.51** | **2.45** | **2.35** | **2.28** | **2.15** | **2.07** | **1.98** | **1.88** | **1.82** | **1.74** | **1.69** | **1.62** | **1.56** | **1.53** |
| 80 | 3.96 | 3.11 | 2.72 | 2.48 | 2.33 | 2.21 | 2.12 | 2.05 | 1.99 | 1.95 | 1.91 | 1.88 | 1.82 | 1.77 | 1.70 | 1.65 | 1.60 | 1.54 | 1.51 | 1.45 | 1.42 | 1.38 | 1.35 | 1.32 |
| | **6.96** | **4.88** | **4.04** | **3.56** | **3.25** | **3.04** | **2.87** | **2.74** | **2.64** | **2.55** | **2.48** | **2.41** | **2.32** | **2.24** | **2.11** | **2.03** | **1.94** | **1.84** | **1.78** | **1.70** | **1.65** | **1.57** | **1.52** | **1.49** |

**Table A.4** The 5 (Roman Type) and 1 (Boldface Type) Percent Points for the *F* Distribution (*continued*)

| Denominator | $N_1$ Degrees of Freedom for Numerator | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $df=N_1$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 14 | 16 | 20 | 24 | 30 | 40 | 50 | 75 | 100 | 200 | 500 | ∞ |
| 100 | 3.94 | 3.09 | 2.70 | 2.46 | 2.30 | 2.19 | 2.10 | 2.03 | 1.97 | 1.92 | 1.88 | 1.85 | 1.79 | 1.75 | 1.68 | 1.63 | 1.57 | 1.51 | 1.48 | 1.42 | 1.39 | 1.34 | 1.30 | 1.28 |
| | **6.90** | **4.82** | **3.98** | **3.51** | **3.20** | **2.99** | **2.82** | **2.69** | **2.59** | **2.51** | **2.43** | **2.36** | **2.26** | **2.19** | **2.06** | **1.98** | **1.89** | **1.79** | **1.73** | **1.64** | **1.59** | **1.51** | **1.46** | **1.43** |
| 125 | 3.92 | 3.07 | 2.68 | 2.44 | 2.29 | 2.17 | 2.08 | 2.01 | 1.95 | 1.90 | 1.86 | 1.83 | 1.77 | 1.72 | 1.65 | 1.60 | 1.55 | 1.49 | 1.45 | 1.39 | 1.36 | 1.31 | 1.27 | 1.25 |
| | **6.84** | **4.78** | **3.94** | **3.47** | **3.17** | **2.95** | **2.79** | **2.65** | **2.56** | **2.47** | **2.40** | **2.33** | **2.23** | **2.15** | **2.03** | **1.94** | **1.85** | **1.75** | **1.68** | **1.59** | **1.54** | **1.46** | **1.40** | **1.37** |
| 150 | 3.91 | 3.06 | 2.67 | 2.43 | 2.27 | 2.16 | 2.07 | 2.00 | 1.94 | 1.89 | 1.85 | 1.82 | 1.76 | 1.71 | 1.64 | 1.59 | 1.54 | 1.47 | 1.44 | 1.37 | 1.34 | 1.29 | 1.25 | 1.22 |
| | **6.81** | **4.75** | **3.91** | **3.44** | **3.14** | **2.92** | **2.76** | **2.62** | **2.53** | **2.44** | **2.37** | **2.30** | **2.20** | **2.12** | **2.00** | **1.91** | **1.83** | **1.72** | **1.66** | **1.56** | **1.51** | **1.43** | **1.37** | **1.33** |
| 200 | 3.89 | 3.04 | 2.65 | 2.41 | 2.26 | 2.14 | 2.05 | 1.98 | 1.92 | 1.87 | 1.83 | 1.80 | 1.74 | 1.69 | 1.62 | 1.57 | 1.52 | 1.45 | 1.42 | 1.35 | 1.32 | 1.26 | 1.22 | 1.19 |
| | **6.76** | **4.71** | **3.88** | **3.41** | **3.11** | **2.90** | **2.73** | **2.60** | **2.50** | **2.41** | **2.34** | **2.28** | **2.17** | **2.09** | **1.97** | **1.88** | **1.79** | **1.69** | **1.62** | **1.53** | **1.48** | **1.39** | **1.33** | **1.28** |
| 400 | 3.86 | 3.02 | 2.62 | 2.39 | 2.23 | 2.12 | 2.03 | 1.96 | 1.90 | 1.85 | 1.81 | 1.78 | 1.72 | 1.67 | 1.60 | 1.54 | 1.49 | 1.42 | 1.38 | 1.32 | 1.28 | 1.22 | 1.16 | 1.13 |
| | **6.70** | **4.66** | **3.83** | **3.36** | **3.06** | **2.85** | **2.69** | **2.55** | **2.46** | **2.37** | **2.29** | **2.23** | **2.12** | **2.04** | **1.92** | **1.84** | **1.74** | **1.64** | **1.57** | **1.47** | **1.42** | **1.32** | **1.24** | **1.19** |
| 1000 | 3.85 | 3.00 | 2.61 | 2.38 | 2.22 | 2.10 | 2.02 | 1.95 | 1.89 | 1.84 | 1.80 | 1.76 | 1.70 | 1.65 | 1.58 | 1.53 | 1.47 | 1.41 | 1.36 | 1.30 | 1.26 | 1.19 | 1.13 | 1.08 |
| | **6.66** | **4.62** | **3.80** | **3.34** | **3.04** | **2.82** | **2.66** | **2.53** | **2.43** | **2.34** | **2.26** | **2.20** | **2.09** | **2.01** | **1.89** | **1.81** | **1.71** | **1.61** | **1.54** | **1.44** | **1.38** | **1.28** | **1.19** | **1.11** |
| ∞ | 3.84 | 2.99 | 2.60 | 2.37 | 2.21 | 2.09 | 2.01 | 1.94 | 1.88 | 1.83 | 1.79 | 1.75 | 1.69 | 1.64 | 1.57 | 1.52 | 1.46 | 1.40 | 1.35 | 1.28 | 1.24 | 1.17 | 1.11 | 1.00 |
| | **6.64** | **4.60** | **3.78** | **3.32** | **3.02** | **2.80** | **2.64** | **2.51** | **2.41** | **2.32** | **2.24** | **2.18** | **2.07** | **1.99** | **1.87** | **1.79** | **1.69** | **1.59** | **1.52** | **1.41** | **1.36** | **1.25** | **1.15** | **1.00** |

*Source:* Reprinted by permission from *Statistical Methods,* 8th ed., by George W. Snedecor and William G. Cochran. Published by Blackwell Publishing, Ames, Iowa 50010.

