Chapter 5

Causation and Experimental Design

Causal Explanation

What Causes What?

Association
Time Order
Nonspuriousness
Mechanism
Context

Why Experiment?

What If a True Experiment Isn't Possible?

Nonequivalent Control Group Designs
Before-and-After Designs
Ex Post Facto Control Group Designs

What Are the Threats to Validity in Experiments?

Threats to Internal Causal Validity
Noncomparable Groups
Endogenous Change
History
Contamination
Treatment Misidentification
Generalizability
Sample Generalizability
Cross-Population Generalizability
Interaction of Testing and Treatment

How Do Experimenters Protect Their Subjects?

Deception
Selective Distribution of Benefits

Conclusion

Identifying causes—figuring out why things happen—is the goal of most social science research. Unfortunately, valid explanations of the causes of social phenomena do not come easily. Why did the homicide rate in the United States drop for 15 years and then start to rise in 1999 (Butterfield, 2000:12)? Was it because of changes in the style of policing (Radin, 1997:B7) or because of changing attitudes among young people (Butterfield, 1996a)? Was it due to variation in patterns of drug use (Krauss, 1996) or to tougher prison sentences (Butterfield, 1996a) or to more stringent handgun regulations (Butterfield, 1996b)? Did better emergency medical procedures result in higher survival rates for victims (Ramirez, 2002)? If we are to evaluate these alternative explanations we must design our research strategies carefully.

This chapter considers the meaning of causation, the criteria for achieving causally valid explanations, the ways in which experimental and quasi-experimental research designs seek to meet these criteria, and the difficulties that can sometimes result in invalid conclusions. By the end of the chapter, you should have a good grasp of the meaning of causation and the logic of experimental design. Most social research, both academic and applied, uses data collection methods other than experiments. But because experimental designs are the best way to evaluate causal hypotheses, a better understanding of them will help you to be aware of the strengths and weaknesses of other research designs that we will consider in subsequent chapters.

CAUSAL EXPLANATION

A cause is an explanation for some characteristic, attitude, or behavior of groups, individuals, or other entities (such as families, organizations, or cities) or for events. For example, Sherman and Berk (1984) conducted a study to determine whether adults who were accused of a domestic violence offense would be less likely to repeat the offense if police arrested them rather than just warned them. Their conclusion that this hypothesis was correct meant that they believed police response had a **causal effect** on the likelihood of committing another domestic violence offense.

Causal effect: The finding that change in one variable leads to change in another variable, *ceteris paribus* (other things being equal). *Example:* Individuals arrested for domestic assault tend to commit fewer subsequent assaults than similar individuals who are accused in the same circumstances but are not arrested.

More specifically, a causal effect is said to occur if variation in the 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 men earn more income than women do. But is this because they are men—or could it be due to higher levels of education, or to longer tenure in their jobs (with no pregnancy breaks), or is it the kinds of jobs men go into as compared to those that women choose? We want to know if men earn more than women, *ceteris paribus*—other things (job, tenure, education, etc.) being 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 exactly the same circumstances except for the variation in the independent variable (King, Keohane, & Verba, 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.

WHAT CAUSES WHAT?

Five criteria should be considered in trying to establish a causal relationship. The first three criteria are generally considered as requirements for identifying a causal effect: (1) empirical association, (2) temporal priority of the independent variable, 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. They must vary together so when one goes up (or down), the other goes up (or down) at the same time. For example: When cigarette smoking goes up, so does lung cancer. The longer you stay in school, the more money you will make later in life. Single women are more likely to live in poverty than married women. When income goes up, so does overall health. 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 you, but empirically it is the case: There is no correlation. So there cannot be a causal relationship.

Time Order

Association is necessary for establishing a causal effect, but it is not sufficient. We must also ensure that the variation in the independent variable came before variation in the dependent variable—the cause must come before its presumed effect. 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 in our terms), you need to really believe in yourself (the independent variable). And it is true that many very successful politicians, actors, and businesspeople 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 actually due to changes in a third variable, so what appears to be a direct connection is in fact not one. 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 else. 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.

