A revolution is taking place in education. Just by reading this book, you’re taking part in that revolution. Whether you’re an educator or studying to be an educator, whether you care deeply about research in education or hope this is the last time you’ll ever hear the word research, this book will change your way of thinking not only about your profession and your part in it but also about educational policy and practice.

The “revolution” is called evidence-based education, but you could also call it “Show me the data.” From the time the first cave mom homeschooled her kids, teachers have taught the best way they know how, learning from their own teachers,
from teacher educators, from wise sages, and most importantly from their students, observing the effects of their own teaching to learn what works and what doesn’t. These sources of knowledge and wisdom are crucial, of course, and they will always exist. However, they are no longer sufficient.

Evidence-Based Education

In education today, teachers and administrators must not only know their craft. In addition, they must know the evidence that supports the decisions they make, and they must be able to demonstrate that their students are learning. Educators at all levels live in an age of accountability. Accountability means having to show that students are learning. They’re doing better on tests, for example, or are graduating at higher rates from high school. Increasingly, accountability means something else as well. It means using programs and practices that have been shown to work in high-quality research. Just as physicians are expected to know and apply the findings of the latest research in medicine, so are educators being asked to know and apply the findings of research in education.

The evidence-based education movement is controversial, and it is taking form in fits and starts. You may or may not like it, but you must come to terms with it. If you are working in any field of education in the twenty-first century, evidence-based education will be a large part of your daily reality.

THE ROLE OF RESEARCH IN EVIDENCE-BASED EDUCATION

Clearly, research plays a central role in evidence-based education. In education, many kinds of research are used to answer all kinds of questions. In the age of accountability, all of these types of research are important, and it is becoming essential for every educator to understand the logic and the language of research. Whether you plan to become a researcher yourself; to do a master’s thesis or a doctoral dissertation; or to be a teacher, principal, central office administrator, or none of the above, you will need to understand how research works, what each type of research contributes to the field, and how to tell valid research from misleading research. In the age of accountability, you’ll need to be a sophisticated consumer of research, regardless of whether you are also a producer of research.
Many developments in education policy support the evidence-based education movement. One is accountability itself.

**Accountability** Today, schools throughout the United States are evaluated based on the achievement of their children. They are rewarded and punished based on test scores, especially in reading and math. Accountability is controversial; many educators are concerned that having an excessive focus on test-based accountability narrows the curriculum to whatever is tested, pushing out other subjects. These educators are concerned about tests focusing teachers on easily measured skills and knowledge and eliminating creativity. They worry that test-based accountability does not allow for teaching the “whole child.” Many worry about unfairness. Schools in more disadvantaged neighborhoods, with high student mobility and insufficient resources, have more trouble meeting state standards than well-funded schools in stable, affluent neighborhoods.

Despite these concerns, accountability is here to stay because the public demands it. No politician ever got elected by promising to lower educational standards or reduce accountability. In the age of accountability, educators need to be savvy about appropriate and inappropriate uses of evidence to indicate that a student, a class, a school, or a district is succeeding or failing, if only in self-defense.

**Effectiveness** A newer movement toward evidence-based education is seen in policies requiring educators in certain circumstances to use programs and practices with strong evidence of effectiveness. For example, in the 2001 No Child Left Behind (NCLB) Act, schools receiving federal funds were directed to use programs and practices “based on scientifically-based research” more than one hundred times. The U.S. Department of Education has established a process to review research in various areas to determine which programs have been validated in rigorous research and which have not. The What Works Clearinghouse (2004) is a key part of the evidence-based education movement, as it provides educators with well-justified evidence on programs that work. Within a few years, as many curricular areas are reviewed in this way, educators will be able to confidently choose programs and practices that have strong evidence of effectiveness, just as physicians can choose medications and procedures that have been rigorously reviewed by the Food and Drug Administration. Evidence-based education puts a new focus on research, not only research to evaluate outcomes of programs but also research on processes that lead to those outcomes.
An optimistic view of the role of research in this age of accountability appears in Figure 1.1. Accountability has made all educators more concerned about demonstrating student outcomes. With greater government investment in research, authoritative reviews of research such as the What Works Clearinghouse, and insistence by policymakers that educators use programs with strong evidence of effectiveness, both research and practice will be transformed to the benefit of children. None of this is certain, however, but there are important movements in this direction.

