CHAPTER 6

Causation and Research Design

Research Question: *How Do Educational Strategies Affect Educational Outcomes?*

Chapter Contents

- Causal Explanation
- Criteria for Causal Explanations
- Types of Research Designs
- True Experimental Designs
- Quasi-Experimental Designs
- Threats to Validity in Experimental Designs
- Nonexperiments

Identifying causes—figuring out why things happen—is the goal of much educational research. Unfortunately, valid explanations of the causes of educational phenomena do not come easily. The importance of early childhood learning is widely accepted. But which school strategies are best for young children, especially the disadvantaged? A connection between poverty and delinquency is well established, and both are linked to low achievement in school. But how exactly does poverty cause delinquency and low achievement, and what can schools do about it? Can early childhood programs such as Head Start have positive effects years later, helping their former students to avoid delinquency and get better grades? Causal questions such as these have stimulated much research.

In this chapter, our goal is to use studies on these and other, related questions to illustrate the ways educational researchers explore questions about causation through careful use of appropriate research methods.

We give special attention to key distinctions in research design that are related to our ability to come to causal conclusions: the criteria for causal explanations, reliance on a cross-sectional or longitudinal design, a focus on individual or group units of analysis, and the use of an experimental or nonexperimental design.

By the end of the chapter, you should have a good grasp of the different meanings of causation and be able to ask the right questions to determine whether causal inferences are likely to be valid, as well as a fuller understanding of research design. You may also have a better idea about the causes of success for early childhood students.

2 Causal Explanation

A cause is an explanation of some characteristic, attitude, or behavior of groups, individuals, other entities (families, organizations), or events. For example, Jeremy Finn and Charles Achilles (1990) conducted a statewide experiment to determine whether smaller class sizes led to long-term improved academic performance in the early grades. They concluded that it did, particularly for minority students. In Tennessee's Student Teacher Achievement Ratio experiment (Project STAR) involving 11,600 kindergarten through third-grade students in 80 elementary schools, they identified a causal effect of smaller class size on improved school performance. Although the original experiment took place more than 20 years ago, it is still the only large-scale, randomized trial study ever undertaken on class size effects. For this reason, many later researchers have used the Tennessee STAR experiment data to explore further causal connections about class size and learning (Schanzenbach, 2006).

More specifically, a causal effect is said to occur if variation in the hypothesized independent variable is followed by variation in the dependent variable, when all other things are equal (**ceteris paribus**). For instance, we know that for the most part, children who participate in early childhood education programs do better in school than children who do not participate in such programs, but this in itself does not establish that early childhood programs improve school performance. It could be that the parents who enroll their children in early childhood education programs also provided their children with more books and educational games before they started early childhood education. Maybe that is the real explanation for their better school performance. Or maybe neighborhoods that have more early childhood programs also have better schools. We just don't know. What we need to figure out is whether children who participate in early childhood programs do better after they enter school than other children, ceteris paribus—when all other things are equal.

We admit that you can legitimately argue that "all" other things can't literally be equal: We can't compare the same people at the same time in the same circumstances except for the variation in the independent variable (King et al., 1994). However, you will see that we can design research to create conditions that are very comparable so that we can isolate the impact of the independent variable on the dependent variable.

2 Criteria for Causal Explanations

Five criteria should be considered in trying to establish a causal relationship. The first three criteria are generally considered requirements for identifying a causal effect: (1) empirical association, (2) appropriate time order, and (3) nonspuriousness. You must establish these three to claim a causal relationship. Evidence that meets the other two criteria—(4) identifying a causal mechanism and (5) specifying the context in which the effect occurs—can considerably strengthen causal explanations.

Research designs that allow us to establish these criteria require careful planning, implementation, and analysis. Many times, researchers have to leave one or more of the criteria unmet and are left with some important doubts about the validity of their causal conclusions, or they may even avoid making any causal assertions.

Association

The first criterion for establishing a causal effect is an empirical (or observed) **association** (sometimes called a *correlation*) between the independent and dependent variables. The variables must vary together such that when one goes up (or down), the other goes up (or down) at the same time. Here are some examples: The longer you stay in school, the more money you will make in life. When income goes up, so does overall health. In the Tennessee STAR Program experiment (Finn & Achilles, 1990; Schanzenbach, 2006), when class size went down, academic performance went up. In all of these cases, a change in an independent variable correlates, or is associated with, a change in a dependent variable. If there is no association, there cannot be a causal relationship. For instance, empirically there seems to be no correlation between the use of the death penalty and a reduction in the rate of serious crime. That may seem unlikely to some people, but empirically it is the case. If there is no correlation, there cannot be a causal relationship.

Time Order

Association is a necessary criterion for establishing a causal effect, but it is not sufficient. We must also ensure that the variation in the dependent variable occurred after the variation in the independent variable—the effect must come after its presumed cause. This is the criterion of **time order,** or the temporal priority of the independent variable. Motivational speakers sometimes say that to achieve success (the dependent variable), you really need to believe in yourself (the independent variable). And it is true that many very successful people seem remarkably confident—there is an association. But it may well be that their confidence is the result of their success, not its cause. Until you know which came first, you can't establish a causal connection.

Nonspuriousness

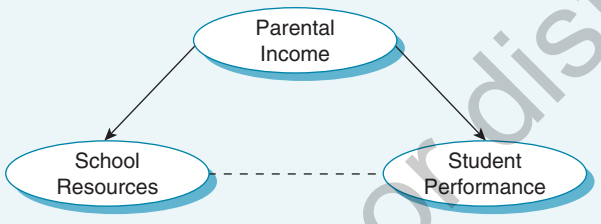
The third criterion for establishing a causal effect is **nonspuriousness.** *Spurious* means false or not genuine. We say that a relationship between two variables is **spurious** when it is due to changes in a third variable. Have you heard the old adage "Correlation does not prove causation"? It is meant to remind us that an association between two variables might be caused by something other than an effect of the presumed independent variable on the dependent variable. If we measure children's shoe sizes and their academic knowledge, for example, we will find a positive association. However, the association results from the fact that older children have larger feet as well as more academic knowledge. A third variable (age) is affecting both shoe size and knowledge so that they correlate, but one doesn't cause the other. Shoe size does not cause knowledge, or vice versa. The association between the two is, we say, spurious.

If you think this point is obvious, consider an education example. Do schools with more resources produce better students? Before you answer the question, consider the fact that parents with more education and higher income tend to live in neighborhoods that spend more on their schools. These parents also are more likely to have books in the home and provide other advantages for their children (see Exhibit 6.1). Do the parents cause variation in both school resources and student performance? If so, there would be an association between school resources and student performance that was at least partially spurious.

Exhibit 6.1 **A Spurious Relationship Revealed**

School resources are associated with student performance; apparently, a causal relation.

But in fact, parental income (a third variable) influences both school resources and student performance, creating the association.



Source: Chambliss and Schutt (2010, p. 134).

Mechanism

A causal **mechanism** is the process that creates the connection between the variation in an independent variable and the variation in the dependent variable it is hypothesized to cause (Cook & Campbell, 1979, p. 35; Marini & Singer, 1988). Many researchers argue that no causal explanation is adequate until a causal mechanism is identified (Costner, 1989).

For instance, there seems to be an empirical association at the individual level between poverty and delinquency: Children who live in impoverished homes seem more likely to be involved in petty crime. But why? Researchers, including Agnew, Matthews, Bucher, Welcher, and Keyes (2008) and Sampson and Laub (1994), have found that children who grew up with such structural disadvantages as family poverty and geographic mobility were more likely to become juvenile delinquents. Their analysis indicates that multiple economic problems and structural disadvantages lead to less parent-child attachment, less maternal supervision, and more erratic or harsh discipline. In this way, figuring out some aspects of the process by which the independent variable influenced the variation in the dependent variable—the causal mechanism—can increase confidence in our conclusion that a causal effect was at work (Costner, 1989).

Context

No cause has its effect apart from some larger **context** involving other variables. When, for whom, and in what conditions does this effect occur? A cause is really one among a set of interrelated factors required for the effect (Hage & Meeker, 1988; Papineau, 1978). Do the causal processes in which we are interested vary across families? Among school systems? Over time? For different types of students and teachers? Identification of the context in which a causal relationship occurs is not itself a criterion for a valid causal relationship, but it can help us to understand the causal relationship.

Awareness of contextual differences helps us to make sense of the discrepant findings from local studies. Always remember that the particular cause on which we focus in a given research design may be only one among a set of interrelated factors required for the effect; when we take context into account, we specify these other factors (Hage & Meeker, 1988; Papineau, 1978).

2 Types of Research Designs

Researchers usually start with a question, although some begin with a theory or a strategy. If you're very systematic, the *question* is related to the *theory,* and an appropriate *strategy* is chosen for the research. All of these, you will notice, are critical defining issues for the researcher. If your research question is trivial (How many shoes are in my closet?), or your theory sloppy (More shoes reflect better fashion sense.), or your strategy inappropriate (I'll look at lots of shoes and see what I learn.), the project is doomed from the start.