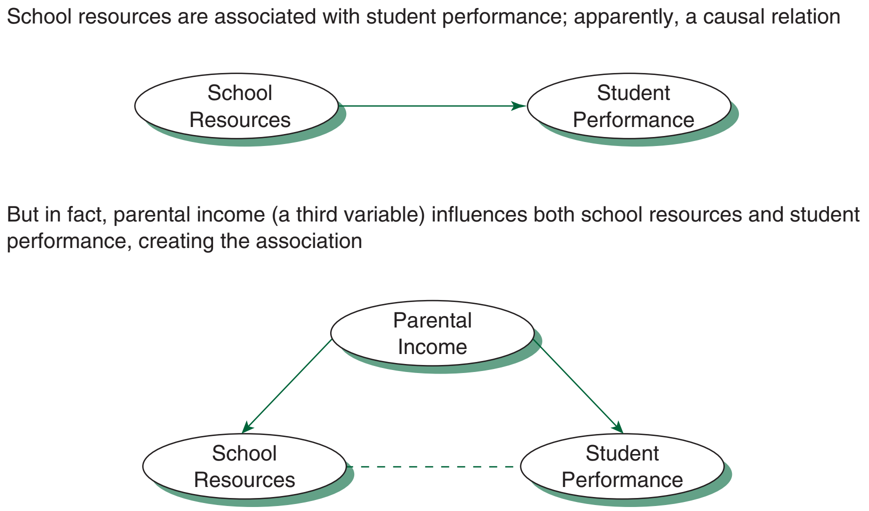
If this point seems obvious, consider a social science example. Do schools with better resources produce better students? There is certainly a correlation, but consider the fact that parents with more education and higher income tend to live in neighborhoods that spend more on their schools. These parents are also more likely to have books in the home and to provide other advantages for their children (see Exhibit 5.1). Maybe parents' income causes variation in both school resources and student performance. If so, there would be an association between school resources and student performance, but it would be at least partially spurious. What we want, then, is *non*spuriousness.

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 that it is hypothesized to cause (Cook & Campbell, 1979:35; Marini & Singer, 1988). Many social scientists (and scientists in other fields) argue that no causal explanation is adequate until a mechanism is identified.

For instance, there seems to be an empirical association at the individual level between poverty and delinquency: Children who live in impoverished homes seem

Exhibit 5.1 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

more likely to be involved in petty crime. But why? Some researchers have argued for a *mechanism* of low parent/child attachment, inadequate supervision of children, and erratic discipline as the means by which poverty and delinquency are connected (Sampson & Laub, 1994). In this way, figuring out some aspects of the process by which the independent variable influenced the variation in the dependent variable can increase confidence in our conclusion that there was a causal effect (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). Identification of the context in which a causal effect occurs is not itself a criterion for a valid causal conclusion, and it is not always attempted; but it does help us to understand the causal relationship.

You may hypothesize, for example, that if you offer employees higher wages to work harder, they will indeed work harder; and in the context of America, this seems to indeed be the case. Incentive pay causes harder work. But in noncapitalist societies, workers often want only enough money to meet their basic needs and would rather work less than drive themselves hard just to have more money.

In America, the correlation of incentive pay with greater effort seems to work; in medieval Europe, for instance, it did not (Weber, 1992).

As another example, in America in the 1960s, children of divorced parents ("from a broken home") were more likely to suffer from a variety of problems; they lived in a context of mostly intact families. In 2006, many parents are divorced, and the causal link between divorced parents and social pathology no longer seems to hold (Coontz, 1997).

WHY EXPERIMENT?

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

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

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" that is a manipulation of the value of the independent variable. This group is termed the **experimental group.** In a simple experiment, there may be one other group that does not receive the treatment; it is termed the **control group.**

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

Control group: A comparison group that receives no treatment.

Consider an example in detail (see the simple diagram in Exhibit 5.2). 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

Exhibit 5.2 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

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 which group improved more. Thus, you may establish *association.*

You may find an association outside the experimental setting, of course, but it won't establish time order. Perhaps good writers hang out in cafés and coffee houses, and then start drinking lots of coffee. So there would be an association, but not the causal relation we're looking for. By controlling who gets the coffee, and when, we 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** that measure the dependent variable before the experimental intervention. A pretest is exactly 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 non-coffee group. Either way, your two groups won't be equivalent at the beginning of the study, and 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 one of them, or by pulling names out of a hat, or by using a random number table as described in the previous chapter. 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 hope to achieve nonspuriousness.

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 5.3). In fact, **random assignment** (**randomization**) does not help at all to ensure that the research subjects are representative of some larger population; instead, representativeness is the goal of random sampling. What random assignment does—create two (or more) equivalent groups—is useful for ensuring internal 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. Matching of individuals in a treatment group with those in a comparison group might involve pairing persons on the basis of similarity of gender, age, year in school, or some other characteristic. The basic problem is that, as a practical matter, individuals can be matched on only a few characteristics; unmatched differences between the experimental and comparison groups may still influence outcomes.