The potential of the evidence-based education movement is truly revolutionary. Education has long been beset by faddism. With any luck, this will soon be replaced by practice based on evidence. And you will be a part of the revolution. The more you know about how evidence is produced, evaluated, and appropriately applied, the better you will be able to operate as an intelligent, critical educator and to contribute to the dialogue. (For more on scientifically based research and practice, see Crawford & Impara, 2001; Mosteller & Boruch, 2002; Shavelson & Towne, 2002; Slavin, 2003.)
Research is organized, systematic inquiry that seeks to answer well-framed questions. One way to understand this is to contrast research and ordinary experience. Day in and day out, everyone tries to make sense of the world by noticing patterns and making generalizations. For example, you might notice that boys raise their hands more often than girls in your class and conclude that boys are more confident, more aggressive, or more knowledgeable than girls. You might notice that students are well behaved before grading but less so afterward, and you might decide that students are motivated by grades. Perhaps you recognize that your preschoolers love to tell stories, and you might decide that building storytelling into your day might help their language development.

In each of these cases, you might be right or wrong, but simply noticing what is going on and making generalizations is not research. Researchers might explore the very same questions, but they will use organized, systematic methods to do so. These methods are set up in advance in an open and public way.

For example, do boys really raise their hands more often than girls in class? Researchers might create an observation form and carefully count the number of hands raised by students in several classes. They might be surprised to find out that girls raise their hands just as often but are not noticed or called on as often by teachers. A series of studies by Sadker, Sadker, and Long (1997) found just this: that

No Child Left Behind and the accompanying state exams have put schools on notice that curriculum and teaching methods need to be proven effective for meeting standards.
teachers who thought they were being completely equal in fact often called on boys much more often than girls.

To explore why boys and girls might have different rates of hand raising, researchers would look at additional data. They might observe whether when called on, boys or girls are more likely to know the answers or analyze test scores to see if students of one or the other gender knows the material better. Researchers might develop a questionnaire or interview form and ask the students themselves why they do or don’t raise their hands in class. The point is, researchers do not rely on general impressions but set out in an organized and
systematic way to collect information that may confirm or disconfirm initial hunches.

One of the hallmarks of good research is **disconfirmability**. What this means is that in good research, the findings can surprise you. You might have an impression that such and such is true, but when you subject your hunch to organized, systematic inquiry, you may find out that you were wrong.

**THE BEST POSSIBLE ANSWER TO THE BEST POSSIBLE QUESTION**

When my daughter, Rebecca, was 2 years old, she could perform amazing feats of mathematics. If you asked her to add 0 and 3, she’d answer “three.” Fourteen minus 7 was 7. Eighty-eight minus 79 was 9. Asked the cube root of 125, she confidently answered 5!

Rebecca’s older brothers, who were 10 and 7, loved to show off their sister’s abilities to unsuspecting friends and relatives. Everyone was truly amazed, until one of the boys made a mistake and posed a question in the wrong order. “Rebecca,” one of them asked, “what is 3 plus 0?” “Zero!” she answered, exposing the trick. (She had just been repeating the last number she heard.)

The case of Rebecca’s mathematical “skills” illustrates one of the most important principles of research: It’s not only the answers you get but the questions you ask that determine the value of a study. The mathematically interesting process involved in the interaction between Rebecca and her brothers lay not in the answers she gave but in posing the questions such that the last digit was always the answer. In research in education, asking questions that are worth asking is perhaps the most important step. A well-framed question is one that is both worth asking and that can be answered in a convincing way using organized, systematic inquiry. Posing well-framed research questions is addressed later in this chapter and in Chapter 12.

**TYPES OF RESEARCH IN EDUCATION**

For the purpose of gaining a general introduction to research, we first look at most research as being subsumed under two large categories: quantitative and qualitative. In reality, however, studies fall somewhere along a continuum between these two types of research, all of which are discussed in more detail throughout the book.

**Quantitative Research** In **quantitative research**, researchers collect numerical data, or information, from individuals or groups and usually subject these data to statistical analyses to determine whether there are relationships among them. Quantitative research usually poses hypotheses that are either supported or disconfirmed by the data.

Within quantitative studies, research designs can be either experimental or nonexperimental. In experimental research, a researcher introduces one or more
independent variables, or treatments, and observes the effect on one or more dependent variables, or outcomes. A treatment is a systematic set of instructions or conditions applied to subjects in a study. For example, a researcher might study the effect of using reading groups in the fifth grade by randomly assigning some teachers to use reading groups and others to use whole-class instruction and then measuring the reading achievement of the students. What makes this an experiment is that the researcher assigned (perhaps by flipping a coin) some teachers to use reading groups and some to use whole-class instruction. The independent variable introduced by the researcher is the teaching method (reading groups versus whole-class instruction); the dependent variable (so called because its value may depend on the value of the independent variables) is reading achievement.