But let's say you've settled these first three elements of a sound research study. Now we must begin a more technical phase of the research: the design of the study. From this point on, we will be introducing a number of terms and definitions that may seem strange or difficult. In every case, though, these terms will help you clarify your thinking. Like precisely the right word in an essay, these technical terms help, or even require, researchers to be absolutely clear about what they are thinking—and to be precise in describing their work to other people.

An overall research strategy can be implemented through several different types of research design. One important distinction between research designs is whether data are collected at one point in time—a **cross-sectional research design**—or at two or more points in time—a **longitudinal research design.** Another important distinction is between research designs that focus on individuals—the individual unit of analysis—and those that focus on groups, or aggregates of individuals—the group unit of analysis.

Cross-sectional research design: A study in which data are collected at only one point in time.

Longitudinal research design: A study in which data are collected that can be ordered in time; also defined as research in which data are collected at two or more points in time.

Individual unit of analysis: A unit of analysis in which individuals are the source of data and the focus of conclusions.

Group unit of analysis: A unit of analysis in which groups are the source of data and the focus of the conclusions.

Cross-Sectional Designs

In a cross-sectional design, all of the data are collected at one point in time. In effect, you take a "cross section"—a slice that cuts across the entire population under study—and use that to see all the different parts, or sections, of that population. Much of the research you have encountered so far in this text—the studies of maternal employment in Chapter 1 and of the academic effects of poverty in Chapter 4—has been cross-sectional. Although each of these studies took some time to carry out, they measured the actions, attitudes, and characteristics of respondents at only one time.

But cross-sectional studies, because they use data collected at only one time, suffer from a serious weakness: They don't take into account the time order of effects. For instance, you may see statistics showing that a certain high school has a very good college sending rate for seniors. You might conclude, then, that seniors' academic success is because of what transpired over time—that is, what they learned while in the school. But in fact, it may be that the school's policies resulted in less academically successful students leaving the school before reaching their senior year, through disciplinary expulsion, being "counseled out," or other reasons. A cross-sectional study of seniors doesn't distinguish if they are succeeding because of

the instructional quality of the school or because, for whatever reason, those students least likely to graduate have already left the school before senior year begins. With a cross-sectional study, we can't be sure which explanation is correct, and that's a big weakness. To study change over time, we need a longitudinal design.

Longitudinal Designs

In longitudinal research, data are collected that can be ordered in time. By measuring the value of cases on an independent variable and a dependent variable at different times, the researcher can determine whether change in the independent variable precedes change in the dependent variable. In a cross-sectional study, when the data are collected all at one time, you can't really show if the hypothesized cause occurs first; in longitudinal studies, though, you can see if a cause occurs and then, later in time, the effect occurs. So if possible to do, longitudinal research is always preferable.

But collecting data two or more times takes time and work. Often researchers simply cannot, or are unwilling to, delay completion of a study for even 1 year to collect follow-up data. But think of the many research questions that really should involve a much longer follow-up period: What is the impact of elementary grade education on high school graduation? How effective is a high school parenting program in improving parenting skills when the students become adults? Under what conditions do traumatic experiences in early childhood result in a special-needs diagnosis in elementary school? It is safe to say that we will not be able to answer many important research questions because there was not enough time for a sufficiently long follow-up period. Nonetheless, the value of longitudinal data is so great that every effort should be made to develop longitudinal research designs of appropriate length when they are required for the research question.

In education, one technique for performing longitudinal studies is to tap into the immense amount of data routinely collected by governmental units such as public school systems and state and federal departments of education. This was the strategy Kathleen J. Skinner (2009) used to study charter schools in Boston, basing her research on longitudinal data from the Massachusetts Department of Education. She was interested in a variation of the question raised in the previous section: How should we view the success of charter schools that claim high rates of academic success for their graduating seniors? She found that although students were initially accepted to charter schools through a lottery system, once they enrolled, there was "significant student attrition resulting from the use of 'pushout' strategies based on student academic and/or behavioral performance" (Skinner, 2009, p. 1). Skinner tracked the number of students enrolled in each Boston charter school year by year to determine what percentage of entering students actually reached Grade 12. Exhibit 6.2 shows the figures from 2004–2009 for students who entered a charter school that claimed a 99% college acceptance rate for its graduates (Skinner, 2009, p. 30).

The final column in Exhibit 6.2 uses a metric called "promoting power" (Balfanz & Legters, 2004), which is simply the number of students in Grade 12 in a given year divided by the number of students who were in Grade 9 four years earlier. For example, for the senior class of 2009, promoting power is computed as 34/72 or 47%—of students who entered in 2005, only 47% made it to senior year. Taking the "promoting power" variable into account, Skinner's longitudinal study reveals a much lower success rate than the 99% graduation figure based on cross-sectional studies of the 12th graders.

Whether you plan to collect the data yourself or use an already existing data source, the following discussion of the three major types of longitudinal designs will give you a sense of the possibilities. (The three types are illustrated in Exhibit 6.3.)

Trend study (repeated cross-sectional design): A type of longitudinal study in which data are collected at two or more points in time from different samples of the same population.

Trend Studies

Studies that use a **repeated cross-sectional design,** also known as **trend studies,** are conducted as follows:

1. A sample is drawn from a population at Time 1, and data are collected from the sample.
2. As time passes, some people leave the population and others enter it.
3. At Time 2, a different sample is drawn from this population.

Exhibit 6.2 Example Charter School, Student Attrition 2000-2009

Graduating Class	Entry Year	Grade				Promoting Power (%)
		9	10	11	12	
2004	2000	78	54	32	27	35
2005	2001	65	50	38	28	43
2006	2002	79	56	24	18	23
2007	2003	49	38	25	20	41
2008	2004	96	72	54	46	48
2009	2005	72	61	46	34	47
Average, 2004–2009		73	55	36.5	29	40

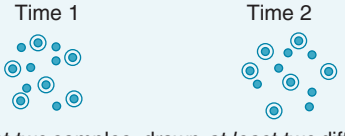
Source: Adapted from Skinner (2009, p. 31). Statistics from Massachusetts Department of Elementary and Secondary Education.

Exhibit 6.3 Three Types of Research Design

1. Cross-Sectional Design

One sample drawn at *one* time (not longitudinal).

2. Trend (or "Repeated Cross-Sectional") Design



At least two samples, drawn *at least two* different times (longitudinal).

3. Panel Design

One sample, measured *at least two* different times (longitudinal).

Source: Chambliss and Schutt (2010, p. 33).

The Gallup polls, begun in the 1930s, are a well-known example of trend studies. One Gallup poll, for instance, asks people what they think is the best way to improve kindergarten through 12th-grade education in the United States. Exhibit 6.4 shows how a 1,010-person sample of American adults answered this question in 2004 and again 5 years later in 2009. The top four items remained the same from 2004 to 2009, but their order and percentage ranking changed slightly.

Exhibit 6.4 **Best Way to Improve Education 2004 and 2009: The Gallup Organization**

Best Way to Improve Kindergarten Through 12th-Grade Education	Mentioning (%)	
	2004	2009
Quality teachers	15	17
Smaller class size	11	6
Basic curriculum	10	10
Improve funding	7	6
More parental involvement	6	5
Better teacher pay	6	6
Better discipline in schools	5	—
Hire more teachers	5	—
Teach about real life	—	5

Source: Gallup education poll, accessed at http://www.gallup.com/poll/1612/education.aspx

Each time the Gallup organization samples, it asks a different, though roughly demographically equivalent, group of people the same question; it isn't talking to the same people every time. Then it uses the results of a series of these questions to analyze change in Americans' opinions about education. This is a trend study.

These features make the trend study (repeated cross-sectional) design appropriate when the goal is to determine whether a population has changed over time. Has racial tolerance increased among Americans in the past 20 years? Are employers more likely to pay maternity benefits today than they were in the 1950s? These questions concern changes in the population as a whole, not changes in individuals.

Fixed-sample panel design (panel study): A type of longitudinal study in which data are collected from the same individuals—the panel—at two or more points in time. In another type of panel design, panel members who leave are replaced with new members.

Panel Designs

When we need to know whether individuals in the population have changed, we must turn to a panel design. Panel designs allow us to identify changes in individuals, groups, or whatever we are studying. This is the process for conducting **fixed-sample panel studies:**

1. A sample (called a panel) is drawn from a population at Time 1, and data are collected from the sample.
2. As time passes, some panel members become unavailable for follow-up, and the population changes.
3. At Time 2, data are collected from the same people as at Time 1 (the panel)—except for those people who cannot be located.