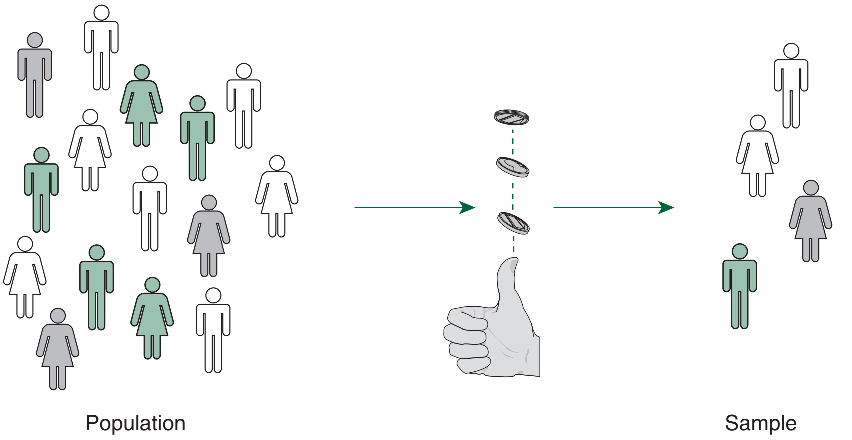
These defining features of true experimental designs give us a great deal of confidence that we can meet the three basic criteria for identifying causes: association, time order, and nonspuriousness. However, we can strengthen our understanding of causal connections, and increase the likelihood of drawing causally valid conclusions, by also investigating causal mechanism and causal context.

WHAT IF A TRUE EXPERIMENT ISN'T POSSIBLE?

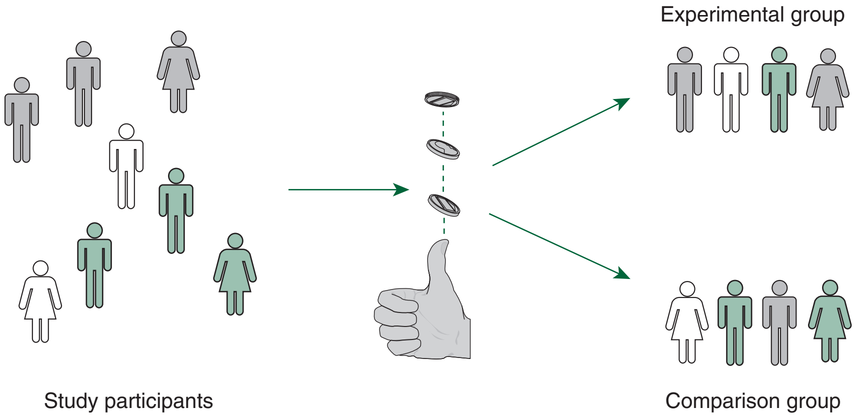
Often, testing a hypothesis with a true experimental design is not feasible. A true experiment may be too costly or take too long to carry out; it may not be ethical to randomly assign subjects to the different conditions; or it may be too late to do

Exhibit 5.3 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.



so. Researchers may instead use "quasi-experimental" designs that retain several components of experimental design but differ in important details.

In **quasi-experimental design**, 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, 1989:313). These research designs are only "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, in order 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, as well as one type—ex post facto (after the fact) control group design—that is often mistakenly termed quasi-experimental (other types can be found in Cook & Campbell, 1979, and Mohr, 1992):

- *Nonequivalent control group designs*—**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*—**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.
- *Ex post facto control group designs*—These designs use nonrandomized control groups designated after the fact.

Exhibit 5.4 diagrams one study using the ex post facto control group design and another study using the multiple group before-and-after design, one type of before-and-after design. (The diagram for an ex post facto control group design is the same as for a nonequivalent control group design, but the two types of experiment differ in how people are able to join the groups.)

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 in order 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, though, 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:

Individual matching—Individual cases in the treatment group are matched with similar individuals in the comparison group. This can sometimes create a comparison

Exhibit 5.4 Quasi-Experimental Designs

Nonequivalent control group design:

Experimental group:		O_1	X_a	O_2
Comparison group 1:		O_1	X_b	O_2
Comparison group 2:		O_1	X_c	O_2
		Pretest	*Treatment*	*Posttest*
Team Interdependence	Group	Team performance	Independent tasks	Team performance
	Hybrid	Team performance	Mixed tasks	Team performance
	Individual	Team performance	Individual tasks	Team performance

Before-and-after design:
Soap-opera suicide and actual suicide (Phillips, 1982)

Experimental group:	O_{11} O_{12} O_{13} O_{14}	X_1 X_2 X_3 X_4	O_{21} O_{22} O_{23} O_{24}
	Pretest	*Treatment*	*Posttest*
	Suicide rate	Soap-opera suicides	Suicide rate

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

Source: Ruth Wageman, 1995. “Interdependence and Group Effectiveness.” *Administrative Science Quarterly, 40:*145–180. Reprinted with permission.

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. However, in many studies it may not be possible to match on the most important variables.

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 must themselves have chosen to be in the treatment group or the control group.

Nonequivalent control group designs allow you to determine whether an association exists between the presumed cause and effect.

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 total coverage programs like Social Security or single-organization studies of the effect of a new management strategy.