In nonexperimental quantitative research (see Chapter 5), the researcher usually observes relationships between two or more variables as they exist, without trying to change them. For example, the researcher could have attempted to answer the question about reading groups and achievement by locating one group of teachers who already use reading groups and one group who already use whole-class instruction and then measuring their students’ gains in achievement over a schoolyear. Alternatively, the researcher might try to determine whether there is a relationship between teachers’ years of experience and their attitudes toward the use of reading groups. In other quantitative nonexperimental research, the researcher simply seeks to describe a certain group in terms of one or more variables—for example, an opinion poll designed to discover what proportion of teachers favors ability grouping in middle school English.

Qualitative Research Qualitative research typically seeks to describe a given setting in its full richness and complexity or to explore reasons that a situation exists. Qualitative research usually begins without a formal hypothesis but produces one over time as events unfold (see Chapters 7 and 8).

Qualitative research seeks primarily to describe a situation as it is, without formal testing of a hypothesis, or statement of relationship between variables. This type of research makes little use of numbers, but rather focuses on “thick description” of social settings. For example, a qualitative researcher might observe teacher–student interactions in high, middle, and low reading groups over an extended period of time to get a sense of how life in these different settings is different for teachers and students. Qualitative research makes few claims regarding representativeness, concentrating instead on explaining social processes in great detail.

Other Types of Research Many studies combine qualitative and quantitative methods in a strategy called mixed-methods research. For example, a quantitative experiment might include observations of what teachers and students actually do and how they perceive their experiences. In fact, many methodologists argue that quantitative research should always include qualitative elements, so the research can provide a deeper understanding of the treatments being studied than numbers.
alone can provide (see, for example, Chelimsky & Shadish, 1997; Tashakkori & Teddlie, 2002).

**Action research** in education is undertaken by teachers and other educators in their own settings to solve real problems and improve real outcomes. Action research may use a quantitative, qualitative, or mixed-methods approach. Chapter 9 explores action research more fully.

---

**Research Design**

Hearing the words *research design* has been known to send chills down the spines of many graduate students in education. Research design is too often seen as a complex, arcane subject that only methodologists and statisticians can possibly understand, and so it is presumed that the experts’ prescriptions must be slavishly followed.

**THE LOGIC OF RESEARCH DESIGN**

The basic logic of research design is quite simple. As summarized in Figure 1.2, all research in education begins with a question worth asking. Most researchers then express a hypothesis, an expectation about what they will find, although many qualitative researchers develop and test their hypotheses over time. Data (whether quantitative or qualitative) are then collected. Quantitative researchers use statistics to test hypotheses, while qualitative researchers use description. Finally, researchers form conclusions and try to show that their explanations for the findings are supported by the data, ruling out alternative explanations.

As a quantitative example, let’s say you have reason to believe that two things are related. For example, you might think that asking students more questions increases their math performance, or that students who are well liked by their classmates are usually more tolerant toward children with special needs than students who are less well liked, or that the more experience teachers have, the better behaved their students will be. If you collect data on any of these pairs of **variables** (questions and math performance, popularity and attitudes toward children with special needs, teacher experience and student behavior), you can use statistics to try to find out whether you were correct in guessing that they are related. If your statistics say that the relationship does exist, your research design must enable you to say confidently that the relationship exists in fact and that your observation of it is not limited to the particular group of students or teachers from whom you collected the data. If your statistics do not show any relationship, you want to be as sure as you can (you can never be certain) that your failure to find a relationship is due to the fact that there really is no relationship, rather than problems in your design, measures, or other aspects of your study.
Similarly, in qualitative research, you want to describe what is going on in a systematic way, testing your ideas against the reality you observe. For example, if you want to know about the experiences of children with disabilities when they are integrated in regular classes, you might observe a few children over several weeks or months and then check your own observations against other independent data. The best research design is one that will add to knowledge, no matter what the results are.

**IMPORTANT ELEMENTS IN RESEARCH**

**Hypotheses** Most scientific investigations begin with a hypothesis, or formalized hunch about the relationship between two or more variables. A clearly stated hypothesis gives a fairly accurate idea of what we would have to do to provide evidence to confirm or disconfirm the hypothesis. Here are a few hypotheses:

1. Use of daily mental arithmetic drills will increase the mathematics performance of fifth-graders more than daily written drills.
2. Eighth-grade students who have brothers and/or sisters will be more popular among their peers than only children.

3. Teachers who belong to unions will be more highly rated by their supervisors than teachers who do not belong to unions.

4. Attending a Montessori school helps preschool children develop better language skills than attending a typical preschool program.