Because a panel design follows the same individuals, it is better than a repeated cross-sectional design for testing causal hypotheses. For example, Eliana Garces, Duncan Thomas, and Janet Currie (2002) used a panel design to study the long-term effects of the federal Head Start program. The panel, sponsored by the National Science Foundation, was a representative sample of U.S. families participating in Head Start; data were collected from the same families and their descendents 34 times between 1968 and 2005 (National Science Foundation, 2005). The researchers found that White Head Start students, in comparison to siblings who were not in the program, were "significantly more likely to complete high school, attend college, and possibly have higher earnings in their early twenties" (Garces et al., 2002, p. 999). They also found positive social and academic benefits for African American program participants (p. 999).

A panel design allows us to determine how individuals change, as well as how the population as a whole has changed; this is a great advantage. However, panel designs are difficult to implement successfully and often are not even attempted because of two major difficulties:

Expense and attrition. It can be difficult and expensive to keep track of individuals over a long period, and inevitably the proportion of panel members who can be located for follow-up will decline over time. Panel studies often lose more than one quarter of their members through attrition (D. C. Miller, 1991, p. 70).

Subject fatigue. Panel members may grow weary of repeated interviews and drop out of the study, or they may become so used to answering the standard questions in the survey that they start giving stock answers rather than actually thinking about their current feelings or actions. This is called the problem of **subject fatigue**.

Because panel studies are so useful, researchers have developed increasingly effective techniques for keeping track of individuals and overcoming subject fatigue. But when resources do not permit use of these techniques to maintain an adequate panel, repeated cross-sectional designs usually can be employed at a cost that is not a great deal higher than that of a one-time-only cross-sectional study. The payoff in explanatory power should be well worth the cost.

Cohort Designs

Trend and panel studies can track both the results of an event (such as the Vietnam War) and the progress of a specific historical generation (e.g., people born in 1985). In this case, the historically specific group of people being studied is known as a **cohort,** and this cohort makes up the basic population for your trend or panel study. Such a study has a **cohort design** (also called an **event-based design**). If you were doing a trend study, the cohort would be the population from which you drew your different samples. If you were doing a panel study, the cohort provides the population from which the panel itself is drawn. Examples include the following:

- *Birth cohorts*—those who share a common period of birth (those born in the 1940s, 1950s, 1960s, etc.)
- *Seniority cohorts*—those who have worked at the same place for about 5 years, about 10 years, and so on
- *School cohorts*—freshmen, sophomores, juniors, seniors

Cohort: Individuals or groups with a common starting point. Examples include individuals who began kindergarten in 1997, the college class of 2009, people who graduated from high school in the 1980s, and teachers who began teaching in 2005. Cohorts can form the initial population for either trend or panel studies.

Cohort design (event-based design): A type of longitudinal study in which data are collected at two or more points in time from individuals in a cohort.

We can see the value of event-based research in a comparison of two studies that estimated the impact of public and private schooling on high school students' achievement test scores. In an initial cross-sectional (not longitudinal) study, James Coleman, Thomas Hoffer, and Sally Kilgore (1982) compared standardized achievement test scores of high school sophomores and seniors in public, Catholic, and other private schools. They found that test scores were higher in the private high schools (both Catholic and other) than in the public high schools.

But was this difference a causal effect of private schooling? Perhaps the parents of higher-performing children were choosing to send them to private schools rather than to public ones.

So James Coleman and Thomas Hoffer (1987) went back to the high schools and studied the test scores of the former sophomores 2 years later, when they were seniors; in other words, the researchers used an event-based panel (longitudinal) design. This time they found that the verbal and math achievement test scores of the Catholic school students had increased more over the 2 years than the scores of the public school students had. Irrespective of students' initial achievement test scores, the Catholic schools seemed to "do more" for their students than did the public schools. The researchers' causal conclusion rested on much stronger ground because they used a cohort design.

Units and Levels of Analysis

Individual and Group Units of Analysis

As a student of educational research, you probably understand by now that groups don't act or think like individuals do. Groups and individuals are different units of analysis. **Units of analysis** are the things that you are studying, whose behavior you want to understand. Often, these are individual people, but they can also be, for instance, classrooms, schools, school systems, or the educational population of whole states. All of these could be units of analysis for your research.

Research on compulsory high-stakes testing, for instance, often uses the individual student as the unit of analysis. The researcher may collect survey data on individual test scores, then analyze the data, and then report on, say, how many individuals passed and how many failed.

Alternatively, units of analysis may instead be groups of some sort, such as grade levels, schools, or school systems. A researcher may analyze testing data published in the newspaper or on the website of the state department of education and find out what percentage of fifth graders passed in each elementary school in town or what percentage of all students in town passed from all grade levels. The researcher can then analyze the relationship between how long students have been in the school system and what happens to their scores. Does the percentage of students reaching competence go up or down the longer they are in the school system? Are math scores stronger than language, or vice versa? Because the data describe the city or town's school system, cities or towns are the units of analysis. In this example, either groups or individuals can be the units of analysis because data are collected from individuals (individual test scores), but taken together, the individual test scores create a profile of achievement in the town.

We also have to know what the units of analysis are to interpret statistics appropriately. Measures of association tend to be stronger for group-level than for individual-level data because measurement errors at the individual level tend to cancel out at the group level (Bridges & Weis, 1989, pp. 29–31).

The Ecological Fallacy and Reductionism

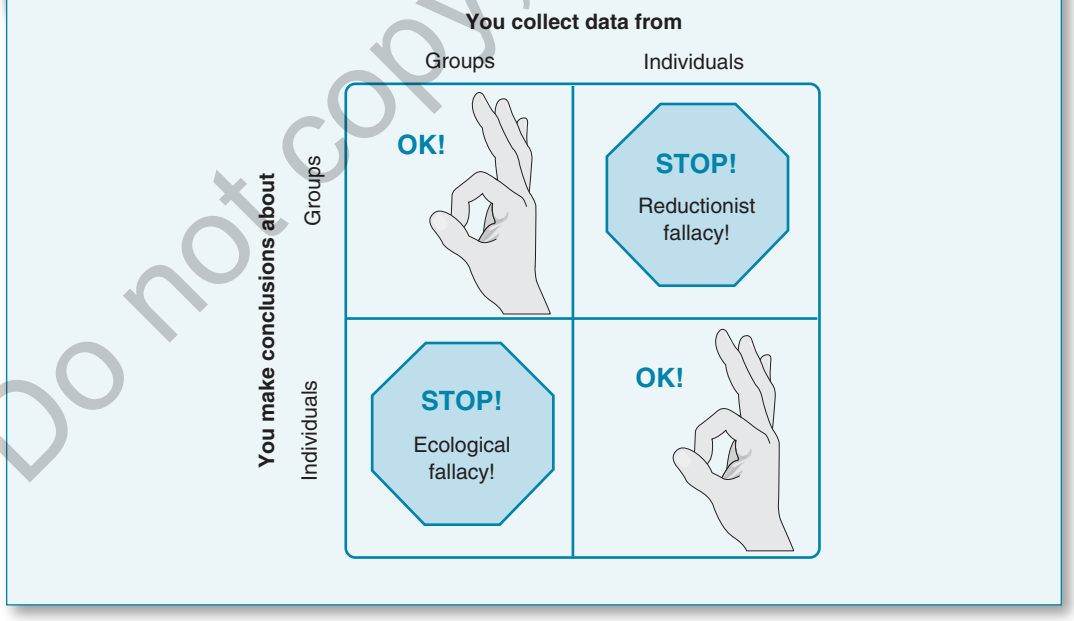
Researchers should make sure that their causal conclusions reflect the units of analysis in their study. Conclusions about processes at the individual level should be based on individual-level data; conclusions about group-level processes should be based on data collected about groups. In most cases, when this rule is violated, we can be misled about the existence of an association between two variables.

A researcher who draws conclusions about individual-level processes from group-level data could be making what is termed an **ecological fallacy** (see Exhibit 6.5). The conclusions may or may not be correct, but we must recognize that group-level data do not necessarily reflect solely individual-level processes. For example, a researcher may examine school records and find that the higher the percentage of male students at the high school, the higher the participation in school-sponsored sports activities. But the researcher would commit an ecological fallacy if she then concluded that boys are more interested in engaging in sports than girls are. This conclusion is about an individual-level causal process (the relationship between individual students and sports participation), even though the data describe groups (schools). It could actually be that prospective female athletes find there are fewer opportunities (not as many teams, lack of coaches) and poorer conditions (no locker rooms or shower facilities, lack of equal access to the gym or playing fields) and so do not participate. This is exactly the scenario that faced girls interested in athletics prior to federal Title IX legislation. Exploding participation in school sports by girls in the wake of Title IX proved that the problem was in the organization of schools, not in the interest of the girls.

Don't be too quick to reject all conclusions about individual processes based on group-level data; just keep in mind the possibility of an ecological fallacy. If we don't have individual-level data, we can't be sure that patterns at the group level will hold at the individual level.