**Table A.5** Table of $x^2$

| *df* | .99 | .98 | .95 | .90 | .80 | .70 | .50 | .30 | .20 | .10 | .05 | .02 | .01 | .001 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | .000157 | .000628 | .00393 | .0158 | .0642 | .148 | .455 | 1.074 | 1.642 | 2.706 | 3.841 | 5.412 | 6.635 | 10.827 |
| 2 | .0201 | .0404 | .103 | .211 | .446 | .713 | 1.386 | 2.408 | 3.219 | 4.605 | 5.991 | 7.824 | 9.210 | 13.815 |
| 3 | .115 | .185 | .352 | .584 | 1.005 | 1.424 | 2.366 | 3.665 | 4.642 | 6.251 | 7.815 | 9.837 | 11.345 | 16.266 |
| 4 | .297 | .429 | .711 | 1.064 | 1.649 | 2.195 | 3.357 | 4.878 | 5.989 | 7.779 | 9.488 | 11.668 | 13.277 | 18.467 |
| 5 | .554 | .752 | 1.145 | 1.610 | 2.343 | 3.000 | 4.351 | 6.064 | 7.289 | 9.236 | 11.070 | 13.388 | 15.086 | 20.515 |
| 6 | .872 | 1.134 | 1.635 | 2.204 | 3.070 | 3.828 | 5.348 | 7.231 | 8.558 | 10.645 | 12.592 | 15.033 | 16.812 | 22.457 |
| 7 | 1.239 | 1.564 | 2.167 | 2.833 | 3.822 | 4.671 | 6.346 | 8.383 | 9.803 | 12.017 | 14.067 | 16.622 | 18.475 | 24.322 |
| 8 | 1.646 | 2.032 | 2.733 | 3.490 | 4.594 | 5.527 | 7.344 | 9.524 | 11.030 | 13.362 | 15.507 | 18.168 | 20.090 | 26.125 |
| 9 | 2.088 | 2.532 | 3.325 | 4.168 | 5.380 | 6.393 | 8.343 | 10.656 | 12.242 | 14.684 | 16.919 | 19.679 | 21.666 | 27.877 |
| 10 | 2.558 | 3.059 | 3.940 | 4.865 | 6.179 | 7.267 | 9.342 | 11.781 | 13.442 | 15.987 | 18.307 | 21.161 | 23.209 | 29.588 |
| 11 | 3.053 | 3.609 | 4.575 | 5.578 | 6.989 | 8.148 | 10.341 | 12.899 | 14.631 | 17.275 | 19.675 | 22.618 | 24.725 | 31.264 |
| 12 | 3.571 | 4.178 | 5.226 | 6.304 | 7.807 | 9.034 | 11.340 | 14.011 | 15.812 | 18.549 | 21.026 | 24.054 | 26.217 | 32.909 |
| 13 | 4.107 | 4.765 | 5.892 | 7.042 | 8.634 | 9.926 | 12.340 | 15.119 | 16.985 | 19.812 | 22.362 | 25.472 | 27.688 | 34.528 |
| 14 | 4.660 | 5.368 | 6.571 | 7.790 | 9.467 | 10.821 | 13.339 | 16.222 | 18.151 | 21.064 | 23.685 | 26.873 | 29.141 | 36.123 |
| 15 | 5.229 | 5.985 | 7.261 | 8.547 | 10.307 | 11.721 | 14.339 | 17.322 | 19.311 | 22.307 | 24.996 | 28.259 | 30.578 | 37.697 |
| 16 | 5.812 | 6.614 | 7.962 | 9.312 | 11.152 | 12.624 | 15.338 | 18.418 | 20.465 | 23.542 | 26.296 | 29.633 | 32.000 | 39.252 |
| 17 | 6.408 | 7.255 | 8.672 | 10.085 | 12.002 | 13.531 | 16.338 | 19.511 | 21.615 | 24.769 | 27.587 | 30.995 | 33.409 | 40.790 |
| 18 | 7.015 | 7.906 | 9.390 | 10.865 | 12.857 | 14.440 | 17.338 | 20.601 | 22.760 | 25.989 | 28.869 | 32.346 | 34.805 | 42.312 |
| 19 | 7.633 | 8.567 | 10.117 | 11.651 | 13.716 | 15.352 | 18.338 | 21.689 | 23.900 | 27.204 | 30.144 | 33.687 | 36.191 | 43.820 |
| 20 | 8.260 | 9.237 | 10.851 | 12.443 | 14.578 | 16.266 | 19.337 | 22.775 | 25.038 | 28.412 | 31.410 | 35.020 | 37.566 | 45.315 |
| 21 | 8.897 | 9.915 | 11.591 | 13.240 | 15.445 | 17.182 | 20.337 | 23.858 | 26.171 | 29.615 | 32.671 | 36.343 | 38.932 | 46.797 |
| 22 | 9.542 | 10.600 | 12.338 | 14.041 | 16.314 | 18.101 | 21.337 | 24.939 | 27.301 | 30.813 | 33.924 | 37.659 | 40.289 | 48.268 |
| 23 | 10.196 | 11.293 | 13.091 | 14.848 | 17.187 | 19.021 | 22.337 | 26.018 | 28.429 | 32.007 | 35.172 | 38.968 | 41.638 | 49.728 |
| 24 | 10.856 | 11.992 | 13.848 | 15.659 | 18.062 | 19.943 | 23.337 | 27.096 | 29.553 | 33.196 | 36.415 | 40.270 | 42.980 | 51.179 |
| 25 | 11.524 | 12.697 | 14.611 | 16.473 | 18.940 | 20.867 | 24.337 | 28.172 | 30.675 | 34.382 | 37.652 | 41.566 | 44.314 | 52.620 |
| 26 | 12.198 | 13.409 | 15.379 | 17.292 | 19.820 | 21.792 | 25.336 | 29.246 | 31.795 | 35.563 | 38.885 | 42.856 | 45.642 | 54.052 |
| 27 | 12.879 | 14.125 | 16.151 | 18.114 | 20.703 | 22.719 | 26.336 | 30.319 | 32.912 | 36.741 | 40.113 | 44.140 | 46.963 | 55.476 |
| 28 | 13.565 | 14.847 | 16.928 | 18.939 | 21.588 | 23.647 | 27.336 | 31.391 | 34.027 | 37.916 | 41.337 | 45.419 | 43.278 | 56.893 |
| 29 | 14.256 | 15.574 | 17.708 | 19.768 | 22.475 | 24.577 | 28.336 | 32.461 | 35.139 | 39.087 | 42.557 | 46.693 | 49.588 | 58.302 |
| 30 | 14.953 | 16.306 | 18.493 | 20.599 | 23.364 | 25.508 | 29.336 | 33.530 | 36.250 | 40.256 | 43.773 | 47.962 | 50.892 | 59.703 |

*Source:* From *Statistical Tables for Biological, Agricultural, and Medical Research,* edited by R. A. Fisher and F. Yates. Pearson Education Limited. Copyright © 1938 by Oliver and Boyd. Reproduced with permission of the publisher.

# Glossary

**ABAB design** A single-subject experimental design in which baseline measurements of the target behavior are made (A) followed by a treatment (B) and then a second baseline measurement (A) and a second treatment (B).

**abstract** A brief summary of the contents of a document.

**accessible population** The population of subjects that is accessible to the researcher for a study, and the one to which the findings can be generalized.

**accidental sampling** A nonprobability sampling technique that simply uses conveniently available subjects without regard to how well the sample represents the population of interest.

**achievement test** A measure of the extent to which a person has acquired certain information or skills, often as a result of specific instruction.

**action plan** In action research, the step taken after analysis of data in which the researcher determines a course of action that incorporates potential solutions to the problem that was the focus of the research.

**action research** A form of research conducted by practitioners to study a particular context and use findings to change practice; typically uses qualitative research strategies.

**active deception** Deception occurring when subjects are deliberately given misinformation about the true purpose of the study.

**active independent variable** An independent variable created through active manipulation in the course of an experiment. Active independent variables are characteristic of experimental research.

**advanced memoing** A process of refining conceptual categories.

**aggregate databases** Large databases that combine single-topic databases.

**agreement coefficient** The percentage of persons for whom the same decision (mastery or nonmastery) is made on two administrations of a *criterion-referenced* test. Used as an index of reliability for criterion-referenced tests.

**alpha coefficient** *See* **Cronbach alpha**.

**alternate form techniques** *See* **equivalent form techniques**.

**alternate hypothesis** Another way of stating the research hypothesis.

**analysis of covariance (ANCOVA)** A statistical technique that provides partial statistical control for one or more variables, removing their influence from the comparison of groups on the dependent variable.

**analysis of variance (ANOVA)** An inferential statistical test used for experimental designs with more than one independent variable or more than two levels of an independent variable.

**anonymity** The practice of not associating individuals' names with the information or measurements obtained from those individuals.

**applied research** Research that aims to solve an immediate practical problem.

**aptitude test** A test that measures general abilities or characteristics believed to indicate a person's ability to learn a future task or to achieve in a particular area.

**attitude scale** A measure of the degree of favorableness or unfavorableness a subject has toward a group, institution, construct, or object.

**attribute independent variable** An independent variable on which subjects differ before a study begins. Attribute independent variables are characteristic of ex post facto research.

**attrition** *See* **experimental mortality**.

**audit trail** A qualitative researcher's documentation of how a study was conducted, including what was done, when, and why. It allows an independent auditor to examine the study from beginning to end and judge the trustworthiness of the outcome.

**authentic student work** Work done by students themselves as a primary source of material in action research.

**axial coding** The process of making connections between and across categories in qualitative data analysis.

**baseline** Measures of the dependent variable taken prior to the introduction of the treatment in a time-series experimental design and used as the standard of comparison.

**basic interpretative study** A form of qualitative research that provides a descriptive account targeted to understanding a phenomenon using data that may be collected in a variety of ways.

**basic qualitative study** *See* **basic interpretative study**.

**basic research** Research that aims to obtain empirical data that can be used to formulate, expand, or evaluate a theory rather than to solve a practical problem.

**Belmont Report** A report that identified three essential ethical principles that must be followed in all research with human subjects.

**between-groups design** An experimental design that uses different groups of subjects for the different treatment conditions being compared.

**biased sample** A sample in which certain elements in the parent population are systematically under- or overrepresented.

**bipolar adjective scale** A scale that presents a respondent with a list of adjectives to choose from that have bipolar or opposite meanings.

**Boolean search** Named after the 19th-century mathematician George Boole, a search allowing the inclusion or exclusion of documents through the use of Boolean operators AND, NOT, and OR.

**bounded system** The unit in a case study defined within specific boundaries or within a specific context.

**bracketing** *See* **epoche**.

**brainstorming** A method of finding a problem for research by listing topics of interest and developing specific questions for each topic.

**bricolage** A puzzle that connects the parts to the whole.

**bricoleur** A maker of quilts.

**canonical correlation** A statistical procedure for determining the relationship between several independent variables and more than one dependent variable.

**CAQDAS** Computer-assisted qualitative data analysis software.

**case study** A qualitative examination of a single individual, group, event, or institution.

**categorical exemptions** Specific exemptions from the current regulations covering projects funded by grants from the Department of Health and Human Services.

**categorical variables** Variables that differ in kind, not in degree or amount.

**categories** In coding, one level of abstraction above the initial codes.

**category scale** A rating scale consisting of a number of categories arranged in an ordered series.

**causal-comparative research** *See* **ex post facto research**.

**cause-and-effect relationship** An indication that one variable causes the observed effect on the other variable in an experiment.