5. A program designed to improve middle school students’ classroom behavior, in which students are given points exchangeable for comic books, will improve their behavior.

The purpose of research design is to determine as unambiguously as possible whether hypotheses such as these are true. Good research design simply rules out the greatest possible number of alternative explanations for a particular outcome.

**The Gremlin** Imagine, if you will, that there is a nasty Gremlin whose sole purpose in life is to expose flaws in educational research. Your job in designing research is to anticipate the Gremlin’s objections and leave him with nothing to say.

Take the last hypothesis from the preceding list, concerning disruptive students. Let’s say we want to evaluate a method of giving points to disruptive students for every two minutes that they are in their seats working on assigned material. Students may turn in their points for comic books or other rewards at the end of each day. Does this program improve the behavior of these students?

Let’s start the program with a single student and observe his behavior. After a week, we find that he is on task (in his seat doing assigned work) 88 percent of the time. Because 88 percent of the time seems like a high figure, we conclude that the program worked. This scenario is represented by Experiment 1 in Figure 1.3 on page 12. But, asks the Gremlin with glee, how much of the time was the student on task before the program began? Since we don’t know, we can’t tell whether the program changed the student’s behavior.

Now let’s assume that we had measured the student’s behavior the day before the program began. He was on task 70 percent of the time on that day, and since he was on task 88 percent of the time during the treatment, we conclude that the program was effective (see Experiment 2 in Figure 1.3). The Gremlin is still not impressed. That one day might have been unusual, so we still can’t be sure that the program worked.
FIGURE 1.3
Ruling Out Alternative Explanations

Experiment 1

Experiment 2

Experiment 3

Experiment 4
As a third alternative, we might have observed the student for two weeks before the beginning of treatment to get a baseline, which represents the average level of behavior in the absence of treatment. During this time, the student was on task only 75 percent of the time. Since his time on task increased to 88 percent during the treatment, the program seems to be effective (see Experiment 3 in Figure 1.3). However, the Gremlin points out that it is still possible that other factors—such as changes at home, changes in other classes, or even the effect of being observed every day—made the student decide to be on task during the week of the treatment. The Gremlin gets that smug look on his face that he gets when he’s really got you.

To rule out this last possibility, we observe for one more week without offering points and rewards at the end of the week. The student’s on task behavior drops back to an average of 76 percent (Experiment 4 in Figure 1.3). Now we can have some confidence that it was the program that made the difference, at least for this one student. Since the changes in the student’s behavior correspond so directly to the introduction and withdrawal of the treatment, it is highly unlikely that factors other than the treatment could explain the changes.

The Gremlin’s smug look turns to panic. “It could have all happened by chance!” he proposes. “Look at the graph,” we reply. “The pattern of changes matched too closely with the beginning and end of the program.” “But . . .” the Gremlin sputters, “but . . .” Just to make him feel better, we admit to the Gremlin that to rule out the possibility that this is the only student for whom the program would work, we would have to replicate the same experiment with other disruptive students.

As you can see, each step in this process of building a research design ruled out one or more explanations for the student’s behavior other than the explanation in which we were interested: that the program made the difference. At each step, it seemed that the program was working, but the Gremlin or any other skeptic could easily have pointed to other explanations. However, the results of Experiment 4 (illustrated in Figure 1.3) made other explanations so unlikely as to be negligible. The Gremlin is nasty but he’s fair; he knows when he’s beaten. Experiment 4 is an example of an ABA, or reversal, design, which is described in detail in Chapter 4. However, similar logic would apply to other kinds of quantitative research, such as experimental (Chapter 2) and correlational (Chapter 5) designs; their purpose is to isolate particular factors and rule out alternative explanations.
Disproving the Null Hypothesis  The hypotheses listed earlier were stated in the form “A is related to B.” This is the hypothesis with which a quantitative researcher usually sets out. However, the logic of scientific method, which explores cause-and-effect relationships, actually demands that we begin each experiment with the hypothesis that “A is not related to B.” This is called the null hypothesis ($H_0$). The researcher’s task is to demonstrate beyond any reasonable doubt that the null hypothesis is incorrect. If we leave any significant possibility that the null hypothesis is correct, then we must continue to believe it. In the case of the experiment evaluating the point system for on-task behavior, Experiments 1, 2, and 3 left open the possibility that the null hypothesis (that the treatment does not affect student behavior) is true. Only Experiment 4 made the null hypothesis highly unlikely.