On the other hand, when data about individuals are used to make inferences about group-level processes, a problem occurs that can be thought of as the mirror image of the ecological fallacy: the **reductionist fallacy,** also known as *reductionism,* or the *individualist fallacy* (see Exhibit 6.5). For example, Robert J. Sampson and William Julius Wilson (1995, pp. 37–38; Wilson, 1987, p. 58) note that we can be misled into concluding from individual-level data that race has a causal effect on violence because there is an association at the individual level between race and the likelihood of arrest for violent crime. However, community-level data reveal that a much higher percentage of poor Blacks live in high-poverty areas, as compared to poor

Exhibit 6.5 Errors in Causal Conclusions



Source: Schutt (2009, p. 193).

Whites. The concentration of African Americans in poverty areas, not the race or other characteristics of the individuals in these areas, may be the cause of higher rates of violence. Explaining violence in this case requires community-level data.

The fact that errors in causal reasoning can be made should not deter you from conducting research with aggregate data or make you unduly critical of researchers who make inferences about individuals on the basis of aggregate data. The solution is to know what the units of analysis and **units of observation** were in a study and to take these into account in weighing the credibility of the researcher's conclusions. The goal is not to reject out of hand conclusions that refer to a level of analysis different from what was actually studied. Instead, the goal is to consider the likelihood that an ecological fallacy or a reductionist fallacy has been made when estimating the causal validity of the conclusions.

2 True Experimental Designs

Experimental research provides the most powerful design for testing causal hypotheses because it allows us to establish confidently the first three criteria for causality—association, time order, and nonspuriousness. True experiments have at least three features that help us meet these criteria:

True experiment: Experiment in which subjects are assigned randomly to an experimental group that receives a treatment or other manipulation of the independent variable and a comparison group that does not receive the treatment or receives some other manipulation. Outcomes are measured in a posttest.

1. Two comparison groups (in the simplest case, an experimental group and a control group), which establishes association
2. Variation in the independent variable before assessment of change in the dependent variable, which establishes time order
3. Random assignment to the two (or more) comparison groups, which establishes nonspuriousness

Experimental group: In an experiment, the group of subjects that receives the treatment or experimental manipulation.

Comparison group: In an experiment, groups that have been exposed to different treatments, or values of the independent variable (e.g., a control group and an experimental group).

Control group: A comparison group that receives no treatment.

We can determine whether an association exists between the independent and dependent variables in a true experiment because two or more groups differ in terms of their value on the independent variable. One group receives some **treatment** (also called an "experimental treatment"), which is an intervention, stimulus, or some other purposely manipulated factor that affects the value of the independent variable. In a drug trial, a treatment can be a new medication. In a school, a treatment might be a new instructional technique or a new textbook. The group receiving the treatment is termed the experimental group. In a simple experiment, there is a second group that does not receive the treatment; it is termed the control group.

Consider an example in detail (see the simple diagram in Exhibit 6.6). Does drinking coffee improve one's writing of an essay? Imagine a simple experiment. Suppose you believe that drinking two cups of strong coffee before class will help you in writing an in-class essay. But other people think that coffee makes them too nervous and "wired" and so doesn't help in writing the essay. To test your hypothesis ("Coffee drinking causes improved performance"), you need to compare two groups of subjects, a control group and an experimental group. First, the two groups will sit and write an in-class essay. Then, the control group will drink no coffee, while the experimental group will drink two cups of strong coffee. Next, both groups will sit and write another in-class essay. At the end, all of the essays will be graded, and you will see whether the experimental group improved more than the control group. Thus, you may establish *association.*

Exhibit 6.6 A True Experiment

Experimental Group:	**R**	**O_1**	**X**	**O_2**
Comparison Group:	**R**	**O_1**		**O_2**

Key: R = Random assignment
O = Observation (pretest [O_1] or posttest [O_2])
X = Experimental treatment

	O_1	X	O_2
Experimental Group	Pretest Essay	Coffee	Posttest Essay
Comparison Group	Pretest Essay		Posttest Essay

Source: Chambliss and Schutt (2010, p. 136).

If you only conduct a survey and find that people who drink coffee score higher on tests, you can't be sure about the time order of effects. Perhaps people who write better have more time on their hands and so are more likely to go to coffeehouses and drink coffee and relax. By controlling who gets the coffee and when, we can establish *time order.*

All true experiments have a posttest—that is, a measurement of the outcome in both groups after the experimental group has received the treatment. In our example, you grade the papers. Many true experiments also have pretests, which measure the dependent variable before the experimental intervention. A pretest is the same as a posttest, just administered at a different time. Strictly speaking, though, a true experiment does not require a pretest. When researchers use random assignment, the groups' initial scores on the dependent variable and on all other variables are very likely to be similar. Any difference in outcome between the experimental and comparison groups is therefore likely to be due to the intervention (or to other processes occurring during the experiment), and the likelihood of a difference just on the basis of chance can be calculated.

Finally, it is crucial that the two groups be more or less equal at the beginning of the study. If you let students choose which group to be in, the more ambitious students may pick the coffee group, hoping to stay awake and do better on the paper. Or people who simply don't like the taste of coffee may choose the noncoffee group. Either way, your two groups won't be equivalent at the beginning of the study, so any difference in their writing may be the result of that initial difference (a source of spuriousness), not the drinking of coffee.

So you randomly sort the students into the two different groups. You can do this by flipping a coin for each student, by pulling names out of a hat, or by using a random number table or a computer program that generates random numbers. In any case, the subjects themselves should not be free to choose, nor should you (the experimenter) be free to put them into whatever group you want. (If you did that, you might unconsciously put the better students into the coffee group, hoping to get the results you're looking for.) Thus, we can achieve nonspuriousness with an experimental design.

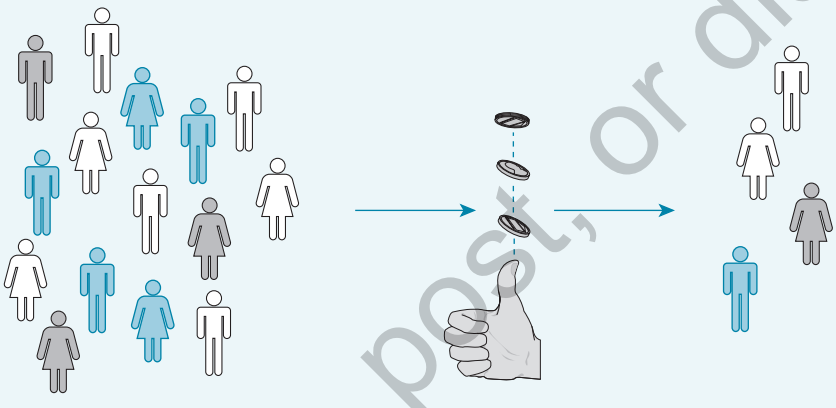
The Tennessee STAR class size project, a true experiment, used randomization to reduce the risk of spuriousness. Students and teachers were randomly assigned to one of three types of classes: small class, regular-size class, or regular-size class with a teacher's aide (A. Krueger & Whitmore, 2001; Schanzenbach, 2006). As a result, the different groups were likely to be equivalent in all respects at the outset of the experiment. In general, the greater the number of cases assigned randomly to the groups, the more likely that the groups will be equivalent in all respects. The STAR experiment involved more than 11,000 students, and because students were randomly assigned, student characteristics such as free lunch status and amount of parental involvement were, on average, the same across class types (A. Krueger & Whitmore, 2001; Schanzenbach, 2006).

Note that the random assignment of subjects to experimental and comparison groups is not the same as random sampling of individuals from some larger population (see Exhibit 6.7). In fact, **random assignment** (randomization) does not help at all to ensure that the research subjects are representative of some larger population—representativeness is the goal of random sampling. What random assignment does—create two (or more) equivalent groups—is useful for ensuring internal (causal) validity, not generalizability.

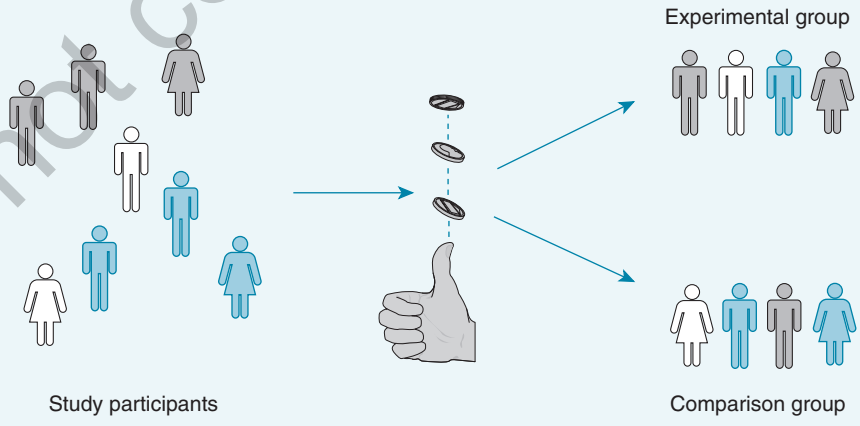
Matching is another procedure sometimes used to equate experimental and comparison groups, but by itself, it is a poor substitute for randomization. One method is to match pairs of individuals (see Exhibit 6.8).