**ceiling effect** An effect that occurs when the performance range on a measure is so restricted on the upper end that subjects cannot perform to their maximum ability.

**census** A survey that includes the entire population of interest.

**central tendency (measures of)** The three averages commonly used to summarize the data in a frequency distribution: mean, median, and mode.

**chain sampling** *See* **snowball sampling**.

**change score** The difference between subjects' pre- and postexperimental scores on the dependent variable(s).

**checklist** A type of survey question presenting a number of possible answers, from which respondents are asked to check those that apply to themselves.

**chi square ($\chi^2$)** An inferential statistic that compares the frequencies of nominal measures actually observed in a study with frequencies expected under a null hypothesis.

**classroom action research** Teachers carrying out research in their classrooms with the object of improving classroom practice or improving practices in the school.

**closed-ended question** A question followed by a fixed set of alternative responses from which the respondent chooses.

**cluster sampling** A probability sampling technique that randomly selects and uses whole naturally occurring groups such as intact classrooms or entire voting precincts as the samples (clusters).

**code–recode system** An intrarater system for assessing dependability, in which the data are coded, then left for a while, and then recoded by the researcher or a peer.

**coding system** A system used to facilitate the categorizing and counting of specific behaviors as they occur.

**coefficient alpha** *See* **Cronbach alpha**.

**coefficient of correlation** *See* **correlation coefficient**.

**coefficient of determination** The square of the correlation coefficient. It indicates the percentage of variance in one variable in common with another variable.

**coefficient of equivalence** A measure reflecting variations in performance from one specific set of items to another.

**coefficient of equivalence and stability** A coefficient reflecting variations in performance from one time to another and from one form of the test to another.

**coefficient of multiple correlation** A coefficient indicating the relationship between the predictor variables in combination and the criterion.

**coefficient of reliability** The proportion of the variance in the observed scores that is free of error.

**coefficient of stability** *See* **test–retest reliability coefficient**.

**Cohen's *d*** A measure of effect size found by dividing the difference between two sample means by the weighted average of their standard deviations.

**cohort study** A longitudinal survey study in which a specific population is studied by taking different random samples from the population at various points in time.

**collaborative action research** Multiple researchers sharing expertise and fostering dialogue among stakeholders.

**collaborative partner** A qualitative term, most often used in action or feminist research, that refers to the partnership between the researcher and the participant in the research process.

**collective case study** A study using several cases selected to further understand and investigate a phenomenon.

**common cause** A variable that influences both the independent variable and the dependent variable.

**comparative rating scale** A scale on which raters make judgments on subjects with reference to the position of other subjects on the same scale.

**comparison group** The group in a study that receives no treatment or a treatment different from the treatment received by the experimental group.

**compensatory demoralization** A threat to internal validity when subjects believe they are receiving less desirable treatment or are being neglected.

**compensatory rivalry** A tendency of the control group subjects who know they are in an experiment to exert extra effort and hence perform above their typical or expected average.

**complementarity** Seeking elaboration, illustration, enhancement, or clarification of findings from one method to another in mixed methods research.

**complete observer** A researcher who is hidden from the group under study or who observes public behavior in a public setting.

**complete participant** A member of a group under study who gathers data on the natural activity of the group without informing the other group members that they are under study.

**completion (fill-in) items** Questions to which respondents must supply their own answers in their own words.

**composite description** A description conveying the overall essence of a phenomenon.

**comprehensive sampling** A form of qualitative sampling in which every unit of interest is included.

**computer search** Using a computer to locate information in databases.

**concept mapping** Plotting the elements in an action research study diagrammatically to visualize what components of the situation relate to the problem under investigation.

**concern for context and meaning** The assumption that human behavior is context bound—that human experience takes its meaning from, and therefore is inseparable from, social, historical, political, and cultural influences.

**concurrent design** When both quantitative and qualitative data are collected separately but at approximately the same time.

**concurrent triangulation** Collecting qualitative and quantitative data separately but at the same time, with the findings converging in the conclusions.

**concurrent validity** The relationship between scores on a measure and criterion scores obtained at the same time.

**confidence band** The range of scores that will, with a given probability, include the individual's true score.

**confidentiality** The practice of keeping the information obtained from an individual in a study secret and private.

**confirmability** A term used in qualitative research, equivalent to validity in quantitative research, related to the degree to which findings in a study can be corroborated by others investigating the same situation.

**confirmatory factor analysis** An advanced factor analysis that allows for formal tests of hypotheses about the number and nature of factors and a comparison of different models or factor solutions.

**confounding** The mixing of variables extraneous to the research problem with the independent variable(s) in such a way that their effects cannot be separated.

**confounding variable** An uncontrolled extraneous variable whose effects on the dependent variable may incorrectly be attributed to the independent variable.

**constant** A characteristic that takes on the same value for all individuals in a study.

**constant comparison** A method of analyzing qualitative data that combines inductive category coding and simultaneous comparison of such units of meaning. Similar units of meaning are grouped together.

**constitutive definition** A definition in which a word is defined by using other words.

**construct** Abstractions that cannot be observed directly but only by their effects, such as creativity, anxiety, or intelligence.

**construct-irrelevant variance** The extent to which test scores are affected by variables that are extraneous to the construct.

**construct underrepresentation** A term that is applied to assessment that is too narrow and fails to include important dimensions of the construct. It threatens the validity of the assessment.

**construct validity (measurement)** The extent to which a test or other instrument measures what the researcher claims it does; the degree to which evidence and theory support the interpretations of test scores entailed by the proposed use of the test.

**construct validity (research)** The extent to which inferences made from the observed subjects, settings, and operations sampled to the constructs that the samples represent are justified.

**content analysis** A research method applied to written or visual materials to analyze characteristics of the material.

**content evidence (of validity)** Evidence based on the degree to which the items of a test representatively sample an intended content domain. Determined mainly by logical analysis.

**context stripping** Isolating human behavior from its context.

**contingency question** A survey question whose answer depends on the answer to a previous question.

**continuous variable** A variable whose measure can take an infinite number of points within a range.

**contrived observation** Observation that takes place when the circumstances have been arranged so that the desired behaviors are elicited.

**control** Steps taken by the researcher to remove the effects of any variable(s) other than the independent variable that might influence the dependent variable.

**control group** The group in a study that does not receive the experimental treatment; it is compared with the experimental group to determine the effects of the treatment.

**control group time-series design** A design that provides measurement of the experimental and control groups with the experimental treatment introduced into the series of measurements.

**convenience sampling** Choosing a sample based on availability, time, location, or ease of access.

**convergent design** Transforming data to a different form before analysis.

**convergent evidence** Evidence for test validity based on the relationship to other measures intended to assess a similar construct. It shows that the intended construct is being measured.

**correlation** A technique for determining the covariation between sets of scores; paired scores may vary directly (increase or decrease together) or vary inversely (as one increases, the other decreases **correlational research** Research that attempts to determine the extent and the direction of the relationship between two or more variables.

**correlation coefficient** A statistic that shows the degree of relationship between two variables.

**correlation matrix** A table that shows the coefficients of correlation between every measure and every other measure.

**corroboration** Agreement between procedures or data sources; used to establish dependability.

**counterbalanced design** An experimental design in which the effects of order are controlled by having all groups receive all treatments, but in a different order.

**covariate** The variable used in ANCOVA to adjust scores.

**cover letter** An introductory letter to respondents, mailed in advance or with a questionnaire, explaining the details about and the purpose of the study.

**covert participant** *See* **complete participant**.

**credibility** In qualitative research, the accuracy or truthfulness of the findings; similar in concept to internal validity in quantitative research.

**criterion-referenced test** An instrument that measures an individual's level of mastery of a predefined content or skill domain.

**criterion-related validity evidence** The degree to which scores on an instrument are related to other indicators of the same thing (the criterion). If the criterion scores are collected at the same point in time, the correlation is evidence of *concurrent validity*. If the criterion scores are collected at a later time, the resulting correlation between scores and criterion is evidence of *predictive validity*.

**criterion sampling** A form of sampling in which the researcher sets a criterion and includes all cases that meet the criterion.

**critical action research** Reseach focusing on evaluating social issues and using the results for social change.

**critical case sampling** Selecting a single unit that provides a crucial test of a theory or a program.

**critical ethnography** When the researcher is acting as an advocate for the group being studied.

**Cronbach alpha** An internal-consistency reliability coefficient that measures the extent to which the scores of the individual items agree with one another. It is especially useful for attitude scales or essay tests.

**cross-case comparison** A strategy used in qualitative research to enhance transferability by investigating more than one case to determine whether findings are similar.

**cross-sectional survey** A survey in which data are collected at one point in time from a specified population.

**cross tabulation** A table showing how frequently various combinations of two or more categorical variables occur, from which one can "see" the relationship (if any) between the variables.

**cross-validation** The process of validating a prediction equation based on one group by using it with a different group.

**cultural portrait** Incorporating the views of both participants and researchers in an ethnographic study.

**cumulative frequency** The number of cases in a distribution with scores equal to or less than a given score.

**curvilinear relationship** A correlational relationship where a scatterplot of the paired scores is better described by a curve than by a straight line.