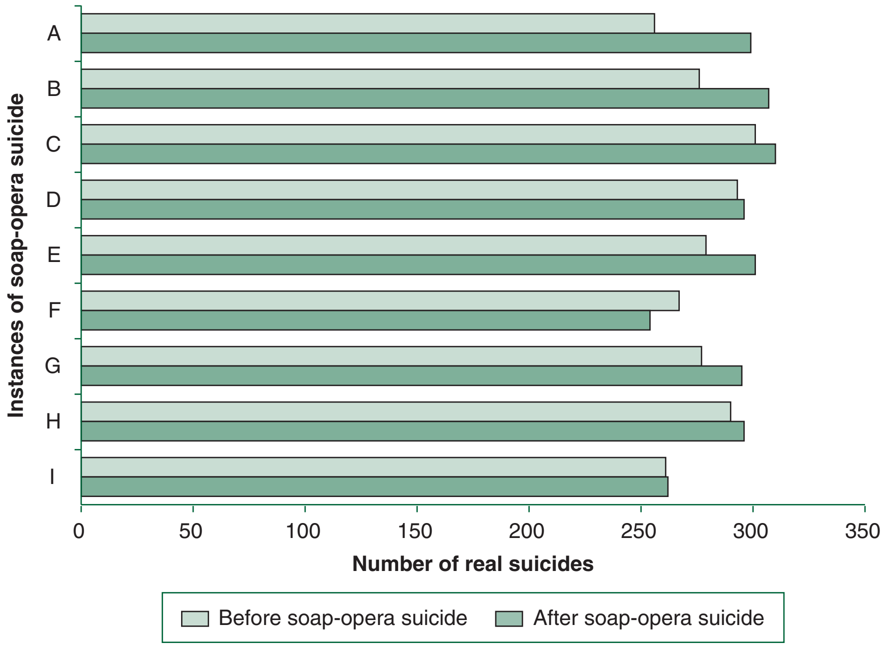
The simplest type of before-and-after design is the fixed-sample panel design. As you may recall from Chapter 2, in a panel design the same individuals are studied over time, the research may entail one pretest and one posttest. However, this type of before-and-after design does not qualify as a quasi-experimental 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.

David P. Phillips's (1982) study of the effect of TV soap-opera suicides on the number of actual suicides in the United States illustrates a more powerful **multiple group before-and-after design.** In this design, before-and-after comparisons are made of the same variables between different groups. Phillips identified 13 soap-opera suicides in 1977 and then recorded the U.S. suicide rate in the weeks prior to and following each TV story. In effect, the researcher had 13 different before-and-after studies, one for each suicide story. In 12 of these 13 comparisons, deaths due to suicide increased from the week before each soap-opera suicide to the week after (see Exhibit 5.5). Phillips also found similar increases in motor-vehicle deaths and crashes during the same period, some portion of which reflects covert suicide attempts. (Despite his clever design, however, some prominent researchers have disputed his findings.)

Another type of before-and-after design involves multiple pretest and posttest observations of the same group. **Repeated measures panel designs** include several pretest and posttest observations, allowing the researcher to study the process by which an intervention or treatment has an impact over time; hence, they are better than a simple before-and-after study.

Time series designs include many (preferably 30 or more) such observations in both pretest and posttest periods. They are particularly useful for studying the impact of new laws or social programs that affect large numbers of people and that are readily assessed by some ongoing measurement. For example, we might use a time series design to study the impact of a new seat-belt law on the severity of injuries in

Exhibit 5.5 Real Suicides and Soap-Opera Suicides



Source: David P. Phillips, 1982. "The Impact of Fictional Television Stories on U.S. Adult Fatalities: New Evidence on the Effect of the Mass Media on Violence." *American Journal of Sociology, 87* (May 1982):1340. Copyright © 1982 by the University of Chicago Press. Reprinted with permission.

automobile accidents, using a monthly state government report on insurance claims. Special statistics are required to analyze time series data, but the basic idea is simple: identify a trend in the dependent variable up to the date of the intervention, then project the trend into the postintervention period. This *projected* trend is then compared to the *actual* trend of the dependent variable after the intervention. A substantial disparity between the actual and projected trends is evidence that the intervention or event had an impact (Rossi & Freeman, 1989:260–261, 358–363).

How well do these before-and-after designs meet the five criteria for establishing causality? The before-after comparison enables us to determine whether an *association* exists between the intervention and the dependent variable (because we can determine whether there was a change after the intervention). They also clarify 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 observe; *spuriousness* may be a problem. Some other event may have occurred during the study that resulted in a change in posttest scores. 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.

Ex Post Facto Control Group Designs

The **ex post facto control group design** appears to be very 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. Like nonequivalent control group designs, this design has experimental and comparison groups that are not created by random assignment. But unlike the groups in nonequivalent control group designs, the groups in ex post facto designs are designated after the treatment has occurred. The problem with this is that if the treatment takes any time at all, people with particular characteristics may select themselves for the treatment or avoid it. Of course, this makes it difficult to determine whether an association between group membership and outcome is spurious. However, the particulars will vary from study to study; in some circumstances we may conclude that the treatment and control groups are so similar that causal effects can be tested (Rossi & Freeman, 1989:343–344).