Exhibit 6.7 Random Sampling Versus Random Assignment

Random sampling (a tool for ensuring generalizability):
Individuals are randomly selected from a population to participate in a study.

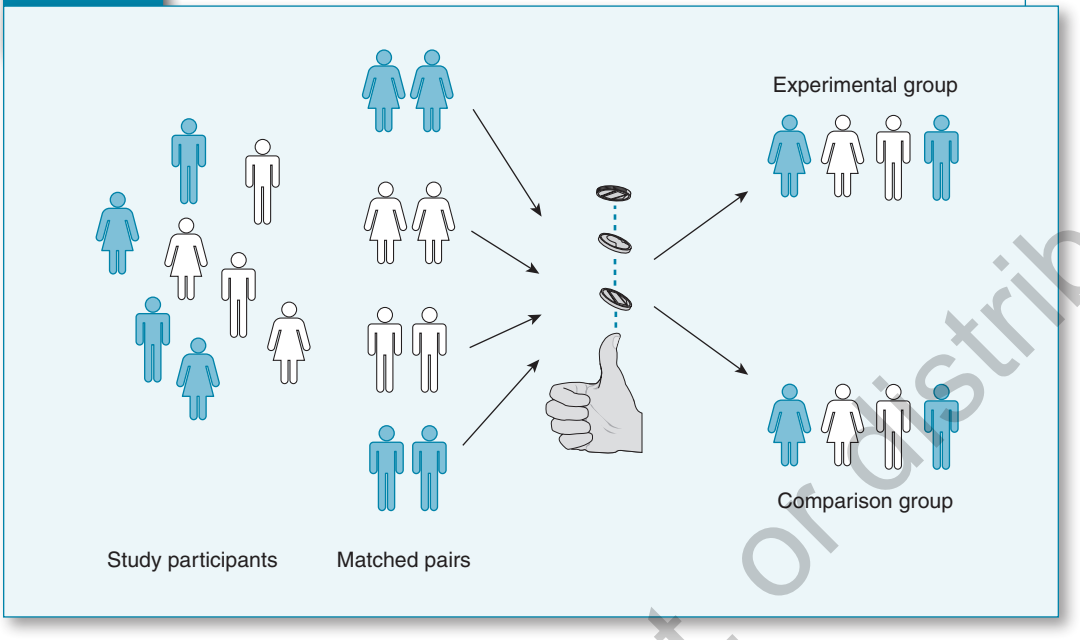


Random assignment, or randomization (a tool for ensuring internal validity):
Individuals who are to participate in a study are randomly divided into an experimental group and a comparison group.



Source: Chambliss and Schutt (2010, p. 138).

Exhibit 6.8 Experimental Design Combining Matching and Random Assignment



Source: Schutt (2009, p. 228).

You start by identifying important characteristics that might affect the study, and then you match pairs of individuals with similar or identical characteristics. In a study of middle school teachers, you might match subjects by gender, education, and years of teaching experience and then assign each member of a pair to the experimental or control groups. This method eliminates the possibility of differences due to chance in the gender, education, and experience composition of the groups. The basic problem is that, as a practical matter, individuals can be matched on only a few characteristics, and so unmatched differences between the experimental and comparison groups may still influence outcomes. However, when matching is combined with randomization, it can reduce the possibility of differences due to chance. A second problem with matching occurs when one member of the matched pair drops out of the study, unbalancing the groups. In this case, researchers will often exclude the findings of the individual who remained in the study.

2 Quasi-Experimental Designs

Despite its advantages for establishing causation, testing a hypothesis with a true experimental design is often not feasible. A true experiment may be too costly or take too long to carry out, it may not be ethical to randomly assign subjects to the different conditions, or it may be too late to do so. For these reasons, researchers may use "quasi-experimental" designs that retain several components of experimental design but do not include randomization.

In quasi-experimental designs, a comparison group is predetermined to be comparable to the treatment group in critical ways, such as being eligible for the same services or being in the same school cohort

(Rossi & Freeman, 1989, p. 313). These research designs are *quasi*-experimental because subjects are not randomly assigned to the comparison and experimental groups. As a result, we cannot be as confident in the comparability of the groups as in true experimental designs. Nonetheless, to term a research design quasi-experimental, we have to be sure that the comparison groups meet specific criteria.

We will discuss here the two major types of quasi-experimental designs (other types can be found in Cook & Campbell, 1979; Mohr, 1992):

Quasi-experimental design: A research design in which there is a comparison group that is comparable to the experimental group in critical ways but subjects are not randomly assigned to the comparison and experimental groups.

Nonequivalent control group design: A quasi-experimental design in which there are experimental and comparison groups that are designated before the treatment occurs but are not created by random assignment.

Before-and-after design: A quasi-experimental design consisting of several before-and-after comparisons involving the same variables but different groups.

- *Nonequivalent control group designs* have experimental and comparison groups that are designated before the treatment occurs but are not created by random assignment.
- *Before-and-after designs* have a pretest and posttest but no comparison group. In other words, the subjects exposed to the treatment serve, at an earlier time, as their own control group.

If quasi-experimental designs are longitudinal, they can establish time order. Where these designs are weaker than true experiments is in establishing the nonspuriousness of an observed association—that it does not result from the influence of some third, uncontrolled variable. On the other hand, because these quasi-experiments do not require the high degree of control necessary to achieve random assignment, quasi-experimental designs can be conducted using more natural procedures in more natural settings, so we may be able to achieve a more complete understanding of causal context. In identifying the mechanism of a causal effect, quasi-experiments are neither better nor worse than experiments.

Nonequivalent Control Group Designs

In this type of quasi-experimental design, a comparison group is selected so as to be as comparable as possible to the treatment group. Two selection methods can be used:

1. *Individual matching*—Individual cases in the treatment group are matched with similar individuals in the comparison group. This can sometimes create a comparison group that is very similar to the experimental group, such as when Head Start participants were matched with their siblings to estimate the effect of participation in Head Start (Garces et al., 2002). However, in many studies, it may not be possible to match on the most important variables.
2. *Aggregate matching*—In most situations when random assignment is not possible, the second method of matching makes more sense: identifying a comparison group that matches the treatment group in the aggregate rather than trying to match individual cases. This means finding a comparison group that has similar distributions on key variables: the same average age, the same percentage female, and so on. For this design to be considered quasi-experimental, however, it is important that individuals themselves have *not* chosen to be in the treatment group or the control group.

Before-and-After Designs

The common feature of before-and-after designs is the absence of a comparison group: All cases are exposed to the experimental treatment. The basis for comparison is instead provided by the pretreatment

measures in the experimental group. These designs are thus useful for studies of interventions that are experienced by virtually every case in some population, such as a whole-school reform program or introduction of a new mathematics curriculum affecting all the mathematics classes in a school.

The simplest type of before-and-after design is the fixed-sample panel design. As you learned earlier, in a panel design, the same individuals are studied over time; the research may entail one pretest and one posttest. However, this simple type of before-and-after design does not qualify as a quasi-experimental design because comparing subjects to themselves at just one earlier point in time does not provide an adequate comparison group. Many influences other than the experimental treatment may affect a subject following the pretest—for instance, basic life experiences for a young subject.

Time-Series Designs

A time-series design typically involves only one group for which multiple observations of data have been gathered both prior to and after the intervention. Although many methodologists distinguish between repeated-measures panel designs, which include several pretest and posttest observations, and time-series designs, which include many (preferably 30 or more) such observations in both pretest and posttest periods, we do not make this distinction here.

A common design is the *interrupted time-series design,* in which three or more observations are taken before and after the intervention. It looks like this:

$$\text{Experimental Group } O_1\ O_2\ O_3\ X\ O_4\ O_5\ O_6$$

As with other designs, there are variations on this basic design, including time-series designs with comparison or control groups and time-series designs in which multiple observations are also gathered during the course of the intervention.

One advantage of a time-series design is that there is only one group, so a second group need not be created. This is very useful when, for instance, one wishes to study a single classroom over the course of a marking period or a year. A second advantage is that, depending on the question, both the pretest and posttest observations need not occur prospectively; rather, the impacts of the programmatic or policy changes can be based on data already collected. For instance, if X in the diagram above referred to adoption of a new teaching strategy for the second marking period, then O_1, O_2, and O_3 could be grades for tests already taken in the first marking period.

A time-series design is based on the idea that, by taking repeated measures prior to an intervention or programmatic change, you have the opportunity to identify a pattern. A pattern may show a trend reflecting an ongoing increase or decline or it may simply stay flat. Having identified the preintervention pattern, the question is whether an intervention or program altered the nature of the pattern to what is considered a more favorable state.