**data analysis in action research** Reviewing the data as they are being collected and attempting to synthesize and make sense out of what is observed.

**data analysis spiral** The repeated use of organizing and familiarizing, coding and reducing, and interpreting and representing data during qualitative analysis.

**data comparison** Comparing data from the qualitative and quantitative sources in mixed methods research.

**data consolidation** Combining both sets of data to create a new set of data or variables.

**data correlation** Comparing the data from different analyses (quantitizing and qualitizing from the originals).

**data display** Using tables and graphs to display quantitative data and other forms, such as matrices, rubrics, and lists, to describe the qualitative data in mixed methods research.

**data integration** Integrating data into a coherent whole from qualitative and quantitative sets of data.

**data interpretation** Focusing on the implications or meanings that emerge from the data analysis in mixed methods research.

**data reduction** Analyzing the qualitative data via theme analysis or thematic coding while also analyzing the quantitative data via descriptive statistics, factor analysis, or other quantitative methods.

**data saturation** In qualitative research, the point at which no new information is forthcoming from additional participants or settings.

**data triangulation** Investigating whether the data collected with one procedure or instrument confirm data collected using a different procedure or instrument.

**deception** *See* **active deception, passive deception**.

**deduction** Beginning with general premises or already known facts and deriving specific logical conclusions.

**deductive hypothesis** A hypothesis derived by deduction from a theory.

**deductive reasoning** A thinking process in which one proceeds from general to specific knowledge through logical argument.

**degrees of freedom (*df*)** The number of observations free to vary around a constant parameter. Each inferential statistic has a defined procedure for calculating its degrees of freedom, which are used to determine the appropriate critical values in statistical tables for determining the probabilities of observed statistics.

**delta (Δ)** Smith and Glass's effect size, defined as the experimental group mean minus the control group mean divided by the control group standard deviation.

**dependability** In qualitative research, the consistency or stability of the results; the extent to which the same general results would occur with different sets of people or in different settings and time periods. Somewhat equivalent to reliability in quantitative research.

**dependent variable** A variable that is a consequence of or dependent on an antecedent (independent) variable.

**descriptive adequacy** The qualitative researcher's responsibility to provide accurate, detailed, and complete descriptions of the context and participants in a study so that the reader can determine trans-ferability.

**descriptive research** Research that asks questions about the nature, incidence, or distribution of variables; it involves describing but not manipulating variables.

**descriptive statistics** Techniques for organizing, summarizing, and describing observations.

**descriptors** Key words used in computer searching of databases. Usually part of a controlled vocabulary, accessible by using a thesaurus specific to that database.

**development** In mixed methods research, using the results from one approach to develop or inform the other approach.

**developmental study** Investigation of how children change as they mature and respond to their environment.

**deviant case sampling** *See* **extreme case sampling**.

**deviation scores** Scores showing the difference between a raw score and the mean of the distribution.

**dichotomous variable** A categorical variable that has only two classes.

**differential selection** Creating experimental and control groups in such a way that they differ before treatment; an internal validity problem.

**diffusion** A threat to internal validity occurring when participants in an experimental group communicate information about the treatment to subjects in the control group in such a way as to influence the latter's behavior on the dependent variable.

**directional hypothesis** A hypothesis that specifies the direction of the expected findings—that is, whether a "greater than" or "less than" result is expected.

**directional test (one-tailed test)** A statistical test of a hypothesis that only states that the population parameter is greater than or less than the value of the parameter specified in the null hypothesis.

**directly administered questionnaire** A questionnaire administered to a group of people assembled at a specified place and time for that purpose.

**direct observation** Observation of subjects to record the frequency and duration of behaviors falling within predetermined categories.

**discourse analysis** *See* **semiotics**.

**discrepant data analysis** *See* **negative case analysis**.

**discriminant analysis** A correlational procedure using a number of predictor variables to classify subjects into two or more distinct groups.

**discriminant evidence** Evidence for test validity based on the relationship between the test scores and measures of purportedly different constructs. It shows that the wrong construct is not being measured.

**documentary analysis** The systematic examination of documents to investigate specific topics or themes.

**double-barreled question** A survey question that actually asks two questions in one.

**double-blind experiment** An experiment in which neither the observers nor the subjects know which is the experimental group and which is the control group.

**double-dipped sample** A sample drawn from the nonrespondents to a survey, who are then interviewed to determine if they differ significantly from respondents.

**dummy variable** A categorical variable that has been recoded as a binary variable for entry into a multiple regression.

**effect size** A measure of the strength of the relationship between two variables shown by the difference between two means calculated by either Cohen's *d* or Smith and Glass's delta.

**effect variable** *See* **dependent variable**.

**elicited text** Information obtained by the researcher.

**elite interviewing** Interviewing individuals selected on the basis of their expertise.

**emergent design** A design typical of qualitative research in which an initial problem or sampling strategy is changed or refined as data are collected and analyzed.

**embedded design** Using one form of data to support another within a single study.

**emic perspective** The perspective of individuals who are part of a case study.

**enquiring** Gathering data by questioning participants.

**epoche** When the researcher intentionally sets aside his or her own experiences or suspends his or her own beliefs.

**equivalent form** Reliability assessment procedure correlating the scores of the same subjects on two tests that are as similar as possible in content, difficulty, length, format, and so on.

**error of central tendency** A tendency for a rater to avoid extreme scale positions, rating all individuals near the middle of the scale.

**error of severity** A tendency to rate all subjects too low.

**error term** The average value expected through chance alone.

**essence** A relational term referring to the intentionalities of our world.

**ethical standards** A set of standards designed specifically to guide the work of researchers, specifying their obligations to their subjects and their profession.

**ethnography** A form of qualitative research that aims for a holistic picture of a cultural group; it uses in-depth interviewing and prolonged participant observation.

**etic perspective** The interpretations made by the researcher in a case study.

**examining** Gathering data by examining materials already existing or collected in the setting.

**expansion** Expanding the breadth and range of the inquiry by using different methods for different components of the study.

**expected frequencies** Theoretical frequencies that would be observed when the null hypothesis is true.

**experiencing** Gathering data through the researcher's own experience.

**experimental design** A plan for an experiment that specifies what independent variables will be applied, the number of levels of each, how subjects are assigned to groups, and the dependent variable.

**experimental group** The group in a research study that receives the experimental treatment.

**experimental mortality (attrition)** A threat to internal validity occurring when there is a different loss of participants from the comparison groups in an experiment.

**experimental research** Research in which the investigator manipulates one or more independent variables (the treatment) and observes the effect on one or more dependent variables.

**experimental variable** *See* **independent variable**.

**experimenter bias** The effects of the experimenter's attitudes, behavior, and expectancies on the behavior of the subjects in an experiment.

**experimenter effect** A threat to internal validity referring to unintentional effects that the researcher may have on the study.

**explanation** A strategy for trying to determine a specific problem for investigation by hypothesizing about factors affecting a situation.

**exploratory factor analysis** A factor analysis in which no formal hypotheses are tested about the nature of the factors that might be found.

**ex post facto research** A type of research that attempts to determine the causes for, or the consequences of, differences that already exist in groups of individuals.

**extant text** Information that the researcher did not shape.

**extended fieldwork** Investigation carried out over a prolonged period of time in the field.

**extended observation** Long-term observation carried out during an ethnographic study by a researcher immersed in the lives of the group being studied.

**extension questions** Questions asking for more detail about a topic.

**external criticism** Evaluation of the authenticity of a document or other concrete evidence in historical research.

**external validity** The extent to which the findings of a particular study can be generalized to other subjects, other settings, and/or other operational definitions of the variables.

**extraneous variable** An uncontrolled variable that may affect the dependent variable of a study; its effect may be mistakenly attributed to the independent variable of the study.

**extreme case sampling** Selecting an atypical or unusual unit for qualitative study.

**face validity** The extent to which examinees believe the instrument is measuring what it is supposed to measure.

**factor analysis** A statistical procedure for analyzing the intercorrelations among three or more measures that reduces the set to a smaller number of underlying factors.

**factorial design** An experimental design that investigates two or more independent variables at the same time in order to study their effects singly or in interaction with each other.

**factor loading** In factor analysis, the correlation between a variable and a factor. The higher the loading, the more the variable is seen as contributing to a particular factor.

**factors** Independent variables in factorial designs.

**falsifiable** A term applied to a theory or hypothesis that is capable of being proven wrong.

**familiarization** The process of becoming familiar with the data obtained in a qualitative study through spending time reviewing the data.

**feminist research** *See* **critical research**.

**field experiment** An experiment conducted in a naturally occurring environment.

**field notes** The researcher's reflections and the written records of observations and conversations made during a qualitative research project.

**field testing** A process of testing a research instrument with a small number of persons in order to identify ambiguities or other problems before the final form is prepared.

**fill-in items** *See* **completion items**.

**floor effect** An effect that occurs when the performance range of a measure is so restricted at the lower end that subjects cannot show their ability to perform. The floor effect occurs when the task is too difficult for the subjects.