In any quantitative study, the researcher’s task is to build an argument ruling out explanations for any findings other than the explanation implied in the theory on which the hypothesis is based. Certain features of the research may buttress that argument considerably, but no matter how sound the design is, the researcher must always justify his or her explanation of what happened by providing overwhelming evidence that the null hypothesis is false. The null hypothesis can never be proved true. It is logically impossible to prove that A is not related to B, because it is always possible that there is a relationship that our methods failed to detect. However, with a strong research design, if the data indicate that A and B are not related, we can establish that it is unlikely that A and B are in fact related, or if they are, that the size of the relationship is very small.

Theory  A theory is essentially an explanation of how one or more variables are related to other variables. In the experiment described earlier, we might have had a theory that explained why we expected the treatment to influence the behavior of the disruptive student. That theory might have been as follows:

1. All organisms act to seek pleasure and avoid pain.
2. Students are organisms, so they seek pleasure and avoid pain.
3. Comic books give students pleasure.
4. If we make the acquisition of comic books dependent on a student’s on-task behavior, the student’s on-task behavior will increase, because increased on-task behavior will earn him or her comic books and thus pleasure.

Each of these statements is separately testable, and they are all very sensible; the first two (on which the others rest) are quite firmly established. Having a theory gives a study meaning beyond the particulars of the procedures and subjects used. Subjects are the individuals whose responses serve as the principal data in a study. In this study, we found out that if points exchangeable for comic books are given as a reward for on-task behavior, the student’s on-task behavior will increase. Our theory would imply that if candy bars give students pleasure, they might work just as
well as comic books, or that giving points based on smiling behavior might increase
smiling behavior, and so on. If we conduct similar experiments using different re-
wards and different behaviors and the results come out the same way, we might find
a general principle of behavior that has considerable explanatory power.

Science is a process of gradually refining theories and making them more gen-
eral while illuminating the conditions under which they may or may not apply. The
experiment described above might be one small step toward a theory that would
go like this:

For any organism, behaviors that are rewarded increase in frequency.

Statistical Significance  One key concept involved in disproving the null hy-
pothesis in quantitative studies is statistical significance, or using statistical analysis
to determine whether a given relationship between variables happened by chance.
To see how this works, consider the following example.

Let’s say a researcher wants to study the effects of a book club activity on the
number of books read at home by third-graders. She takes a group of 30 children
and randomly assigns them to a book club group or a non–book club group by put-
ting their names in pairs and then flipping a coin repeatedly to see which group
each child will be in. During a special period, the book club group goes with a
student teacher to discuss library books they have read at home. The remaining
students, the control group, are also encouraged to take books home but do not
have book club discussions.

Two possible sets of outcome data are depicted in Figure 1.4 on page 16. In the
figure, each X represents the number of books read by a single child. Note that in
Outcome A and Outcome B, the average number of books read by the book club
group is seven and by the non–book club group is five. In both cases, participating
in the book club appears to have led to children reading more books. But are seven
and five different enough to conclude that the different treatments had different
effects, or could this difference be due to random, meaningless variation? In other
words, do we have sufficient evidence to reject the null hypothesis?

In Outcome B, we cannot reject the null hypothesis. The number of books
read varies from zero to fourteen in one group and zero to twelve in the other; one-
third (five) of the children in the non–book club read more books than the average
in the book club group. There is no clear pattern, even though the means appear
to be different. However, in Outcome A, the book club group clearly read more
books than the control group. The two groups hardly overlap at all; only one child
from the control group has a score above the mean of the book club group. The
null hypothesis is highly unlikely in this case, so we can reject it.

Of course, Outcome A and Outcome B are extreme cases. Ordinarily, we can-
ot just look at a graph to see whether two groups differ. To make this comparison,
we would use statistics that essentially test whether the difference between the two
means is large compared to the amount of dispersion, or spread, in the scores. If
the amount of dispersion of scores around the mean is small (as in Outcome A), a
small difference between means can be judged to be statistically significant (that is, reliably different), while a much larger difference is needed when the dispersion is greater. The statistics themselves are discussed in Chapters 13 and 14, but the concept of statistical criteria for rejecting the null hypothesis is what is important here.
False Positive and False Negative Errors  In testing a hypothesis about the relationship between two or more variables, there are two ways to be wrong. One can be called false positive error (or Type I error). This type of error occurs when a relationship does not really exist, but your analysis claims that it does. The second, a false negative error (or Type II error), occurs when the relationship does exist, but your analysis fails to show that it does. This is illustrated in Figure 1.5.

To understand false positive and false negative errors, consider an experiment. Imagine that you and your friend get into an argument on a fishing trip. You say that if you drop a big rock into a small lake, the waves from the rock will peter out not far from the boat. Your friend says the waves will go all the way to the shore. You bet a pizza on the outcome and decide to test it out. You get a big rock and row out to the middle of a small lake on a calm day. Your friend stays on shore to watch the waves. “Drop!” you shout as you throw in the rock. A minute later, your friend sees the waves lapping the shore. “I was right!” he says.