What can we say about causality when using a time-series design? The before-and-after comparison enables you to determine whether an *association* exists between the intervention and the dependent variable. You can determine whether the change in the dependent variable occurred after the intervention, so *time order* is not a problem. However, there is no control group, so we cannot rule out the influence of extraneous factors as the actual cause of the change we observed; *spuriousness* may be a problem. Some other event may have occurred during the study that resulted in a change in posttest scores. What you *can* determine is that a trend caused by other factors did not cause the change in the dependent variable from before to after the exposure to the independent variable. Overall, the longitudinal nature of before-and-after designs can help to identify causal mechanisms, while the loosening of randomization requirements makes it easier to conduct studies in natural settings, where we learn about the influence of contextual factors.

2 Threats to Validity in Experimental Designs

Experimental designs, like any research design, must be evaluated for their ability to yield valid conclusions. Remember, there are three kinds of validity: internal (causal) validity, external validity (generalizability), and measurement. True experiments are good at producing internal validity, but they fare less well in achieving external validity (generalizability). Quasi-experiments may provide more generalizable results than true experiments but are more prone to problems of internal invalidity. Measurement validity is a central concern for both kinds of research, but even a true experimental design offers no special advantages or disadvantages in measurement.

In general, nonexperimental designs, such as those used in survey research and field research, offer less certainty of internal validity, a greater likelihood of generalizability, and no particular advantage or disadvantage in terms of measurement validity. In this section, we focus on the ways in which experiments help (or don't help) to resolve potential problems of internal validity and generalizability.

Threats to Internal Causal Validity

The following sections discuss threats to validity (also referred to as "sources of invalidity") that occur frequently in social science research, including educational research (see Exhibit 6.9). These "threats" exemplify five major types of problems that arise in research design.

Noncomparable Groups

The problem of noncomparable groups occurs when the experimental group and the control group are not really comparable—that is, when something interferes with the two groups being essentially the same at the beginning (or end) of a study.

- *Selection bias*—Occurs when the subjects in your groups are initially different. If the ambitious students decide to be in the "coffee" group, you'll think their performance was helped by coffee—but it could have been their ambition.

Everyday examples of selection bias are everywhere. Harvard graduates are very successful people, but Harvard *admits* students who are likely to be successful anyway. Maybe Harvard itself had no effect on them. A few years ago, a psychotherapist named Mary Pipher wrote a best seller called *Reviving Ophelia* (1994) in which she described the difficult lives of—as she saw it—typical adolescent girls. Pipher painted a stark picture of depression, rampant eating disorders, low self-esteem, academic failure, suicidal thoughts, and even suicide itself. Where did she get this picture? From girls who selected themselves to be her patients—that is, from adolescent girls who were in deep despair or at least were unhappy enough to seek help. If Pipher had talked with a comparison sample of girls who hadn't sought help, perhaps the story would not have been so bleak.

- *Mortality*—Even when random assignment works as planned, the groups can become different over time because of mortality, or differential attrition; this can also be called "deselection." That is, the groups become different because subjects are more likely to drop out of one of the groups for various reasons. At some colleges, satisfaction surveys show that seniors are more likely to rate their colleges positively than are freshmen. But remember that the freshmen who really hated the place may have transferred out, so their ratings aren't included with senior ratings. In effect, the lowest scores are removed; that's a mortality problem.

Exhibit 6.9 Threats to Internal Validity

Problem	Example	Type
Selection	Girls who choose to see a therapist are not representative of population.	Noncomparable Groups
Mortality	Students who most dislike college drop out, so aren't surveyed.	Noncomparable Groups
Instrument Decay	Interviewer tires, losing interest in later interviews, so poor answers result.	Noncomparable Groups
Testing	If someone has taken the SAT before, they are familiar with the format, so do better.	Endogenous Change
Maturation	Everyone gets older in high school; it's not the school's doing.	Endogenous Change
Regression	The lowest-ranking students on IQ must improve their rank; they can't do worse.	Endogenous Change
History	The O. J. Simpson trial affects members of diversity workshops.	History
Contamination	"John Henry" effect; people in study compete with one another.	Contamination
Experimenter Expectation	Researchers unconsciously help their subjects, distorting results.	Treatment Misidentification
Placebo Effect	Fake pills in medical studies produce improved health.	Treatment Misidentification
Hawthorne Effect	Workers enjoy being subjects and work harder.	Treatment Misidentification

Source: Chambliss and Schutt (2010, p. 145).

Note that whenever subjects are not assigned randomly to treatment and comparison groups, the threat of selection bias or mortality is very great. Even if the comparison group matches the treatment group on important variables, there is no guarantee that the groups were similar initially in terms of either the dependent variable or some other characteristic. However, a pretest helps the researchers to determine and control for selection bias.

- *Instrument decay*—Measurement instruments of all sorts wear out, producing different results for cases studied later in the research. An ordinary spring-operated bathroom scale, for instance, may become "soggy" after some years, showing slightly heavier weights than would be correct. Or a college teacher—a kind of instrument for measuring student performance—gets tired after reading too many papers one weekend and starts giving everyone a B. Research interviewers can get tired or bored, too, leading perhaps to shorter or less thoughtful answers from subjects. In all these cases, the measurement instrument has "decayed" or worn out and so would result in a pretest to posttest change that is not due to the experimental treatment itself.

Endogenous Change

The next three problems, subsumed under the label *endogenous change,* occur when natural developments in the subjects, independent of the experimental treatment itself, account for some or all of the observed change between pretest and posttest.

- *Testing*—Taking the pretest can itself influence posttest scores. As the Kaplan SAT prep courses attest, there is some benefit just to getting used to the test format. Having taken the test beforehand can be an advantage. Subjects may learn something or may be sensitized to an issue by the pretest and, as a result, respond differently the next time they are asked the same questions on the posttest.
- *Maturation*—Changes in outcome scores during experiments that involve a lengthy treatment period may be due to maturation. Subjects may age, gain experience, or grow in knowledge—all as part of a natural maturational experience—and thus respond differently on the posttest than on the pretest. In many high school yearbooks, seniors are quoted as saying, for instance, "I started at West Geneva High School as a boy and leave as a man. WGHS made me grow up." Well, he probably would have grown up anyway, high school or not. WGHS wasn't the cause.
- *Regression*—Subjects who are chosen for a study because they received very low scores on a test may show improvement in the posttest, on average, simply because some of the low scorers on the pretest were having a bad day. Whenever subjects are selected for study because of extreme scores (either very high or very low), the next time you take their scores, they will likely "regress," or move toward the average. For instance, suppose you give an IQ test to third graders and then pull the bottom 20% of the class out for special attention. The next time that group (the 20%) takes the test, they'll almost certainly do better—and not just because of testing practice. In effect, they *can't* do worse—they were at the bottom already. On average, they must do better. A first-time novelist writes a wonderful book and gains worldwide acclaim and a host of prizes. The next book is not so good, and critics say, "The praise went to her head." But it may not have; she *couldn't* have done better. Whenever you pick people for being on an extreme end of a scale, odds are that next time, they'll be more average. This is called the *regression effect.*

Regression effect: A source of causal invalidity that occurs when subjects who are chosen for a study because of their extreme scores on the dependent variable become less extreme on the posttest due to natural cyclical or episodic change in the variable.

Testing, maturation, and regression effects are generally not a problem in experiments that have a control group because they would affect the experimental group and the comparison group equally. However, these effects could explain any change over time in most before-and-after designs because these designs do not have a comparison group. Repeated measures, panel studies, and time-series designs are better in this regard because they allow the researcher to trace the pattern of change or stability in the dependent variable up to and after the treatment. Ongoing effects of maturation and regression can thus be identified and taken into account.

History

History, or external events during the experiment (things that happen outside the experiment), could change subjects' outcome scores. Examples are newsworthy events that concern the focus of an experiment and major disasters to which subjects are exposed. If you were running a series of diversity workshops for some insurance company employees while the notorious 1995 O. J. Simpson murder trial was taking place, for instance, participants' thoughts on race relations at the end of the workshops may say less about your training course than about O. J. Simpson or about their own relationship with the judicial system. This problem is often referred to as a history effect—history during the experiment, that is. It is a particular concern in before-and-after designs.

Causal conclusions can be invalid in some true experiments because of the influence of external events. For example, in an experiment in which subjects go to a special location for the treatment, something at that location unrelated to the treatment could influence these subjects. External events are a major concern in studies that compare the effects of programs in different cities or states (Hunt, 1985, pp. 276–277).

Contamination

Contamination occurs in an experiment when the comparison and treatment groups somehow affect each other. When comparison group members know they are being compared, they may increase their efforts just to be more competitive. This has been termed compensatory rivalry, or the John Henry effect, named after the "steel-driving man" of the folk song, who raced against a steam drill in driving railroad spikes and killed himself in the process. Knowing that they are being denied some advantage, comparison group subjects may as a result increase their efforts to compensate. On the other hand, comparison group members may become demoralized if they feel that they have been left out of some valuable treatment, performing worse than expected as a result. Both compensatory rivalry and demoralization thus distort the impact of the experimental treatment.