Susan Cohen and Gerald Ledford's (1994) study of the effectiveness of self-managing teams used a well-constructed ex post facto design. They studied a telecommunications company with some work teams that were self-managing and some that were traditionally managed (meaning that a manager was responsible for the team's decisions). Cohen and Ledford found the self-reported quality of work life to be higher in the self-managing groups than in the traditionally managed groups.

WHAT ARE THE THREATS TO VALIDITY IN EXPERIMENTS?

Experimental designs, like any research design, must be evaluated for their ability to yield valid conclusions. Remember, there are three kinds of validity: internal (or causal), external (or 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 also a central concern for both kinds of research, but even 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

Exhibit 5.6 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.	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

validity. We will introduce survey and field research designs in the following chapters; 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 10 threats to validity (also referred to as "sources of invalidity") that occur frequently in social science research (see Exhibit 5.7). 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 bestseller 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 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.

In the Sherman and Berk (1984) domestic violence experiment in Minneapolis, some police officers sometimes violated the random assignment plan when they thought the circumstances warranted arresting a suspect who had been randomly assigned to receive just a warning; thus, they created a selection bias in the experimental group.

- *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. This is not a likely problem in a laboratory experiment that occurs in one session, but some laboratory experiments occur over time, and so differential attrition can become a problem. Subjects who experience the experimental condition may become more motivated to continue in the experiment than comparison subjects.

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 scales, for instance, becomes "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.

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 to just 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 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 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 football team that goes 0–12 one season almost has to improve. 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 didn't; 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 **regression.**

Regression effects: A source of causal validity 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 have to do with 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 O. J. Simpson trial was taking place, for instance, participants' thoughts on race relations at the end of the workshops may say less about you 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, 1985:276–277).

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 and may perform worse than expected as a result. Both compensatory rivalry and demoralization thus distort the impact of the experimental treatment.

The danger of contamination can be minimized if the experiment is conducted in a laboratory, if members of the experimental group and the comparison group have no contact while the study is in progress, and if the treatment is relatively brief. Whenever these conditions are not met, the likelihood of contamination increases.

Treatment Misidentification

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

Expectancies of experiment staff—Change among experimental subjects may be due to the positive expectancies of the 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.** However, in experiments on the effects of treatments such as medical drugs, **double-blind procedures** can be used: Staff delivering the treatments do not know which subjects are getting the treatment and which are receiving a placebo—something that looks like the treatment but has no effect.

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: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.

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 famous productivity experiment at the Hawthorne electric plant outside Chicago. No matter what conditions the researchers changed in order 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 more recent scholars suggest 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, 1985:272–274). Periodic measures are taken throughout an experiment to assess whether the treatment is being delivered as planned. For example, Drake et al. (1996) collected process data to monitor the implementation of two employment service models that they tested. One site did a poorer job of implementing the individual placement and support model than the other site, although the required differences between the experimental conditions were still achieved. Process analysis is often a special focus in evaluation research because of the possibility of improper implementation of the experimental program.

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.

Sample Generalizability

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 are unlikely to be a representative sample of any large population of interest to social scientists. 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, 1979:73–74). The more artificial the experimental arrangements, the greater the problem (Campbell & Stanley, 1966:20–21).

In some limited circumstances, a researcher may be able to sample subjects randomly for participation in an experiment and thus select a generalizable sample—one that is representative of the population from which it is selected. This approach is occasionally possible in **field experiments.** For example, some studies of the effects of income supports on the work behavior of poor persons have randomly sampled persons within particular states before randomly assigning them to experimental and comparison groups. Sherman and Berk's (1984) field experiment about the impact of arrest in actual domestic violence incidents (see Chapter 2) used a slightly different approach. In this study, all eligible cases were treated as subjects in the experiment during the data collection periods. As a result, we can place a good deal of confidence in the generalizability of the results to the population of domestic violence arrest cases in Minneapolis.

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, 1979:76–77).

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. Price, Van Ryn, and Vinokur (1992) found that intensive job-search assistance reduced depression among individuals who were at high risk for it because of other psychosocial characteristics; however, the intervention did not influence the rate of depression among individuals at low risk for depression. This is an important limitation on the generalizability of the findings, even if the sample taken by Price et al. was representative of the population of unemployed persons.

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. For example, Sherman and Berk's research (see Chapter 2) found that arrest did not deter subsequent domestic violence for unemployed individuals; arrest also failed to deter subsequent violence in communities with high levels of unemployment.

There is always an implicit tradeoff 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. College students are easy to recruit and to assign to artificial but controlled manipulations, but both practical and ethical concerns preclude this approach with many groups and with respect to many treatments. 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 (Campbell & Russo, 1999:143).