Now imagine it was a windy day and the lake was choppy. “Drop!” you shout as you drop the rock. A minute passes, but your friend can’t see any difference in the waves. In fact, the rock had exactly the same effect as it did on a calm day, but it would take very sophisticated instruments to detect it—to separate the rock-caused waves from all the other waves. This would be a false negative error. The rock had an effect, but you failed to detect it.

Now imagine a third possibility. It’s still a windy day. You row the boat to the middle of the lake and shout “Drop!” Your friend, who is nearsighted, can’t see that you were fooling him—you didn’t drop the rock. A minute later, your friend thinks he sees the effect of the rock in the wavy water. “I win!” he says. He’s not lying; he is just perceiving something that’s not there in a confusing situation. (Besides, he does want that pizza.) Your friend’s seeing waves from the rock when you didn’t drop it would be an example of a false positive error, which means that there was in fact no effect but your research methods claimed there was. Figure 1.6 illustrates this experiment.
Correct: The rock caused waves.

False negative error: The rock caused waves, but our friend did not detect them.

Correct: No rock, no waves.

False positive error: No rock was dropped, but our friend thought he detected waves anyway.
In educational research, we live on “wavy lakes.” Children, teachers, schools, and school systems are very diverse, creating “turbulence.” Children have many influences on them, other than the treatment a researcher has in mind, so you have to drop a very big “rock” (i.e., a powerful treatment) or use sophisticated measures and analyses to detect the effects of treatments in educational research.

Statistics and other aspects of quantitative research design are directed primarily at minimizing the possibility of false positive error. That is, we want to be conservative about rejecting the null hypothesis. The risk of missing some true relationships is preferable to the risk of cluttering up our understanding of important variables with false relationships. We reject the null hypothesis only when the evidence against it is overwhelming. Statistical conventions generally demand that there be less than 1 chance in 20 (5 percent) that a difference between two means could have happened by chance. In some instances, researchers are not satisfied unless there is less than 1 chance in 100 (1 percent) that random variation could account for the findings. The more stringent we are in setting criteria for rejecting the null hypothesis, the more we reduce the possibility of getting a false positive error, but by doing so, we also increase the possibility of getting a false negative error (i.e., of missing a true relationship).

However, we can reduce the possibility of getting a false negative error without increasing the possibility of getting a false positive error. One means of doing this is to use large numbers of subjects. In the book club experiment, if there had been 150 children instead of 15 and if the pattern of scores had been the same, it may have been reliably clear in Outcome B that more books were read in the book club group than in the non–book club control group. Another means of reducing the possibility of getting a false negative error is to use more reliable measures (see Chapter 11). The wide dispersion depicted in Outcome B in Figure 1.4 might have been caused by difficulties in determining whether children actually read the books or just said they did. The “number of books read” measure would then be unreliable; that is, inaccuracies in determining the number of books read would create some degree of meaningless variation. The “books read” measure might have been made more reliable by giving the students quizzes on their books to determine if they had really read them, for example.

As another example, suppose a researcher hypothesizes that English language learners (ELLs) are more creative than English proficient students. To test this hypothesis, he gives a test of creativity to 10 ELLs from one school and to 10 English proficient students from another school. Let’s say he finds no statistically significant difference between the groups. Having failed to reject the null hypothesis, can the researcher safely conclude that there is in fact no difference between the two groups? Not at all. The small sample size (10 per group) makes statistical significance difficult to achieve. Inadequate measurement of creativity (which is difficult to define and measure) might also account for the findings. The chances are thus very high that a false negative error is responsible for the failure to find significant results.
What if the researcher had found a significant difference between the groups in the hypothesized direction? Since small sample size and unreliable measures work against finding statistical significance, they are less of a problem if significant findings are obtained. However, there are several potential sources of bias in this study. The main hypothesis concerned differences between ELLs and English proficient students. There are millions of students in the United States alone. The two groups of 10 chosen for this study (from only two schools) are certainly not representative of all such students. They may not even be comparable to each other. The ELL students might be drawn from better-educated families than the English proficient students. They might have better teachers or teachers who place greater emphasis on creativity. Any number of factors other than the students’ English proficiency might account for their higher scores. If the ELL students were significantly lower in creativity, the same problems would make this result questionable. In either case, if we decide to believe that the finding of a statistically significant difference in measured creativity between ELL and English proficient students is due to language proficiency, chances are good that we will be making a false positive error. The Gremlin would have no problem discrediting any conclusions from this study.