Treatment Misidentification

Sometimes the subjects experience a "treatment" that wasn't intended by the researcher. The following are three possible sources of treatment misidentification:

1. *Expectancies of experiment staff*—Change among experimental subjects may be due to the positive expectancies of experiment staff who are delivering the treatment rather than to the treatment itself. Even well-trained staff may convey their enthusiasm for an experimental program to the subjects in subtle ways. This is a special concern in evaluation research, when program staff and researchers may be biased in favor of the program for which they work and are eager to believe that their work is helping clients. Such positive staff expectations thus create a self-fulfilling prophecy.
2. *Placebo effect*—In medicine, a *placebo* is a chemically inert substance (a sugar pill, for instance) that looks like a drug but actually has no direct physical effect. Research shows that such a pill can actually produce positive health effects in two thirds of patients suffering from relatively mild medical problems (Goleman, 1993, p. C3). In other words, if you wish that a pill will help, it often actually does. In social science research, such placebo effects occur when subjects think their behavior should improve through an experimental treatment and then it does—not from the treatment, but from their own belief. Researchers might then misidentify the treatment as having produced the effect.
3. *Hawthorne effect*—Members of the treatment group may change in terms of the dependent variable because their participation in the study makes them feel special. This problem could occur when treatment group members compare their situation to that of members of the control group who are not receiving the treatment, in which case it would be a type of contamination effect. But experimental group members could feel special simply because they are in the experiment. This is termed a *Hawthorne effect* after a classic worker productivity experiment conducted at the Hawthorne electric plant outside Chicago in the 1920s. No matter what conditions the researchers changed to improve or diminish productivity (for instance, increasing or decreasing the lighting in the plant), the workers seemed to work harder simply because they were part of a special experiment. Oddly enough, some later scholars suggested that in the original Hawthorne studies, there was actually a selection bias, not a true Hawthorne effect—but the term has stuck (see Bramel & Friend, 1981). Hawthorne effects are also a concern in evaluation research, particularly when program clients know that the research findings may affect the chances for further program funding.

Process analysis is a technique for avoiding treatment misidentification (Hunt, 1985, pp. 272–274). Periodic measures are taken throughout an experiment to assess whether the treatment is being delivered as planned. Process analysis is often a special focus in evaluation research because of the possibility of improper implementation of the experimental program. For example, many school reform initiatives attempt to replicate their model program design in widely diverse school contexts. If we want to evaluate the impact of the innovation, we need to monitor whether the adopting school is implementing the model, which can be regarded as a "treatment," fully and correctly.

Generalizability

The need for generalizable findings can be thought of as the Achilles heel of true experimental design. The design components that are essential for a true experiment and that minimize the threats to causal validity make it more difficult to achieve sample generalizability—being able to apply the findings to some clearly defined larger population—and cross-population generalizability—generalizing across subgroups and to other populations and settings.

Subjects who can be recruited for a laboratory experiment, randomly assigned to a group, and kept under carefully controlled conditions for the duration of the study may not be representative of any large population of interest to educational researchers. Can they be expected to react to the experimental treatment in the same way as members of the larger population? The generalizability of the treatment and of the setting for the experiment also must be considered (Cook & Campbell, 1979, pp. 73–74). The more artificial the experimental arrangements, the greater the problem (D. T. Campbell & Stanley, 1966, pp. 20–21).

Cross-Population Generalizability

Researchers often are interested in determining whether treatment effects identified in an experiment hold true across different populations, times, or settings. When random selection is not feasible, the researchers may be able to increase the cross-population generalizability of their findings by selecting several different experimental sites that offer marked contrasts on key variables (Cook & Campbell, 1979, pp. 76–77).

Within a single experiment, researchers also may be concerned with whether the relationship between the treatment and the outcome variable holds true for certain subgroups. This demonstration of "external validity" is important evidence about the conditions that are required for the independent variable(s) to have an effect. School- and student-based research studies may not involve participants that are diverse in terms of income level and cultural/ethnic background, making it even more important for researchers to examine the relationship between the independent and dependent variable for all subgroups, not just one or two.

Finding that effects are consistent across subgroups does not establish that the relationship also holds true for these subgroups in the larger population, but it does provide supportive evidence. We have already seen examples of how the existence of treatment effects in particular subgroups of experimental subjects can help us predict the cross-population generalizability of the findings.

There is always an implicit trade-off in experimental design between maximizing causal validity and generalizability. The more that assignment to treatments is randomized and all experimental conditions are controlled, the less likely it is that the research subjects and setting will be representative of the larger population. However, although we need to be skeptical about the generalizability of the results of a single experimental test of a hypothesis, the body of findings accumulated from many experimental tests with different people in different settings can provide a very solid basis for generalization (D. T. Campbell & Russo, 1999, p. 143).

Interaction of Testing and Treatment

A variant on the problem of external validity occurs when the experimental treatment has an effect only when particular conditions created by the experiment occur. One such problem occurs when the treatment has an effect only if subjects have had the pretest. The pretest sensitizes the subjects to some issue so that when they are exposed to the treatment, they react in a way they would not have reacted if they had not taken the pretest. In other words, testing and treatment interact to produce the outcome. For example, answering questions in a pretest about racial prejudice may sensitize subjects so that when they are exposed to the experimental treatment, seeing a film about prejudice, their attitudes are different from what they would have been. In this situation, the treatment truly had an effect, but it would not have had an effect if it were repeated without the sensitizing pretest. This possibility can be evaluated by using the Solomon Four-Group Design to compare groups with and without a pretest (see Exhibit 6.10). If testing and treatment do interact, the difference in outcome scores between the experimental and comparison groups will be different for subjects who took the pretest and those who did not.

Exhibit 6.10 **Solomon Four-Group Design Testing the Interaction of Pretesting and Treatment**

Experimental group:	R	O_1	X	O_2
Comparison group:	R	O_1		O_2
Experimental group:	R		X	O_2
Comparison group:	R			O_2

Key: R = Random assignment
O = Observation (pretest or posttest)
X = Experimental treatment

Source: Chambliss and Schutt (2010, p. 154).

As you can see, no single procedure establishes the external validity of experimental results. Ultimately, we must base our evaluation of external validity on the success of replications taking place at different times and places and using different forms of the treatment.

Limitations of True Experimental Designs

The distinguishing features of true experiments—experimental and comparison groups, pretests (which are not always used) and posttests, and randomization—do not help researchers identify the mechanisms by which treatments have their effects. In fact, this question of causal mechanisms often is not addressed in experimental research. The hypothesis test itself does not require any analysis of mechanism, and if the experiment was conducted under carefully controlled conditions during a limited span of time, the causal effect (if any) may seem to be quite direct. But attention to causal mechanisms can augment experimental findings. Evaluation researchers often focus attention on the mechanisms by which an educational program has its effect (Mohr, 1992, p. 25–27; Scriven, 1972). The goal is to measure the intermediate steps that lead to the change that is the program's primary focus.

True experimental designs also do not guarantee that the researcher has been able to maintain control over the conditions to which subjects are exposed after they are assigned to the experimental and comparison

groups. If these conditions begin to differ, the variation between the experimental and comparison groups will not be what was intended. Such unintended variation is often not much of a problem in laboratory experiments, where the researcher has almost complete control over the conditions. But control over conditions can become a very big concern for field experiments, experimental studies that are conducted in the field, in real-world settings.

2 Nonexperiments

All of the other research designs we study are, of course, "nonexperimental." One of these designs, the ex post facto control group design, is often called quasi-experimental, but that's really not correct. Other designs are covered in other chapters under the headings of "cross-sectional" and "longitudinal" designs. Here, we'll briefly contrast these nonexperimental designs with experimental and quasi-experimental designs.

Ex Post Facto Control Group Designs

The ex post facto control group design is similar to the nonequivalent control group design and is often confused with it, but it does not meet as well the criteria for quasi-experimental designs. This design has experimental and comparison groups that are not created by random assignment, but unlike nonequivalent control group designs, individuals may decide themselves whether to enter the "treatment" or "control" group. As a result, in *ex post facto* (after the fact) designs, the people who join the treatment group may differ because of what attracted them to the group initially, not because of their experience in the group. However, in some studies, we may conclude that the treatment and control groups are so similar at the outset that causal effects can be tested (Rossi & Freeman, 1989, pp. 343–344).

One-Shot Case Studies and Longitudinal Designs

Cross-sectional designs, termed *one-shot case studies* in the experimental design literature, are easily able to establish whether an association exists between two variables, but we cannot be anywhere near as confident in their conclusions about appropriate time order or nonspuriousness as with true experiments or even quasi-experiments. Longitudinal designs improve greatly our ability to test the time order of effects, but they are unable to rule out all extraneous influences.