**focused coding** Sorting codes to facilitate synthesizing and integrating theoretical ideas.

**focused interview** A qualitative research data-gathering technique employing open and flexible questions that respondents are free to answer in their own words.

**focus group** A data-gathering tool in which a researcher interviews a small group of people to obtain different perspectives on a particular issue.

**focus of inquiry** In qualitative research, the initial topic (problem) that a researcher pursues. It may be presented in the form of a question or a statement.

**follow-up study** An investigation of subsequent development of subjects after a specified treatment or program.

**foreshadowed problems** Problems that emerge from the general area of a particular concern or topic.

***F* ratio** The between-groups mean square/within-groups mean square ratio.

**frequency distribution** A tabular representation of the scores obtained by a group of individuals.

**frequency polygon** A graphic representation of data, preferred when data are of a continuous nature.

***F* test** A statistical procedure used for testing hypotheses about differences among two or more means, and for other purposes.

**fully mixed design** A design involving mixing of the qualitative and quantitative approaches in an interactive way throughout the study.

**generic questions** Asking samples about broader population. More typically used in qualitative research.

**generosity error** A tendency for a rater to give every subject the benefit of the doubt and, when uncertain, to give high ratings.

**grand-tour questions** Global questions allowing participants to describe something in their own terms.

**graphic scale** A scale in which respondents place a check on the appropriate point on a horizontal line that runs from one extreme to the other of the behavior in question.

**grounded theory** Theory derived inductively from the data collected in a natural setting rather than from a priori ideas or theories

**halo effect** The tendency for a rater's general impression of a subject to influence the rating given for specific aspects or behaviors of the subject.

**Hawthorne effect** An effect on the dependent variable resulting from the treatment group's knowledge that the members are participating in an experiment.

**histogram** A graph in which the frequency distribution of scores is represented by vertical bars, preferred when data are discrete.

**historical research** A systematic attempt to establish facts and arrive at conclusions about the past.

**history effect in qualitative research** The fact that unique historical experiences of the participants may militate against comparison, threatening transferability.

**history effect in quantitative research** A threat to internal validity when extraneous events occurring at the same time as the treatment could produce the observed outcome even without treatment.

**homogeneity measures** Used to assess the inter-item consistency of the items on a test.

**homogeneous group** A set of subjects who are the same or similar on a relevant variable.

**homogeneous sampling** Selecting subsets of units for qualitative study because they are similar in the phenomenon of interest.

**homogeneous selection** Employing subjects who are as alike as possible in regard to relevant characteristics, to reduce the influence of extraneous variables.

**horizontalization** Identifying significant statements or quotes in phenomenological research.

**human as instrument** The investigator as the means of data gathering in qualitative inquiry.

**hypothesis** A tentative proposition suggested as a solution to a problem; a statement of the researcher's expectations about the relationship among the variables of a study.

**hypothesis testing** Collection of observations to determine whether these observations confirm or fail to confirm a hypothesized relationship.

**imperfect induction** The process of inferring from observations made on a sample of a group to what is characteristic of the whole group.

**implementation threat** A threat to the internal validity of a study resulting from possible variations in implementing the treatment.

**independent variable** A variable that is antecedent to the dependent variable.

**induction** Gaining knowledge through systematic study and observation of specific facts or episodes and then arriving at generalized conclusions.

**inductive analysis** A process of beginning with observations (data) and then proceeding to hypothesis formation and refinement and finally to theory.

**inductive coding** An initial stage of coding individual data pieces with simultaneous comparison of units of meaning.

**inductive data analysis** Proceeding from data to theory or interpretation.

**inductive hypothesis** A hypothesis derived through generalization from observation.

**inductive reasoning** Reaching a conclusion by generalizing from examples to the whole class or category.

**inferential statistics** Procedures that permit one to make tentative generalizations from sample data to the population from which the sample was drawn.

**informants** The term used in ethnographical studies in preference to participants.

**informational adequacy** Ensuring that the researcher will understand the setting thoroughly, precisely, and adequately.

**informed consent, right to** The right of a subject in a research study to know the nature and purpose of the study and to give or withhold consent to participate.

**initial coding** Sorting data by attaching labels to segments. The first step in coding.

**initial memos** Preliminary analytic notes.

**initiation** Discovering paradoxes or contradictions that might lead to the reframing a theory.

**institutional review board (IRB)** A committee that determines whether proposed research meets federal and other legal and ethical standards.

**instrument** A device for operationally defining a variable.

**instrumental case study** A case selected because it can help provide insight into some other issue under investigation.

**instrumentalist questions** Questions formulated in terms of observable, measurable data.

**instrumentation threat** A threat to internal validity occurring as a result of changes in the instruments used in the study.

**intangibles** Constructs that are not directly observable but, rather, must be inferred from indirect measures.

**intensity sampling** Selecting units for qualitative study because they exhibit different levels of a phenomenon of interest.

**interaction** The outcome of a factorial experiment when the effects of one independent variable on a dependent variable change at different levels of the second independent variable.

**interaction effect** The effect of different combinations of two independent variables on a dependent variable.

**interdisciplinary triangulation** Using other disciplines to infer processes and understandings of the findings of a qualitative study.

**internal consistency measures** Procedures to determine whether all the items in a test are measuring the same thing.

**internal criticism** Evaluation of the truthfulness of a document in historical research.

**internal structure evidence of validity** Evidence based on an analysis of the internal structure of a test; it examines the intercorrelations among test items to determine whether the test is measuring a single construct and whether the interrelationships conform to the theory behind the construct being measured.

**internal validity** The extent to which observed differences on the dependent variable in an experiment are the result of the independent variable, not of some uncontrolled extraneous variable(s).

**interobserver agreement** A qualitative strategy for enhancing the dependability of findings by comparing coding of a peer observer with the original coded data.

**interobserver reliability** *See* **interrater reliability**.

**interpretation** Reflecting on and bringing out meaning, telling the story, and providing an explanation of the observations in a qualitative study.

**interpretive accuracy** *See* **referential accuracy**.

**interpretive adequacy** The degree to which the participants' viewpoints, thoughts, feelings, intentions, and experiences are accurately understood and portrayed.

**interpretive rigor** *See* **rigor**.

**interpretive study** A basic form of qualitative research that provides detailed accounts targeted to understanding a process, a phenomenon, or a particular point of view.

**interrater agreement** The degree to which the ratings of two independent raters agree.

**interrater reliability** The extent to which two or more observers produce similar quantitative results when observing the same individual during the same time period.

**interval scale** A scale of measurement that orders objects or events and has points equidistant from one another.

**interview** Oral questioning of a subject.

**interviewer bias** A bias that occurs in interviews when the interviewer's own feelings or attitudes influence the way questions are asked or interpreted.

**interview schedule** A document used by an interviewer that contains the instructions, the questions in a fixed order, and any transition phrases.

**intrarater agreement** The degree to which the ratings made by a single rater agree with each other.

**intrinsic case study** A case study conducted to understand a particular case.

**invariant structure** *See* **composite description**.

**inventory** A collection of statements to which subjects respond by indicating whether each statement describes them or not; used in assessing personality.

**investigator triangulation** Having multiple researchers collect data independently and comparing their collected data.

***in vivo* codes** Codes named from the actual words of respondents.

**item analysis** Analysis of a test determining for each item the number and proportion of correct responses and the correlation of scores on that item with total test scores.

**John Henry effect** An effect that occurs when the control group performs above its usual average when it perceives itself in competition with an experimental group that is using a new method or procedure.

**journaling** Making notes to assist the reflective process.

**judgment sampling** *See* **purposive sampling**.

**kappa coefficient** The proportion of consistent classifications observed *beyond* that expected by chance alone.

**known-groups technique** A method for determining the validity of a measure by seeing whether groups known to differ on the construct also differ on the measure itself.

**Kruskal–Wallis test** A one-way analysis of variance carried out on ranks.

**Kuder–Richardson formulas** Formulas for determining the internal consistency reliability (homogeneity) of an instrument from a single administration and without splitting the test into halves.

**laboratory experiment** An experiment conducted in a highly controlled environment.

**law of the single independent variable** Rule stating that to unambiguously determine the effect of an independent variable on a dependent variable, the independent variable must be the only difference between the experimental and control groups before the dependent variable is measured.

**level of significance** The largest probability of error acceptable for rejection of the null hypothesis.

**Likert scale** A measurement scale consisting of a series of statements followed by five response categories, typically ranging from "strongly agree" to "strongly disagree."

**Likert-type item** A statement similar to those in a Likert scale, where the response options are on a continuum from "strongly agree" to "strongly disagree."

**linear relationship** A correlational relationship where the plotted points in a scatterplot of the paired scores fit around a straight line.

**longitudinal survey** A survey in which data are collected several times over an extended period of time.

**low-inference descriptor** Verbatim or direct quotations from the participants in qualitative research that help a reader experience the participants' world.

**main effect** The effect that an independent variable has on the dependent variable in a factorial experiment irrespective of the influence of the other independent variables.

**manipulated variable** The variable that is systematically altered by an experimental researcher in order to observe the effect on a dependent variable.