Exhibit 5.7 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

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 5.7). 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 compared to those who did not.

As you can see, there is no single procedure that 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.

HOW DO EXPERIMENTERS PROTECT THEIR SUBJECTS?

Social science experiments often involve subject deception. Primarily because of this feature, some experiments have prompted contentious debates about research ethics. Experimental evaluations of social programs also pose ethical dilemmas because they require researchers to withhold possibly beneficial treatment from some of the subjects just on the basis of chance. Such research may also yield sensitive information about program compliance, personal habits, and even illegal activity—information that is protected from legal subpoenas only in some research concerning mental illness or criminal activity (Boruch, 1997). In this section, we will give special attention to the problems of deception and the distribution of benefits in experimental research.

Deception

Deception occurs when subjects are misled about research procedures in order to determine how they would react to the treatment if they were not research subjects. Deception is a critical component of many social experiments, in part because of the difficulty of simulating real-world stresses and dilemmas in a laboratory setting. Stanley Milgram's (1965) classic study of obedience to authority provides a good example. Volunteers were recruited for what they were told was a study of the learning process. The experimenter told the volunteers they were to play the role of "teacher" and to administer an electric shock to a "student" in the next room when the student failed a memory test. The shocks were phony (and the students were actors), but the real subjects, the volunteers, didn't know this. They were told to increase the intensity of the shocks, even beyond what they were told was a lethal level. Many subjects continued to obey the authority in the study (the experimenter), even when their obedience involved administering what they thought were potentially lethal shocks to another person.

But did the experimental subjects actually believe that they were harming someone? Observational data suggest they did: "Persons were observed to sweat, tremble, stutter, bite their lips, and groan as they found themselves increasingly implicated in the experimental conflict" (Milgram 1965:66).

Verbatim transcripts of the sessions also indicated that participants were in much agony about administering the "shocks." So it seems that Milgram's deception "worked"; moreover, it seemed "necessary," since Milgram could not have administered real electric shocks to the students, nor would it have made sense for him to order the students to do something that wasn't so troubling, nor could he have explained what he was really interested in before conducting the experiment. The real question: Is this sufficient justification to allow the use of deception?

Aronson and Mills's study (1959) of severity of initiation (at an all-women's college in the 1950s) provides a very different example of the use of deception in experimental research—one that does not pose greater-than-everyday risks to subjects. The students who were randomly assigned to the "severe initiation" experimental condition had to read a list of embarrassing words. Even in the 1950s, reading a list of potentially embarrassing words in a laboratory setting and listening to a taped discussion were unlikely to increase the risks to which students were exposed in their everyday lives. Moreover, the researchers informed subjects that they would be expected to talk about sex and could decline to participate in the experiment if this requirement would bother them. No one dropped out.

To further ensure that no psychological harm was caused, Aronson and Mills (1959) explained the true nature of the experiment to the subjects after the experiment, in what is called **debriefing**. The subjects' reactions were typical:

> None of the Ss expressed any resentment or annoyance at having been misled. In fact, the majority were intrigued by the experiment, and several returned at the end of the academic quarter to ascertain the result. (1959:179)

Although the American Sociological Association's *Code of Ethics* does not discuss experimentation explicitly, one of its principles highlights the ethical dilemma posed by deceptive research:

> (a) Sociologists do not use deceptive techniques (1) unless they have determined that their use will not be harmful to research participants; is justified by the study's prospective scientific, educational, or applied value; and that equally effective alternative procedures that do not use deception are not feasible, and (2) unless they have obtained the approval of institutional review boards or, in the absence of such boards, with another authoritative body with expertise on the ethics of research.
>
> (b) Sociologists never deceive research participants about significant aspects of the research that would affect their willingness to participate, such as physical risks, discomfort, or unpleasant emotional experiences. (American Sociological Association, 1997:3)

Selective Distribution of Benefits

Field experiments conducted to evaluate social programs also can involve issues of informed consent (Hunt, 1985:275–276). One ethical issue that is somewhat unique to field experiments is the **distribution of benefits:** How much are subjects harmed by the way treatments are distributed in the experiment? For example, Sherman and Berk's (1974) experiment, and its successors, required police to make arrests in domestic violence cases largely on the basis of a random process. When arrests were not made, did the subjects' abused spouses suffer? Price et al. (1992) randomly assigned unemployed individuals who had volunteered for job-search help to an intensive program. Were the unemployed volunteers who were assigned to the comparison group at a big disadvantage?

Is it ethical to give some potentially advantageous or disadvantageous treatment to people on a random basis? Random distribution of benefits is justified when the researchers do not know whether some treatment actually is beneficial or not—and, of course, it is the goal of the experiment to find out. Chance is as reasonable a basis for distributing the treatment as any other. Also, if insufficient resources are available to fully fund a benefit for every eligible person, distribution of the benefit on the basis of chance to equally needy persons is ethically defensible (Boruch, 1997:66–67).