Christopher Brown (2009) used a one-shot case study design to explore ways in which the child-centered approach used in prekindergarten programs is being incorporated into the accountability-centered environment found in elementary schools. Brown hypothesized that, as more public school systems began to offer prekindergarten programs, there would be a disconnect between their approach and the academic accountability expected in grades after kindergarten. His case study examined implementation of an assessment tool designed to "align the academic achievement expectations of the prekindergarten with those in the corresponding elementary schools" (p. 202). At first, the tool did not work. It was modified to give a more accurate picture of skills students were meant to acquire in kindergarten. The difficulties in aligning the two ways of looking at children and instruction showed how complex it is to merge a child-centered orientation with test and standards-centered approaches and that the process requires effort and compromise on both sides.

Summary: Causality in Nonexperiments

How well do nonexperimental designs allow us to meet the criteria for causality identified earlier in this chapter?

Association: Nonexperiments can provide clear evidence of association between the independent and dependent variables.

Time order: For the most part, cross-sectional designs cannot establish time order. Longitudinal designs, even when nonexperimental, do allow identification of time order.

Nonspuriousness: Nonexperimental designs only weakly address the need to ensure nonspurious relationships because it is unlikely that we will be able to control for all potential **extraneous variables** that may confound the relationship between the independent and dependent variables.

Mechanism: Nonexperimental designs have no particular advantages or disadvantages for establishing causal mechanisms, although qualitative research designs facilitate investigations about causal process.

Context: Because they make it easy to survey large numbers of widely dispersed persons or organizations, one-shot cross-sectional studies facilitate investigation of **contextual effects.**

2 Conclusions

In this chapter, you have studied the five criteria used to evaluate the extent to which particular research designs may achieve causally valid findings. You have learned how our ability to meet these criteria is shaped by research design features such as units of analysis, use of a cross-sectional or longitudinal design, and use of randomization to deal with the problem of spuriousness. You have also seen why the distinction between experimental and nonexperimental designs has so many consequences for how, and how well, we are able to meet criteria for causation.

We began this chapter by posing the general question, "How do educational strategies affect educational outcomes?" Throughout the chapter, you have seen a variety of research approaches to this question. What conclusions were reached by some of these studies? The Tennessee STAR study (Finn & Achilles, 1990; Schanzenbach, 2006) was a unique, large-scale, randomized trial of the effects of reducing class size in Grades K–3. It reached the conclusion that, other things being equal, smaller classes meant more learning, especially for disadvantaged students. A longitudinal panel study (Garces et al., 2002) concluded that the Head Start program has positive academic and social effects lasting into adolescence and early adulthood. A one-shot case study (C. P. Brown, 2009) showed that a "mismatch" in educational approaches (child centered vs. accountability centered) can create transition problems between kindergarten and elementary grades unless educators from both grade levels work together to integrate their approaches.

We also looked at researchers (Agnew et al., 2008; Sampson & Laub, 1994; Sampson & Wilson, 1995) who used a variety of methods to explore the relationship between poverty, delinquency, and low school achievement and concluded that poverty alone does not necessarily result in delinquency and low achievement. Rather, economic factors increase the likelihood of familial and social breakdowns that, in some cases but not others, lead to negative outcomes. A longitudinal study based on years of public data (Skinner, 2009) looked at charter schools in the city of Boston and found evidence that "push-out" strategies in some

charter schools caused low achievers and problem students to leave the school before senior year, creating the impression that a higher percentage of students was graduating than was actually the case.

We should reemphasize that the results of any particular study are part of an always changing body of empirical knowledge about educational reality. Thus, our understandings of causal relationships are always partial. Researchers always wonder whether they have omitted some relevant variables from their controls, whether their experimental results would differ if the experiment were conducted in another setting, or whether they have overlooked a critical historical event. But by using consistent definitions of terms and maintaining clear standards for establishing the validity of research results—and by expecting the same of others who do research—educational researchers can contribute to a growing body of knowledge that can reliably guide educational policy and understanding.

When you read the results of an educational study, you should now be able to evaluate critically the validity of the study's findings. If you plan to engage in educational research, you should now be able to plan an approach that will lead to valid findings. And with a good understanding of three dimensions of validity (measurement validity, generalizability, and causal validity) under your belt, and with sensitivity also to the goal of "authenticity," you are ready to focus on the major methods of data collection used by educational researchers.

Key Terms

Highlights

- Three criteria are generally viewed as necessary for identifying a causal relationship: association between the variables, proper time order, and nonspuriousness of the association. In addition, the basis for concluding that a causal relationship exists is strengthened by identification of a causal mechanism and the context for the relationship.
- Association between two variables is in itself insufficient evidence of a causal relationship. This point is commonly made with the expression "Correlation does not prove causation."
- Experiments use random assignment to make comparison groups as similar as possible at the outset of an experiment to reduce the risk of spurious effects due to extraneous variables.
- Nonexperimental designs use statistical controls to reduce the risk of spuriousness. A variable is controlled when it is held constant so that the association between the independent and dependent variables can be assessed without being influenced by the control variable.
- Ethical and practical constraints often preclude the use of experimental designs.
- Longitudinal designs are usually preferable to cross-sectional designs for establishing the time order of effects. Longitudinal designs vary in terms of whether the same people are measured at different times, how the population of interests is defined, and how frequently follow-up measurements are taken. Fixed-sample panel designs provide the strongest test for the time order of effects, but they can be difficult to carry out successfully because of their expense as well as subject attrition and fatigue.

- We do not fully understand the variables in a study until we know what units of analysis they refer to.
- Invalid conclusions about causality may occur when relationships between variables measured at the group level are assumed to apply at the individual level (the ecological fallacy) and when relationships between variables measured at the level of individuals are assumed to apply at the group level (the reductionist fallacy). Nonetheless, many research questions point to relationships at multiple levels and so may profitably be investigated at multiple units of analysis.

Student Study Site

To assist in completing the web exercises, please access the study site at www.sagepub.com/check, where you will find the web exercise with accompanying links. You'll find other useful study materials such as self-quizzes and e-flashcards for each chapter, along with a group of carefully selected articles from research journals that illustrate the major concepts and techniques.

Discussion Questions

1. Review articles in several newspapers, copying down all causal assertions. These might range from assertions that the stock market declined because of uncertainty in the Middle East to explanations about why a murder was committed or why test scores are declining in U.S. schools. Inspect the articles carefully, noting all evidence used to support the causal assertions. Which criteria for establishing causality are met? What other potentially important influences on the reported outcome have been overlooked?
2. Select several research articles in professional journals that assert, or imply, that they have identified a causal relationship between two or more variables. Are all of the criteria for establishing the existence of a causal relationship met? Find a study in which subjects were assigned randomly to experimental and comparison groups to reduce the risk of spurious influences on the supposedly causal relationship. How convinced are you by the study?

Practice Exercises

1. Search the *American Educational Research Journal* (*AERJ*) or another similar source for several articles on studies using any type of longitudinal design. You will be searching for article titles that use words such as *longitudinal, panel, trend,* or *over time.* How successful were the researchers in carrying out the design? What steps did the researchers who used a panel design take to minimize panel attrition? How convinced are you by those using repeated cross-sectional designs that they have identified a process of change in individuals? Did any researchers use retrospective questions? How did they defend the validity of these measures?
2. Propose a hypothesis involving variables that could be measured with individuals as the units of analysis. How might this hypothesis be restated so as to involve groups as the units of analysis? Would you expect the hypothesis to be supported at both levels? Why or why not? Repeat the exercise, this time starting with a different hypothesis involving groups as the units of analysis and then restating it so as to involve individuals as the units of analysis.

Web Exercises

1. Try out the process of randomization. Go to the website http://www.randomizer.org. Now type numbers into the randomizer with two groups and 20 individuals per group. Repeat the process for four groups and 10 individuals per

group. Plot the numbers corresponding to each individual in each group. Does the distribution of numbers within each group truly seem to be random?

2. Go to the website of the U.S. Department of Education (http://www.ed.gov) and type *user friendly guide* into the search box. Click on the first item in the list that comes up, which should bring you to the publication *Identifying and Implementing Educational Practices Supported by Rigorous Evidence: A User Friendly Guide.* Open the pdf and read sections I, II, and III, which deal with randomized control trials, quasi-experimental designs, and comparison group studies. Do you agree with the designations "strong" and "possible" levels of effectiveness?
3. Read section IV of the "User Friendly Guide," concerning classroom implementation. As an educator, how helpful do you think this section is in giving you guidance for your own work?

Developing a Research Proposal

How will you try to establish the causal effects you hypothesize?

1. Identify at least one hypothesis involving what you expect is a causal relationship.
2. Identify key variables that should be controlled in your survey design to increase your ability to avoid arriving at a spurious conclusion about the hypothesized causal effect. Draw on relevant research literature and social theory to identify these variables.
3. Add a longitudinal component to your research design. Explain why you decided to use this particular longitudinal design.
4. Review the criteria for establishing a causal effect and discuss your ability to satisfy each one. Include in your discussion some consideration of how well your design will avoid each of the threats to experimental validity.