**Mann–Whitney test** A statistical test for the difference in the group means for two independent samples when the dependent variable is ranked data.

**margin of error** An estimate of the extent to which sample results are likely to deviate from the population value.

**matching** Identifying pairs of subjects who are as alike as possible before a study begins.

**matrix sampling** Randomly selecting respondents, each of whom is administered a random subset of questions.

**maturation** Naturally occurring changes in subjects due to the passage of time, which may threaten internal validity.

**maximum variation sampling** Including units that will maximize differences on specific characteristics.

**mean** A measure of central tendency for a distribution of interval data; the sum of the scores divided by the number of scores in the distribution; the arithmetic average.

**median** The point in a distribution below which are 50 percent of the scores; used with ordinal or interval data.

**member check** A process in which a qualitative researcher asks the participants in a study whether they have accurately and realistically described their experience. The participant feedback contributes to the trustworthiness of qualitative inquiry.

**meta-analysis** The systematic combination of quantitative data from a number of studies investigating the relationship between the same variables. A weighted average of effect sizes.

**methodological rigor** *See* **rigor**.

**methods triangulation** Using more than one method on the assumption that the combination results in better evidence.

**mixed methods research** Combining quantitative and qualitative research in different ways, with each approach adding something to the understanding of the phenomenon.

**mixed model research** When mixing of qualitative and quantitative approaches occurs in all stages of a study or across stages of a study.

**mode** The score that occurs most frequently in a distribution of scores; used with nominal, ordinal, interval, or ratio data.

**monomethod research** Using one method, either qualitative or quantitative, with corresponding data collection, analysis, and accompanying procedures to answer the research question.

**mortality** A threat to internal validity that occurs when some participants drop out of a study before it is completed.

**Moses effect** The gap occurring when researchers hand down edicts with the expectation that teachers will be passive recipients of these edicts.

**multifactor analysis of variance** An analysis of variance with more than one independent variable.

**multimethod research** Using two different data collection procedures from the same research tradition (quantitative or qualitative) or combining two research methods from the same research tradition.

**multiple-baseline design** Observations made on one or several participants using two or more AB units.

**multiple case study** *See* **collaborative case study**.

**multiple regression** The prediction of a criterion using two or more predictor variables in combination. Each predictor is weighted in proportion to its contribution to prediction accuracy. The equation showing the weights assigned to each predictor is the *multiple regression equation*.

**multitrait–multimethod matrix** A way to simultaneously evaluate the convergent and discriminant validity of a construct.

**narrative study** A form of qualitative research, also called *biography, life stories*, and *life narratives*, that provides accounts of a person's experiences.

**naturalistic inquiry** The study of subjects in their normal environment without predetermined hypotheses.

**naturalistic observation** A qualitative research term that indicates that the researcher is not attempting to alter a situation in any way but is merely observing and recording events as they naturally occur.

**natural setting** A research environment in which activities occur in the ordinary course of events as opposed to an environment in which events are contrived or manipulated for the purpose of the research.

**negative case analysis** A strategy used in qualitative research that involves intentionally searching for data that contradict the emerging category or pattern.

**negative case sampling** Selecting units that are examples of exceptions to expectations. Can be used to control researcher bias.

**negative correlation** A correlation with high scores on one variable associated with low scores on the other variable.

**negative evidence** Evidence that contradicts a theory and thus leads to its rejection or revision.

**negatively skewed curve** A polygon showing many scores piled up at the high (right) end and the lower scores spread out toward the low (left) end; the tail of the curve extends to the left.

**negative results** Results opposite to those hypothesized.

**network sampling** *See* **snowball sampling**.

**neutrality** In qualitative research, the extent to which the research is free of bias.

**nominal scale** A scale of measurement that classifies objects or individuals into categories that are qualitatively but not quantitatively different.

**nondirectional hypothesis** A hypothesis that states that a relationship between variables will be observed but does not specify the direction of the expected findings.

**nondirectional test (two-tailed test)** A statistical test that takes into account differences in either direction (greater or less than) from the value specified in the null hypothesis; the null hypothesis is rejected if the difference is large enough in *either* tail of a sampling distribution of the statistic.

**nonexperimental research** Research that does not involve the manipulation of variables.

**nonparticipant observation** A research method in which the observer does not participate in the situation being studied but is only an observer.

**nonprobability sampling** Sampling through other than random selection.

**nonrandomized control group pretest–posttest designs** Designs using groups already organized into classes or other intact groups.

**nonresponse** Failure to return the completed instrument in a survey.

**normal curve** A hypothetical symmetrical, bell-shaped distribution of scores used as a model for many naturally occurring distributions and for many statistical tests.

**normal distribution** A spread of cases resembling the normal curve, with most cases concentrated near the mean.

**norm-referenced test** A test that enables one to compare an individual's performance with the performance of others who have previously taken the test.

**notation system** A method of indicating when qualitative and quantitative methods are used during the investigation.

**null hypothesis** A hypothesis that states there is no effect, no difference, or no relationship between variables; it is a negation of the research hypothesis—hence one that the researcher hopes to reject.

**numeric scale** *See* **graphic rating scale**.

**observed frequencies** The actual frequencies obtained by observation.

**observer bias** An effect that occurs when the observer's personal attitudes and values affect the observation and/or the interpretation of the observation.

**observer effect** The impact that the observer's presence has on a study's participants so that they behave differently from the way they normally would.

**observer expectation** A tendency for an observer who knows the participants he or she is studying to expect certain behaviors and hence to see or interpret actions or events in a certain way.

**observer participant** An observer who interacts with subjects enough to establish rapport but does not really become involved in the behavior or activities of the group. The person's status as an observer/researcher is known to the group.

**one-group pretest–posttest design** A preexperimental design involving a pretest, experimental treatment, and posttest to a single group of subjects.

**one-group time-series design** Periodic measurements of one group, introducing treatment into this series of measurements.

**one-tailed test** *See* **directional test**.

**open coding** The process of breaking down and categorizing qualitative data into manageable segments.

**open-ended question** A question that does not have fixed response alternatives but allows the respondent to respond as he or she chooses.

**operational definition** A definition that specifies the procedure or operation to be followed in producing or measuring a concept.

**opportunistic sampling** Taking advantage of all new leads or opportunities.

**ordinal scale** A scale of measurement that rank-orders objects and events according to the extent to which they possess the characteristic of interest, but in which the distance between points cannot be assumed to be equal.

**organization** Putting data together for easy retrieval.

**organizational categories** Categories that may have been anticipated or that have been established prior to data collection.

**outcome variables** *See* **dependent variables**.

**panel study** A longitudinal survey study in which data are gathered from the same sample of people at different points in time.

**parallel design** A design in which data are collected and analyzed separately.

**parallel forms technique** *See* **equivalent forms technique**.

**parameter** A characteristic of a population, such as the population mean ($\mu$) or the population standard deviation ($\sigma$).

**parental consent** *See* **informed consent**.

**parsimony principle** *See* **principle of parsimony**.

**partial control** Some but not complete control over relevant extraneous variables.

**partial correlation** A statistical procedure for describing the relationship between two variables, with the correlation between these variables and a third variable removed.

**participant observation** A research method in which the researcher becomes a part of and participates in the activities of the group or situation being studied.

**participatory action research** Involving stakeholders in action research.

**particularizing questions** Questions asking about a specific context.

**passive deception** Deception occurring when subjects are not given sufficient information about the purpose of the study.

**path analysis** A statistical procedure for investigating the causal relationships among correlated variables.

**pattern matching** Making predictions based on theory and then determining how well patterns in the data fit the predicted pattern.

**pause** A few moments of silence in an interview until the respondent answers.

**Pearson product moment coefficient (Pearson *r*)** An index of correlation for interval or ratio data; it is the mean of paired *z* score products of the two variables.

**peer debriefers** Members of a qualitative research team who check one another's work and look for evidence of bias. They act to keep one another honest.

**peer review** Discussion among the researcher's peers to determine whether his or her interpretation of the data is reasonable.

**percentile rank** A measure of relative position indicating the percentage of scores in a distribution that fall below a given score point.

**perfect induction** Can only occur when all examples are examined before generalizing to the whole category.

**performance assessment** A type of assessment that requires subjects to show what they know by performing specified tasks rather than by filling out a paper-and-pencil instrument.

**performance test** *See* **performance assessment**.

**personal interview** When the interviewer reads the questions to the respondent and records the answers in a face-to-face setting.

**phenomenological approach** A philosophy of research that focuses on understanding the meaning events have for people in particular situations.

**phenomenological interviewing** Examining lived experience through a series of in-depth interviews.

**phenomenology** A type of qualitative research designed to gain an understanding of how participants experience and give meaning to an event, concept, or phenomenon.

**phi coefficient** A correlation coefficient for use when both variables are measured on a nominal scale.

**pilot study** A trial run with a few subjects to assess the appropriateness and practicability of the procedures and data-collecting instruments.

**placebo** A chemically inert or neutral material given to subjects in an experiment to make them believe they are receiving a drug or experimental treatment.