CONCLUSION

Causation and the means for achieving causally valid conclusions in research is the last of the three legs on which the validity of research rests. In this chapter, you have learned about the five criteria used to evaluate the extent to which particular research designs may achieve causally valid findings. You have been exposed to the problem of spuriousness and the way that randomization deals with it. You also have learned why we must take into account the units of analysis in a research design in order to come to appropriate causal conclusions.

True experiments help greatly to achieve more valid causal conclusions—they are the "gold standard" for testing causal hypotheses. Even when conditions preclude use of a true experimental design, many research designs can be improved by adding some experimental components. However, although it may be possible to test a hypothesis with an experiment, it is not always desirable to do so. Laboratory experiments may be inadvisable when they do not test the real hypothesis of interest but test instead a limited version that is amenable to laboratory manipulation. It also does not make sense to test the impact of social programs that cannot actually be implemented because of financial or political problems (Rossi & Freeman, 1989:304–307). Yet the virtues of experimental designs mean that they should always be considered when explanatory research is planned.

We emphasize that understandings of causal relationships are always partial. Researchers must always wonder whether they have omitted some relevant variables from their controls or whether their experimental results would differ if the experiment were conducted in another setting or at another time in history. But the tentative nature of causal conclusions means that we must give more—not less—attention to evaluating the causal validity of social science research whenever we need to ask the simple question, "What caused variation in this social phenomenon?"

KEY TERMS

Association
Before-and-after design
Causal effect
Ceteris paribus
Compensatory rivalry (John Henry effect)
Contamination
Context
Control group
Debriefing
Differential attrition (mortality)
Distribution of benefits
Double-blind procedure
Endogenous change
Evaluation research
Ex post facto control group design
Expectancies of experimental staff
Experimental group
External event
Field experiment
Hawthorne effect
History effect
Matching
Mechanism
Multiple group before-and-after design
Nonequivalent control group design
Nonspuriousness
Placebo effect
Posttest
Pretest
Process analysis
Quasi-experimental design
Random assignment
Randomization
Regression effects
Repeated cross-sectional design
Repeated measures panel design
Selection bias
Self-fulfilling prophecy
Spurious relationship
Time order
Time series design
Treatment misidentification
True experiment

HIGHLIGHTS

- Three criteria generally are 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.

- Association between two variables by itself is insufficient evidence of a causal relationship. This point is commonly made by the expression "Correlation does not prove causation."

- The independent variable in an experiment is represented by a treatment or other intervention. Some subjects receive one type of treatment; others may receive a different treatment or no treatment. In true experiments, subjects are assigned randomly to comparison groups.

- Experimental research designs have three essential components: use of at least two groups of subjects for comparison, measurement of the change that occurs as a result of the experimental treatment, and use of random assignment. In addition, experiments may include identification of a causal mechanism and control over experimental conditions.

- Random assignment of subjects to experimental and comparison groups eliminates systematic bias in group assignment. The odds of there being a difference between the experimental and comparison groups on the basis of chance can be calculated. They become very small for experiments with at least 30 subjects per group.

- Random assignment and random sampling both rely on a chance selection procedure, but their purposes differ. Random assignment involves placing predesignated subjects into two or more groups on the basis of chance; random sampling involves selecting subjects out of a larger population on the basis of chance. Matching of cases in the experimental and comparison groups is a poor substitute for randomization because identifying in advance all important variables on which to make the match is not possible. However, matching can improve the comparability of groups when it is used to supplement randomization.

- Ethical and practical constraints often preclude the use of experimental designs.

- Quasi-experimental designs can be either a nonequivalent control group design or a before-and-after design. Nonequivalent control groups can be created through either individual matching of subjects or matching of group characteristics. In either case, these designs can allow us to establish the existence of an association and the time order of effects, but they do not ensure that some unidentified extraneous variable did not cause what we think of as the effect of the independent variable. Before-and-after designs can involve one or more pretests and posttests. Although multiple pretests and posttests make it unlikely that another, extraneous influence caused the experimental effect, they do not guarantee it.

- Ex post facto control group designs involve a comparison group that individuals could decide to join precisely because they prefer this experience rather than what the experimental group offers. This creates differences in subject characteristics between the experimental and control groups that might very well result in a difference on the dependent variable. Because of this possibility, this type of design is not considered a quasi-experimental design.

- 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 may profitably be answered by studying different units of analysis.

- Causal conclusions derived from experiments can be invalid because of selection bias, endogenous change, the effects of external events, cross-group contamination, or treatment misidentification. In true experiments, randomization should eliminate selection bias and bias due to endogenous change. External events, cross-group contamination, and treatment misidentification can threaten the validity of causal conclusions in both true experiments and quasi-experiments.