The researcher could have predicted that the chance of finding statistically significant differences between the groups was small (even if true differences did exist) and that even if he found statistically significant differences, their origin would be unclear. He should have found a different way to answer the question he posed.

**Internal Validity**  
Internal validity refers to the degree to which a study rules out any explanation for the study’s findings other than the one claimed by the researcher. If a researcher wants to compare Treatment A with Treatment B, she wants to be sure that if her study shows Treatment A to be superior to Treatment B, the reason for this will be that Treatment A really is better than Treatment B. In other words, she wants to be sure that any difference observed is due to a true difference between the treatments, not to defects in the study. The same logic applies to a finding that Treatments A and B do not differ. In a study high in internal validity, we can be relatively confident that if no differences are found between the two treatments, then none exist. We can never be sure that a failure to find statistically significant differences means that two treatments do not differ, but in a well-designed study high in internal validity, we can make that argument more confidently than in a poorly designed study. In the fictitious example mentioned earlier, we learned little or nothing about ELLs and creativity because the study was too low in internal validity; any number of factors other than a true difference between the students could have explained the results.

Determining whether a study is high in internal validity is largely a matter of common sense, aided by experience gained through reading and interpreting.
research. For example, let’s say a researcher measures students’ attitudes toward school in January and then introduces a program in which students receive colorful stickers for doing well on their classwork. In May, the researcher gives the students the same attitude scale again and finds that students’ attitudes have significantly improved. Is the sticker program effective? We have no idea, because we don’t know what would have happened to students’ attitudes without the stickers. Perhaps students are happier in May. Perhaps something else changed to improve their attitudes. Perhaps students wanted to help their teacher look good by responding positively on the attitude scale. On the other hand, it is possible that attitudes did improve because of the sticker program.

Regardless, if the research cannot rule out all of the alternative “perhapses” beyond a reasonable doubt, then the study has little internal validity and therefore little informative value. To put it another way, we have to convince a very skeptical Gremlin. The burden of proof is always on the researcher to argue that his or her explanation for a study’s results is the only explanation that has any real chance of being true. Common threats to internal validity in educational research are discussed in Chapter 10.

**External Validity**  In social science research, it is never enough to say that for a particular set of subjects, Variable A and Variable B are related or that Group A is different from Group B. The next question is, So what? If, for example, we find that in Ms. Jackson’s twelfth-grade physics class, students’ heights are correlated with their physics grades, we will know very little. Will that relationship hold up in other physics classes? For other subjects? At other grade levels? Are the findings due to that fact that twelfth-grade boys tend to be taller than girls, that Ms. Jackson favors boys in her grading, or to some other explanation?

As researchers, we have very little interest in Ms. Jackson’s physics class unless we can learn something in it that has meaning for some larger set of individuals, such as all physics students, all science students, all twelfth-graders, or all high school students. The degree to which the results of a study can be generalized or applied to a population in which we are interested is called external validity or generalizability.

External validity is not as cut-and-dried as internal validity, because we never know for certain that a finding has external validity until we assess it on the entire population in which we are interested. If the population we wish to understand is that of a small Pacific island, we might be able to study the entire population. However, even when anthropologists study such a population, they are really trying to learn principles that will have some application to understanding other Pacific Islanders or even to understanding human society in general. In research in schools, we almost always intend our research to have meaning for a much larger population and under a much wider range of conditions than the particular sample and conditions we study. For this reason, external validity is a major concern. Threats to external validity that are common in education research are discussed in Chapter 10.
Essentials of Research Design

Research design is really quite straightforward, since it is based on logic and common sense. The critical skill in research design is deciding on a question that is important and then choosing research methods that will answer that question as unambiguously as possible, given limited resources. Letting research methods determine our question or following research design formulas instead of thinking through what we are trying to learn diminishes the usefulness of research in informing us about the issues we want to understand. If a researcher can, in all honesty, answer the following questions in the affirmative, then he or she knows all that is necessary about research design:

1. Is the problem I am planning to study an important one?
2. Do I have a sensible theory that links the variables I plan to study?
3. If the data confirm my hypothesis or expectation, can I be confident (a) that the relationship I hypothesized does in fact exist, (b) that it exists for the reason I say it exists, and (c) that the finding has meaning beyond the particular group I studied?
4. If the data fail to confirm my hypothesis, can I be confident that the relationship I hypothesized does not in fact exist?
5. Is the study feasible, given my resources?