**plausibility** *See* **theoretical adequacy**.

**polygon** A graph showing scores and frequency of scores by connecting the intersections of each of the scores with their frequencies.

**population** The larger group to which a researcher wishes to generalize; it includes *all* members of a defined class of people, events, or objects.

**portraiture** A form of qualitative research whose goal is to paint a vivid story that reflects the perspectives of both the researcher and the participants.

**positive correlation** A correlation with high scores on one variable associated with high scores on the other variable, and low scores associated with low scores.

**positively skewed distribution** A polygon showing many scores piled up at or near the low (left) end and higher scores spread out toward the high (right) end; the tail extends to the right.

**positivism** A philosophy of research characterized by objective inquiry based on measurable variables, believing that science should be primarily concerned with the explanation and the prediction of observable events. It underlies quantitative research.

**post hoc fallacy** Mistakenly attributing causation based on a relationship between two variables.

**power** The ability of a significance test to find significant differences when differences truly exist; the ability to avoid making a Type II error.

**power calculation** A mathematical procedure for determining the sample size needed to reject the null hypothesis at a given level of significance with a given effect size.

**practitioner research** *See* **action research**.

**pragmatic approach** Focusing on "what works," rather than choosing between methods considered to be pragmatically incompatible, to answer the research question.

**prediction studies** Using one variable to predict another when the two variables are correlated.

**predictive validity evidence** The relationship between scores on a measure and criterion scores available at a future time.

**predictor** An indicator of future performance based on previous performance.

**preexperimental research** Experimental designs with little or no control of extraneous variables and therefore little internal validity.

**pretest sensitization** The effect of a pretest on subjects that causes them to respond differently to the treatment from the way they would without the pretest.

**primary source** Original documents, relics, remains, or the records of an eyewitness used in historical research.

**principle of parsimony** The principle that states that the simplest explanation of a phenomenon is to be preferred over a more complicated explanation.

**privacy** Specifies that any data that identifies a student may not be made available without written permission.

**proactive ex post facto research** When the subjects are grouped on the basis of an independent variable.

**probability sampling** Sampling employing random selection, which means that every element in the population has a nonzero chance of being selected.

**probe** A follow-up question used during an interview to elicit clearer and more complete responses from the interviewee.

**problem analysis** Using visual representations to show antecedents to and consequences of the problem under investigation in action research.

**process questions** Questions asking about how things happen.

**Professional Development School** A partnership between a college/university and a K–12 school designed to enhance professional development, provide clinical experiences for preservice teachers, conduct inquiry to improve practice, and enhance student achievement.

**projective technique** A method for assessing personality by analyzing subjects' responses to ambiguous material.

**prompt questions** Questions used to elicit more details about a topic.

**purposive sampling** A nonprobability sampling technique in which subjects judged to be representative of the population are included in the sample.

**qualitative research** A generic term for a variety of research approaches that study phenomena in their natural settings, without predetermined hypotheses.

**qualitization** Changing quantitative data into qualitative terms.

**quantitative research** Inquiry employing operational definitions to generate numeric data to answer predetermined hypotheses or questions.

**quantitization** Changing qualitative data into quantitative data.

**quasi-experimental research** Research in which the investigator can control the treatment and the measurement of the dependent variable but cannot control assignment of the subjects to treatment.

**questionnaire** An instrument in which respondents provide written responses to questions or mark items that indicate their responses.

**quota sampling** A nonprobability sampling technique that determines the size of each relevant subgroup in the population and then nonrandomly selects subjects to produce a sample in which the proportion of each subgroup is the same in the sample as it is in the population.

**random assignment** Assigning members of a sample to experimental or control groups through a chance procedure.

**random errors of measurement** Errors that are the result of pure chance.

**randomization** *See* **random assignment**.

**randomized experiment** An experiment in which subjects are randomly assigned to groups.

**randomized matched subjects, posttest only control group design** A design in which subjects are matched and randomized and a control group is used, but with no pretest.

**randomized matching** A procedure in which subjects are first paired (matched) on relevant variables and then a chance procedure is used to assign one member of each pair to the experimental group and the other to the control group.

**randomized subjects, posttest only control group design** A powerful design involving the use of randomization and a control group with no pretest.

**random purposeful sampling** Randomly selecting a sample from a previously selected larger group.

**random sample** A sample selected by a chance procedure so that every member of the population has an equal probability of being selected.

**random selection** The process of selecting a sample by chance means so that every member of the population has an equal probability of being included.

**range** A nominal measure of dispersion; the difference between the highest and lowest scores plus one unit of measure.

**ranking item** A type of survey question that asks respondents to indicate the order of their preference among a number of options.

**ratio scale** A scale of measurement that provides a true zero point as well as equal intervals.

**reactivity** The effect of the research on the participants.

**realist ethnography** Providing an objective account of a situation.

**realist questions** Questions asked about unobserved phenomena that are treated as real.

**reducing** The sorting of codes into categories and categories into themes.

**reduction** A phenomenological device aiming to bring aspects of meaning into focus.

**redundancy** *See* **data saturation**.

**referential accuracy** A term used in qualitative research to refer to the degree to which participants' viewpoints, thoughts, feelings, and experiences are accurately understood.

**reflection** A strategy for identifying problems in action research and continually reviewing the data as the action research process unfolds.

**reflective log** Notes made while researchers are familiarizing themselves with their data.

**reflexivity** Qualitative researchers' use of self-reflection to recognize and actively search for their own biases.

**regression** *See* **statistical regression**.

**regression threat** An internal validity problem that arises when results in a study are caused by a tendency for groups selected on the basis of extreme scores to move (regress) toward the average on subsequent measures, regardless of the experimental treatment.

**reliability** The extent to which a measure yields consistent results; the extent to which scores are free of random error.

**reliability coefficient** *See* **coefficient of reliability**.

**repeated-measures design** *See* **within-subjects design**.

**replication logic** A process whereby one conducts a qualitative study in multiple locations or with multiple groups in order to assess the dependability of the findings.

**representation** Methods for presenting data from qualitative research studies.

**researcher biases** Biases held by the investigator that may be a source of invalidity in qualitative research studies.

**researcher-made test** A test constructed by a researcher for a specific study.

**research hypothesis** A statement of the relationships one expects to find as a result of the research.

**research plan** A preliminary proposal for research indicating the steps to be followed.

**research proposal** A step-by-step plan for conducting and completing a research study.

**researcher's relationship to the participant** The changing relationship over time between the researcher and the participant, from researcher to friend.

**researcher's role** The researcher's level of involvement in the research setting.

**response process evidence (of validity)** Evidence about what a test measures obtained by looking at the mental processes and skills that subjects use when responding to the items on a test.

**response rate** The proportion of the selected sample that agrees to be interviewed or returns a completed questionnaire.

**response set** A habitual way of responding that is independent of the content of a particular survey or test item, as when a subject always checks "undecided" regardless of the statement.

**restorying** Retelling a participant's story.

**retroactive ex post facto research** Research seeking possible antecedent causes (independent variables) for a preexisting dependent variable.

**reverse causality** An alternative hypothesis stating that the supposed dependent variable is in actuality the independent variable ($Y$ causes $X$, not $X$ causes $Y$).

**rich, thick description** *See* **thick description**.

**rigor** In qualitative research, somewhat the equivalent of reliability and validity in quantitative research.

**sample** A group selected from a population for observation in a study.

**sample survey** A survey done on only a portion (a sample) of a population, often using a questionnaire or interview.

**sampling error** The difference between a sample statistic and a population parameter.

**sampling frame** A complete list of all individuals in the population.

**scale** A continuum, usually having quantitative units, that measures the extent to which individuals exhibit certain traits or characteristics.

**scaled items** A set of concepts, events, or situations presented to respondents together with a scale of responses on dimensions such as quantity or intensity of feelings or how much, how well, how often.

**scatterplot** A figure using plotted points to represent the intersection of individuals' paired $X$ and $Y$ scores; used to show the relationship between the $X$ and $Y$ variables.

**scholastic aptitude test** A test that measures abilities that are predictive of success in academic tasks.

**school-based study group** A means for collaboration and a source for energy and support for action research in schools.

**scientific approach** A way of seeking knowledge that involves both inductive and deductive reasoning to develop hypotheses that are then subjected to rigorous and objective testing.

**secondary source** Secondhand information, such as a description of a historical event written by someone other than an eyewitness.

**selection** A threat to internal validity in quantitative research occurring when there are important differences between the experimental and control groups before the experiment begins.

**selection effects** In qualitative research, the fact that the constructs being investigated may be unique to a single group, which affects transferability.

**selection–instrumentation interaction** A problem occurring as the result of a ceiling effect limiting the range of achievement on a test.

**selection–maturation interaction** An internal validity problem created when subjects are selected in such a way that the experimental and control groups mature at different rates.

**selection–regression interaction** A problem that may occur if groups are drawn from populations having different means.

**selective coding** The process of systematically reviewing qualitative data to look for a specific category or theme.

**self report inventory** *See* **inventory**.

**semiotics** A qualitative study of linguistic units to examine the relationship between words and their meanings.

**sense-making** Considering how the data fit together and what stands out.

**sensitizing concepts** General concepts that provide a place to begin discussion of a grounded theory research study.

**sequential design** A design in which the analysis of data collected and examined in an earlier phase of a research study will influence choices made in the next phase.