- Process analysis can be used in experiments to identify how the treatment had (or didn't have) an effect—a matter of particular concern in field experiments. Treatment misidentification is less likely when process analysis is used.

- The generalizability of experimental results declines if the study conditions are artificial and the experimental subjects are unique. Field experiments are likely to produce more generalizable results than experiments conducted in the laboratory.

- The external validity of causal conclusions is determined by the extent to which they apply to different types of individuals and settings. When causal conclusions do not apply to all the subgroups in a study, they are not generalizable to corresponding subgroups in the population; consequently, they are not externally valid with respect to those subgroups. Causal conclusions can also be considered externally invalid when they occur only under the experimental conditions.

- Subject deception is common in laboratory experiments and poses unique ethical issues. Researchers must weigh the potential harm to subjects and debrief subjects who have been deceived. In field experiments, a common ethical problem is selective distribution of benefits. Random assignment may be the fairest way of allocating treatment when treatment openings are insufficient for all eligible individuals and when the efficacy of the treatment is unknown.

To assist you in completing the Web Exercises, please access the Study Site at http://www.pineforge.com/mssw2 where you'll find the Web Exercises with accompanying links. You'll find other useful study materials like 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 presented in the book.

EXERCISES

Discussing Research

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 of the five criteria for establishing causality are met? What other potentially important influences on the reported outcome have been overlooked? Can you spot any potentially spurious relationships?

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 each 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?

3. The practice CD-ROM contains lessons on units of analysis. Choose the Units of Analysis lesson from the main menu. It describes several research projects and asks you to identify the units of analysis in each.

4. The National Institutes of Health provides a tutorial for learning about current ethical standards in research. Complete this tutorial at http://cme.nci.nih.gov/intro.htm. Be prepared to spend one-half to one hour completing the tutorial. You must register as a college student and provide a bit of other information. Indicate that you do not need a certificate of completion. After you complete the registration fields, begin with the section on History. In this section, you will find a subsection on "The Development of Codes of Research Ethics." When you get to the heading in this subsection on the "Belmont Report," you will find a link to the federal "Common Rule" document. Click on this link and take the time to print the document out and read it. When you are finished with the tutorial and have read the Common Rule, you will be well on your way to becoming an expert on human subjects regulations. Identify the human subjects rules that are most important for research on human subjects.

Finding Research

1. Read an original article describing a social experiment. (Social psychology "readers," collections of such articles for undergraduates, are a good place to find interesting studies.) Critique the article, using as your guide the article review questions presented in Exhibit 10.2. Focus on the extent to which experimental conditions were controlled and the causal mechanism was identified. Did inadequate control over conditions or inadequate identification of the causal mechanism make you feel uncertain about the causal conclusions?

2. Go to the Web site of the Community Policing Consortium at www.communitypolicing.org/about2.html. What causal assertions are made? Pick one of these assertions and propose a research design with which to test this assertion. Be specific.

3. Go to Sociosite at www.pscw.uva.nl/sociosite. Choose "Subject Areas." Choose a sociological subject area you are interested in. Find an example of research that has been

done using experimental methods in this subject. Explain the experiment. Choose at least five of the Key Terms listed at the end of this chapter that are relevant to and incorporated in the research experiment you have located on the Internet. Explain how each of the five Key Terms you have chosen plays a role in the research example you found on the Web.

Critiquing Research

1. From newspapers or magazines, find two recent studies of education (reading, testing, etc.). For each study, list in order what you see as the most likely sources of internal invalidity (e.g., selection, mortality, etc.).

2. Select a true experiment, perhaps from the *Journal of Experimental and Social Psychology,* the *Journal of Personality and Social Psychology*, or sources suggested in class. Diagram the experiment using the exhibits in this chapter as a model. Discuss the extent to which experimental conditions were controlled and the causal mechanism was identified. How confident can you be in the causal conclusions from the study, based on review of the threats to internal validity discussed in this chapter: selection bias, endogenous change, external events, contamination, and treatment misidentification? How generalizable do you think the study's results are to the population from which the cases were selected? To specific subgroups in the study? How thoroughly do the researchers discuss these issues?

3. Repeat the previous exercise with a quasi-experiment.

4. Critique the ethics of one of the experiments presented in this chapter, or some other experiment you have read about. What specific rules do you think should guide researchers' decisions about subject deception and the selective distribution of benefits?

Doing Research

1. Try out the process of randomization. Go to the Web site www.randomizer.org. Now just type numbers into the randomizer for an experiment with two groups and 20 individuals per group. Repeat the process for an experiment with 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. Participate in a social psychology experiment on the Internet. Go to www.socialpsychology.org/expts.htm. Pick an experiment in which to participate and follow the instructions. After you finish, write a description of the experiment and evaluate it using the criteria discussed in the chapter.

3. Volunteer for an experiment. Contact the psychology department and ask about opportunities for participating in laboratory experiments. Discuss the experience with your classmates.