All too often, researchers in education latch on to a particular methodology and stick with it throughout their careers. This means that they must either ask a limited set of questions that lend themselves to that methodology or bend questions that really require different research methods to fit the methods they know (as in the old saying that “To a person who only has a hammer, all problems look like nails”). The purpose of this book is to provide future researchers and readers of research with a conceptual understanding and practical guide to many different types of research, so that they can feel comfortable with many different approaches to answering questions. The intention is to allow both producers and consumers of research to free themselves from a narrow view of how research must be done and help them to focus on the quality of the question first and only then to look for appropriate methods to search for correct and useful answers.

The task of the researcher is to get the best possible answer to the best possible question. The best possible answer may not always involve a perfect research design. It is often better to compromise on certain aspects of the research design than to ask a less interesting question that can be answered with a more perfect design. There is no formula for good research; procedures that are appropriate for one question might be fatally flawed for another. Ideally, the researcher uses a variety of methods to provide a deeper understanding of the topic. This book presents many ways of doing research and discusses the conditions under which each might be appropriate to particular questions. The following chapters are guides to answering these questions.
### RESEARCH NAVIGATOR

**Key Terms**

<table>
<thead>
<tr>
<th>Term</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>action research</td>
<td>9</td>
</tr>
<tr>
<td>baseline</td>
<td>13</td>
</tr>
<tr>
<td>bias</td>
<td>20</td>
</tr>
<tr>
<td>data</td>
<td>7</td>
</tr>
<tr>
<td>dependent variable</td>
<td>8</td>
</tr>
<tr>
<td>disconfirmability</td>
<td>7</td>
</tr>
<tr>
<td>evidence-based education</td>
<td>1</td>
</tr>
<tr>
<td>external validity</td>
<td>21</td>
</tr>
<tr>
<td>false negative error/type II error</td>
<td>17</td>
</tr>
<tr>
<td>false positive error/type I error</td>
<td>17</td>
</tr>
<tr>
<td>generalizability</td>
<td>21</td>
</tr>
<tr>
<td>hypothesis</td>
<td>10</td>
</tr>
<tr>
<td>independent variable</td>
<td>8</td>
</tr>
<tr>
<td>internal validity</td>
<td>20</td>
</tr>
<tr>
<td>mixed-methods research</td>
<td>8</td>
</tr>
<tr>
<td>null hypothesis ($H_0$)</td>
<td>14</td>
</tr>
<tr>
<td>qualitative research</td>
<td>8</td>
</tr>
<tr>
<td>quantitative research</td>
<td>7</td>
</tr>
<tr>
<td>reliable measures</td>
<td>19</td>
</tr>
<tr>
<td>research</td>
<td>5</td>
</tr>
<tr>
<td>research design</td>
<td>9</td>
</tr>
<tr>
<td>scientific method</td>
<td>14</td>
</tr>
<tr>
<td>statistical significance</td>
<td>15</td>
</tr>
<tr>
<td>study</td>
<td>7</td>
</tr>
<tr>
<td>subjects</td>
<td>14</td>
</tr>
<tr>
<td>theory</td>
<td>14</td>
</tr>
<tr>
<td>treatment</td>
<td>14</td>
</tr>
<tr>
<td>variables</td>
<td>9</td>
</tr>
</tbody>
</table>

**Activity**

If you have access to Research Navigator, located on the MyLabSchool website (www.mylabschool.com), enter the following keywords to find articles related to the content of this chapter.

- Accountability in education
- Evidence-based education
- What Works Clearinghouse
- Research design

### EXERCISES

A researcher conducted a study to determine whether having football games improved students’ attachment to the school. He stopped a group of 10 students in the hall and gave each a questionnaire involving one question about school spirit one day after a football game. After the football season was over, he stopped another group of 10 students and gave each the same questionnaire.
1. State the null hypothesis of this study.

2. How does the sample size affect the chances of getting a false positive or false negative error?

3. How does the method of choosing students for the study affect the chances of getting a false positive or false negative error?

4. How does the structure of the questionnaire chosen affect the chances of getting a false positive or false negative error?

5. Suppose the researcher in this study found a statistically significant difference between the two groups of students who were given the questionnaire after football season. The researcher concludes that having football games improves school spirit. What alternative explanations for this can you supply?

6. Suppose the researcher found no differences between the groups who had been given the questionnaire. What can you conclude from the study?

7. Be the Gremlin. Comment on the internal and external validity of this study.

8. Be the Gremlin again. Critique Demetra Powell’s study, described in the Research with Class box (see page 6). Is it research? What is her question? Does her study answer the question conclusively? Why or why not?

FURTHER READING

Learn more about the concepts discussed in this chapter by reviewing some of the research cited.

**Scientifically Based Research and Practice**