**serendipitous discoveries** Unhypothesized relationships appearing in a study.

**setting effects** The fact that results may be a function of the specific context under investigation, which threatens transferability.

**significance fallacy** An incorrect assumption that the statistical significance of results also indicates practical significance.

**simple random sampling** When all member of a population have an equal and independent chance of being include in the random sample.

**single-blind experiment** An experiment in which subjects are unaware of the treatment condition they are in, although the researcher knows.

**single-subject designs** A type of experimental design for use with one subject or a group treated as one subject.

**single-subject research** An experiment with only one subject or intact group.

**situational differences** A threat to internal validity caused by extraneous variables.

**situational test** *See* **contrived observation**.

**skewed distribution** A distribution that is pulled out to one end by the presence of extreme scores.

**snowball sampling** A sampling procedure used in qualitative research that occurs when the initially selected subjects suggest the names of others who would be appropriate for the sample, the latter suggest others, and so on.

**social desirability bias** A bias that occurs in interviews when subjects give responses to enhance their image rather than honest responses.

**Solomon four-group design** An experimental design that involves random assignment of subjects to each of four groups; two groups are pretested, two are not; one of the pretested and one of the nonpretested groups receive the treatment; and all four groups are posttested.

**Solomon three-group design** An experimental design in which subjects are randomly assigned to three groups: (A) pretest plus treatment, (B) pretest plus no treatment, or (C) no pretest plus treatment.

**Spearman–Brown formula** A statistical procedure employing the correlation between split halves of a test to estimate the reliability of the whole test.

**Spearman rho** An index of correlation for use with ordinal data.

**split-half reliability** A type of internal-consistency reliability obtained by artificially splitting a test into two halves and correlating individuals' scores on the two half tests.

**spurious result** An apparent relationship between two variables that is not genuine but due to an extraneous variable influencing both.

**standard deviation** A measure of the extent to which individual scores deviate from the mean of the distribution; the square root of the variance; a measure of dispersion used with interval data.

**standard error of estimate** A measure of the expected difference between predicted scores and actual scores.

**standard error of measurement** An index of the amount of measurement error in test scores; theoretically, the standard deviation of the distribution of observed scores around an individual's true score.

**standard error of the mean** The standard deviation of sampling errors of the mean. It indicates how much the means of random samples drawn from a single population can be expected to differ through chance alone.

**standardized test** A test with specified content, prescribed directions for administering and scoring, norms, and reliability and validity information derived from administration to representative samples.

**standard score** A transformed score expressed in terms of the number of standard deviation units the score is

above or below the mean; standard scores have a mean and a standard deviation set to some arbitrary standard.

**stanine scores** Standardized scores developed to help interpret *z* scores, avoiding negative numbers and decimals.

**static group comparison** A preexperimental design that involves at least two nonequivalent groups; one receives a treatment and both are posttested.

**statistic** A characteristic of a sample.

**statistical conclusion validity** The validity of the inferences about the covariation between two variables in an experimental study.

**statistical equivalence** The condition occurring when any difference between groups is a function of chance alone.

**statistically significant result** A result that has less than a previously specified probability of being merely a function of chance.

**statistical regression** The tendency for subjects who scored extremely high or extremely low on a pretest to score closer to the mean on a posttest.

**stepwise replication** A process in which two investigators divide the data from a qualitative inquiry, analyze them independently, and then compare results.

**stratified purposeful sampling** A form of stratified sampling that ensures that all subgroups are represented.

**stratified sampling** A probability sampling technique that first divides a population into subgroups by relevant variables such as age, social status, or education and then randomly selects subjects from each subgroup.

**structural corroboration** The use of different sources of qualitative data and different methods to determine if there is agreement (corroboration) in the description and interpretation of a state of affairs.

**structural description** Illuminating the context that influenced the experience.

**structural equation modeling** A generic term for a group of advanced techniques used in the analysis of causality. It includes path analysis and confirmatory factor analysis.

**style manual** A book detailing form and style for theses, dissertations, and other formal written works.

**subject** A person in a study.

**subject descriptors** *See* **descriptors**.

**subject effects** Attitudes developed by subjects in response to the research situation.

**substantive categories** Categories that are primarily descriptive, not generally related to more abstract theories.

**summated rating scale** *See* **Likert scale**.

**survey** Study of a sample to investigate the incidence and distribution of variables.

**symmetrical distribution** A distribution that can be represented by a polygon whose right side is a mirror image of its left side.

**systematic approach to coding** When the constant-comparative method is the primary analysis technique.

**systematic errors of measurement** Errors that inflate or depress scores of identifiable groups in a predictable way.

**systematic sampling** A probability sampling in which every *k*th element of the population list is selected for the sample.

**tangibles** Items that can be directly assessed by observation.

**target population** The total group to which the researcher would like to generalize the results of a study.

***t* curves** The probability distribution for degrees of freedom.

**teacher inquiry** *See* **action research**.

**teacher-made test** *See* **researcher-made test**.

**teacher research** *See* **action research**.

**test** A systematic procedure for measuring a sample of behavior presumed to represent an educational or psychological characteristic.

**testable** A term applied to a theory or hypothesis that can be verified through empirical observation.

**test content-related evidence** *See* **content-related evidence**.

**testing effect** The problem that arises when taking a test once may affect the subject's performance when the test is taken again, regardless of any treatment.

**test of significance** A statistical test used to determine whether the obtained results are likely to be a function of chance.

**test–retest reliability** A measure of reliability obtained by correlating scores from two administrations of the same measure to the same subjects.

**textual description** A description illuminating what was observed.

**themes** The level of abstraction above categories in the coding process.

**theoretical adequacy** The degree to which a theoretical explanation developed from the study fits the data and is defensible.

**theoretical categories** More abstract categories that can be developed from prior theory or from inductively developed theory.

**theoretical coding** The stage in coding in which relationships among categories are posited.

**theoretical sampling** A qualitative sampling strategy that involves continuing to select new cases as the research unfolds, based on any emerging theories developed from the data.

**theoretical saturation** The process of testing the explanatory adequacy of theoretical constructs until it no longer contributes anything new.

**theory** A set of interrelated propositions or hypotheses that presents an explanation of some phenomenon.

**theory triangulation** Considering how the phenomenon under investigation might be explained by multiple theories.

**thick description** A term used in qualitative research that refers to descriptions that are written to create vivid images of the setting and behaviors being observed.

**time-series design** An experiment in which dependent variable measures are made several times before and after treatment.

**transcription** Recording in words all forms of data obtained in a qualitative study, with additional notes made as necessary.

**transferability** In qualitative research, the degree to which the findings of a study can be generalized to other contexts or to other groups.

**treatment group** The group in an experimental study that receives the treatment.

**treatment variable** *See* **independent variable**.

**trend study** A longitudinal survey in which data are gathered from different samples of a general population at different points in time to investigate changes over time.

**triangulation** Confirming data by using multiple data-gathering procedures, multiple sources of data or multiple observers.

**true experimental design** An experimental design in which the researcher controls the treatment and uses randomization to assign subjects to treatments.

**true score** The hypothetical error-free score for an individual.

**trustworthiness** *See* **dependability**.

***T* score** A standard score with a mean of 50 and a standard deviation of 10.

***t* test** A statistical procedure for testing hypotheses concerning the difference between two means; also used for other purposes.

**two-tailed test** *See* **nondirectional test**.

**Type I error** The error that occurs when a researcher rejects a null hypothesis that is in fact true.

**Type II error** The error that occurs when a researcher fails to reject a null hypothesis that is in fact false.

**typical case sampling** Selecting units that are considered typical in terms of the phenomenon of interest.

**unhypothesized relationships** Relationships appearing in the study that were not anticipated in the hypothesis.

**unique case sampling** *See* **extreme case sampling**.

**unit** A single occurrence, group, or individual that the researcher is interested in examining in a case study.

**universal determinism** The belief that all natural phenomena have antecedent factors.

**unobtrusiveness** A research setting characterized by the fact that the actions observed are not influenced by the presence of the researcher.

**using subjects as their own controls** A procedure used to avoid threats to internal validity by assigning all subjects to all experimental treatments.

**validity** The extent to which a measure actually taps the underlying concept that it purports to measure.

**validity coefficient** The coefficient of correlation between test scores and criterion.

**variability** The dispersion or spread in a distribution of scores.

**variable** A representation of a construct that takes on a range of values.

**variance** The mean of squared deviation scores; an interval measure of dispersion of scores around the mean.

**variance questions** Questions that ask to what extent or about differences.

**web-based surveys** Questionnaires placed on websites.

**Wilcoxon test** A statistical test for ranked data that serves the same purpose as the dependent *t* test.

**within-groups design** An experimental design in which the same subjects participate in all of the different treatment conditions.

***z* score** A standard score that indicates how far a score is above or below the mean score in terms of standard deviation units.

# Index

## C

## D



## E

## N

## O



## P

## Q

## R

## S